Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Extracting Chinese Question-Answer Pairs from Online Forums*

Baoxun Wang Bingquan Liu Chengjie Sun Xiaolong Wang Lin Sun
School of Computer Science and Technology
Harbin Institute of Technology
Harbin, China
{bxwang, liubq, cjsun, wangxl, lsun}@insun.hit.edu.cn

***Abstract*—Extracting question-answer pairs from online forums is a meaningful work due to the huge amount of valuable user generated resource contained in forums. In this paper we consider the problem of extracting Chinese question-answer pairs for the first time. We present a strategy to detect Chinese questions and their answers. We propose a sequential rule based method to find questions in a forum thread, then we adopt non-textual features based on forum structure to improve the performance of answer detecting in the same thread. Experimental results show that our techniques are very effective.**

***Keywords*—question answering, labeled sequential rules, non-textual features, classification, information extraction**

I. INTRODUCTION

Online forums are web applications for people to hold discussions in certain domains, such as techniques, travel, sports, etc. For the researchers focusing question-answer systems, forums contain a huge number of question-answer pairs generated by users, which can be excellent knowledge resources for building question answering systems. So extracting question-answer pairs from the human generated contents in forum threads is of great value.

In this paper we focus on the mining of knowledge from Chinese online forums in the form of QA pairs. The value of QA pairs is obvious to both commercial QA services and artificial intelligence researches. Acquiring question-answer pairs automatically has the following two meanings. First, for most commercial QA services, such knowledge becomes essential to provide QA search and community-based question-answering functions. The latter application, always called CQA for short, has become more and more popular nowadays, such as Yahoo! Answers and Baidu[1]. At the same time large QA data perform a key role in offering exact answers, thus a CQA system with larger QA data attracts more users. Second, large number of QA pairs is the base of the research on automatic question-answer systems, which has been a hotspot in artificial intelligence field. Now most researchers take the QA pairs as their training and testing data. Large scale of QA pairs can meet the researchers' requirement to learn valuable knowledge, extract useful patterns and finally locate the answers which users really need. Recently many researchers have realized the importance of the resources existing in the CQA systems, but much less work has notice the data embedded in the forums.

Having investigated the main Chinese popular forums we find large number of forum users tend to ask for help or get knowledge from others by asking questions, at the same time people are also willing to share their own knowledge by means of offering answers. Although it is highly valuable to detect and extract QA pairs from online forums, some attributes of post contents bring difficulties in solving this problem to researchers. The characteristics of contents in forums can be concluded as follows:

- A human generated post usually includes a very short content, which always have much fewer sentences than that of web pages. This means that some useful models for similarity computing become not that powerful any longer when facing forum contents, and vector space model is an example for it, the reason of which we will explain in section III. Moreover, the short contents can not provide enough semantic or logical information for deep language processing.
- When a person is initializing or replying a post, he/she tends to use an informal tone, which is more closed to his/her oral habit. Actually this phenomenon is naturally determined by forums' function of virtual community. Compared with western languages, Chinese is more flexible, which makes parsing the sentences of Chinese forum contents with a set of rules become very hard.
- A forum usually shows a simple simulated structure of real community. Although this kind of structure is not as complex as real society, we can still obtain some helpful features to solve the problem of question-answer pair extraction. Using such virtual community based features to do answer detection is one of the points of this paper.

An example thread from real forum data is given in figure 1, where a question about how to realize a stack using C++ is put

* This investigation was supported by the project of the High Technology Research and Development Program of China (grants No. 2006AA01Z197 and 2007AA01Z172), the project of the National Natural Science Foundation of China (grants No. 60435020 and 60673037) and the project of the Natural Science Foundation of Heilongjiang Province (grant No. E200635).
[1] Baidu: http://zhidao.baidu.com

forward in the first post, and people give their opinions about this question in the following posts. From this example we can clearly see the characteristics of forum contents which have been listed above. The content in italic is the question in this thread, and the ones in bold and italic are its answers. We can see all of them are short and informal. Among the posts, P3 is the direct answer to P0, because we can identify it by similarity features; but we must agree that P4 is also an answer to the question which can not be found by using ordinary textual features, however, the content in bold in P5 is helpful for us to extract P4 as an answer. Although P1 and P2 are not answers to this question, they can still provide some information on question detection. From P1 and P2 we can see that locating question words simply is not a good idea.

Thread title: 初学者请教一个问题

P0(zhou): *大家好，我是一个 C++的初学者，请教各位一个问题：如何用 C++实现栈？*

P1(Bull): 太初级了，大家都知道栈***如何***实现。

P2(cobras): 楼上的你***怎么***可以这样说，太伤人了。

P3(samon): ***C++的 STL 中有栈结构的实现，可以直接拿来用，很方便。***

P4(cookie): ***楼上正解，具体方法可以参考《C++ Primer》这本书***

P5(zhou): **感谢楼上，已经找到了！**

Figure 1. An example thread from real forum data

In this paper, we present an approach to detect question-answer pairs in threads of Chinese forums, which consists of two components: question detection and answer detection. The target of question detection is to find all the questions within a thread. This work seems easy to complete by identifying question marks at first, but the real forum data show this problem is not so trivial. In this paper we propose a rule based question detecting algorithm, which has the same effect as the pattern method. Taking the questions detected as basic information, we aim to find the corresponding answers for each question in the same thread. The first two characteristics of forum data above are a little disappointing, but the last one is quite exciting, because some useful non-textual features may be introduced into our method. In this paper we take answer detection as a classification problem with a set of textual and non-textual features.

The contributions of this paper are as follows:

- Our work considers Chinese question-answer pairs for the first time. To our knowledge, there is no previous work for this problem.
- We employ virtual community structure based non-textual features by mixing them with the semantic similarity features. The non-textual features are important because they are not related to the post contents, so they can not be affected by the diversity of user generated data.
- We conducted experiments on real Chinese forum data. Experimental results show that 1) our sequential rule based question detecting method outperforms the simple question words strategy and the question mark strategy, and the coverage rate of the rules is satisfying; 2) our answer detection using mixed feature outperforms the traditional IR approach and the pure semantic similarity method, which demonstrates that non-textual features improve answer detection.

The rest of the paper is organized as follows: Section II discusses the related work. Section III presents the propose techniques, including sequential rule based question detection and answer detection using mixed features. We evaluate our techniques in section IV. Section V concludes this paper and discusses the future work.

II. RELATED WORK

Until now, little work has been done to extract question-answer pairs from Chinese forums. There is some research work on acquiring knowledge from discussion threads. Huang et al. [1] extracted input-reply pairs by using classification method. Feng et al [2] used cosine similarity to match users' query with reply posts for discussion-bot. Although their work focuses on mining knowledge in QA field, the definition of our problem is different from theirs due to the difference of application background.

Techniques on question-answer retrieval have also attracted more researchers' attention, the main task of which is to find similar questions and their corresponding answers with the question given by a user. Current research mainly focuses on retrieval of CQA [3] and FAQ [4, 5]. The work in this field is not directly related to ours, but how to bridge the lexical gap is the common problem these papers talking about. In this paper, when trying to locate the answers of a found question, we have to find a way to overcome the lexical chasm between questions and answers. [6] retrieves question-answer pairs from FAQ pages, whose task is easier than ours.

Comparing with question-answer retrieval, extracting QA pairs for email summarization seems more related to our work. Early email summaries are formed by collecting overview sentences [7, 8]. Shrestha and McKeown's work [9] is closer to QA pair extracting, because they organize email summaries by gathering the question and answer sentences of email threads with a classification method. The reason we consider email summarization as a related problem is that we both pay attention to locate and extract essential sentences to summarize a topic or answer a question. We also notice the work of Carenini et al. [10], which tries to improve the quality of email summaries using a more complex method based on clue words.

Exploratory work on mining question-answer pairs from English forums has been done by S. Ding [11] and G. Cong [12]. 2-demension conditional random fields (2-D CRF) model is introduced in [11] to extract answers and their contexts, where the 2-D CRF has promoted the precision but brought higher time consumption. [12] presents a graph based propagation method to detect answers. Because of the unsupervised strategy, [12] achieves higher efficiency. Both of

the papers above claim that question detecting is a non-trivial problem, and [12] uses a pattern based method to solve it.

Both [11] and [12] focus on English QA pairs, instead we try to extract Chinese QA pairs from forum data. To our best knowledge, there has been no previous work dealing with this problem. In this paper we adopt a sequential rule based method, and the experiment shows our method has the same effect. To detect answers, we focus on both textual and non-textual feature collection base on the structure of virtual communities. Similar work has been done in J. Jeon[13]'s work to predict the quality of answers.

III. Question and Answer Detection

A. Sequential Rule Based Question Detection

Chinese has a flexible grammar system. On one side, we can not hope to detect all the questions in forums by identifying question mark and question words simply, on the other, establishing a rule set is not a wise choice. [12] proposes a question detecting algorithm based on labeled patterns and achieves a rather high precision. In contrast, questions in Chinese forums have such great diversity that we may need a very huge amount of patterns to solve this problem in the same way as [12] does. To learn a pattern for every question that appears possibly is unpractical. Besides, the elements labeled by part-of-speech (PoS) in a pattern can not provide much information to find a question, but they increase the consumption of preprocessing time.

Next we will explain our sequential rule based approach to question detection. Let BW be the set of beginning words, e.g. "请问", "请教" and "求助" etc, QW be the set of question words corresponding to 5W1H words in English, MW be the set of mood words which usually appear in questions and QM be the singleton set containing only the question mark. The sequential rules can be expressed as follows:

Rule #1: $BW \times QW \times MW \times QM$

Rule #2: $QW \times MW \times QM$

Rule #3: $BW \times QW \times QM$

Rule #4: $BW \times QW \times MW$

Rule #5: $QW \times MW$

Rule #6: $BW \times QW$

Rule #7: $QW \times QM$

It is clear that locating question words and the question mark is the most popular idea, but the precision is usually not satisfying. After a survey of real forum data we find this rule can cover only a small number of questions. In the research of question detection, beginning words and mood words are easily ignored. Beginning words are used to lead a question, and most people think they are not a part of the trunk of the sentences. Mood words, which compose a part of Chinese stop word list, are always removed by preprocessing module. In fact both of the two kinds of words are helpful to judge whether a sentence is a question or not.

Our method comes from the analysis of the posting habits of Chinese people in virtual communities. In forums people prefer colloquial to written style. To be polite people like to use beginning words to lead a question, while mood words are sometimes put at the end of a question to strengthen the tone. Contrasting with the labeled patterns the rules we present need no tagging work while keeping a high predict precision. At the same time, any necessary words can be inserted into the blanks between the elements of the rules, and this property makes the rules have a strong adaptability.

B. Answer Detection Using Mixed Features

In this part we present our techniques to find answers in forums for detected questions. We take the threads with annotated questions as input and the output is a list of ranked answers for each question. As we have mentioned in section I, the contents of the posts in forum are mostly short and informal, so pure textual features can not lead to a high precision on answer detection. In this paper we extract a collection of non-textual features based on the community characteristic of forums. A new kind of feature vector is formed by combining the non-textual and textual features, which improves the precision of classifier.

The basic reason of introducing textual features to detect answers is that an answer should be similar with its corresponding question, so the most intuitive idea to find an answer to a given question is computing the similarity between them. Some classical information retrieval models may be introduced to solve the problem, among which cosine similarity is the most popular. Given a question q and a candidate answer a, their cosine similarity can be computed as follows:

$$\cos(q,a) = \frac{\sum_{k=1}^{n} w_{q_k} \times w_{a_k}}{\sqrt{\left(\sum_{k=1}^{n} w_{q_k}^2\right)} \times \sqrt{\left(\sum_{k=1}^{n} w_{a_k}^2\right)}} \quad (1)$$

Where w_{q_k} and w_{a_k} stand for the weight of a given word in question and answer respectively, which can be computed by the product of term frequency (*tf*) and inverse document frequency (*idf*). This model has achieved significant success in IR field, but for our problem it is not so powerful. The reason is that the number of words shared by a QA pair is always rather small, at the same time the word frequency in the content is mostly 0 or 1, which leads to the fact that sometimes the similarity between a question and a candidate answer is very high, but their cosine similarity is still a value approaching 0.

Taking lexical gap as the main reason that causes the failure of cosine similarity, we try to bridge the gap by using semantic similarity. In this paper we adopt two kinds of textual features to replace the less powerful cosine similarity feature. Firstly we count the notional words shared by a question and its candidate answer, since we consider the shared words of great value.

Secondly we introduce semantic similarity based on HowNet[2], which can be computed as follows:

$$sim(q,a)=\frac{\sum_{k=1}^{n} sem\_sim\left(w_{q_k},w_{a_k}\right)}{\sum_{k=1}^{n} idf_{q_k}} \quad (2)$$

Where the function $sem\_sim\left(w_{q_k},w_{a_k}\right)$ computes the semantic similarity between two corresponding words based on HowNet using a greedy algorithm.

The characteristics of forum data determine that even textual features such as semantic similarity can not deal with the problem alone. In order to enrich enough information for classifier, e.g. SVM in this paper, we extract some non-textural features. The following is a detailed explanation of each individual non-textual feature.

- **Answer Length** The length of the answer. We take this feature as a non-textual one because it doesn't need any serious analysis of the content to get the length of the answer, which is helpful in measuring the quality of online information.
- **Distance between Question and Answer** The post including the answer to a given question in a forum thread should not be too far away from the one where the question appears.
- **Answerer's Activity** If a user answers many times in a forum, the user may have more knowledge in a certain field, and his/her candidate answers are more likely to become the real answer.
- **Quote Counts** The number of times that users quote a post. An answer has more opportunities to be quoted in a thread if it contains useful information.
- **Presence of acknowledgement** Due to the basic manners in virtual communities, Questioners always express acknowledgement for getting help from the answers. If a candidate answer receives acknowledgement, it is likely a real answer to the question.

We list all the features we use for answer extraction in the table below, including 4 textual features and 5 non-textual features. The last column shows numerical types of the features in our programs. We will use the features above to form a feature vector and train a classifier by SVM.

TABLE I. FEATURES FOR CANDIDATE ANSWERS

Features	*Kind*	*Type*
Number of words shared with question	Textual	Integer
Similarity with question based on HowNet	Textual	Float
Number of words shared with thread topic	Textual	Integer
Similarity with thread topic based on HowNet	Textual	Float
Answer length	Non-textual	integer
Distance from question	Non-textual	Integer
Answerer's activity	Non-textual	Integer
Quote Counts	Non-textual	Integer
Presence of acknowledgement	Non-textual	Binary

[2] HowNet: an electronic world knowledge system, which serves as a powerful tool for meaning computation in HLT. Detail information can be found in: http://www.keenage.com/

Support Vector Machines (SVM) are a set of supervised learning algorithms first introduced by Vapnik [14]. Given a training set of instance-label pairs (x_i,y_i), $i=1,2,...,l$, where $x_i \in R^n$ and $y \in \{-1,1\}^l$, the SVM require the solution to the following optimization problem:

$$\min \quad \frac{1}{2}w^T w + C\sum_{i=1}^{l}\varepsilon_i \quad (3)$$

Subject to:

$$y_i\left(w^T\varphi(x_i)+b\right)\geq 1-\varepsilon_i \quad (4)$$

$$\varepsilon_i \geq 0 \quad (5)$$

Here training vectors x_i are mapped into a higher dimensional space by the function Φ. A hard classifier implementing the optimal separating hyperplane in the feature space is given by:

$$f(x)=sign\left(\sum \alpha_i K(x_i,x)+b\right) \quad (6)$$

Where $K(x_i,x)=\phi(x)\phi(x_i)$ is the kernel function, in this paper we chose RBF kernel as our kernel function as following:

$$K(x,x_i)=\exp\left(-\gamma\left\|x-x_i\right\|^2\right),\gamma>0 \quad (7)$$

For training the SVM, we use the libsvm [15] toolbox. The parameter C of SVM is chosen through 5-fold cross validation.

IV. EXPERIMENTS

In this section, we will design experiments to evaluate the techniques for question detection and answer detection.

A. Corpus

We obtained 80,000 threads from CSDN[3] forum, and randomly selected 600 threads. On average each thread in our corpus consists of about 3.5 posts and each contains at least two posts. For training and testing five annotators were asked to tag questions and their answers in each thread. The main

[3] CSDN: http://community.csdn.net

language is Chinese in our corpus, which has comparable scale with that in [11] and [12]. For metrics, we calculate precision, recall for both our method and baselines.

B. Experimental Results

To evaluate the performance of our question detection method, we first compute the coverage rate of the identification rules presented in section III. Table II lists the coverage of our seven rules.

TABLE II. COVERAGE OF RULES

Rules	*Coverage (%)*
Rule #1	29.6
Rule #2	17.2
Rule #3	18.7
Rule #4	14.5
Rule #5	5.1
Rule #6	5.8
Rule #7	5.6

From the result we can make some intuitive conclusions. Firstly, there are about 25% of the questions without the question mark, which means the method of identifying question mark could not achieve a high recall. Secondly, the rules can not cover all the forms of questions in forums, about 3% of the questions can not be found by our rules. Thirdly, most questions contain beginning words, and we attribute this fact to Chinese tradition of courtesy.

Table III shows the performance of question detection methods. The *question words* method is to locate question words in a sentence to decide whether the sentence is a question or not. From the results we can see although *question words* method achieves good recall, its precision is low. The reason is most questions contain question words, while there are a number of sentences with question words which are not real questions, e.g. "这个问题***能不能***解决已经不太重要了。 (*Whether* this problem can be solved is not so important.) ", "大家都知道栈***如何***实现。 (Everyone knows *how* to realize a stack.) " and "你***怎么***可以这样说。 (*How* can you say that.) " The precision of *question mark* method is rather satisfying because the sentence with a question mark is very likely to be a question, but there are about 25% of questions without a question mark according to our corpus. Our rules can cover 97% of the questions, and the sequential rule based method outperforms the simple rule approaches, thus the precision and the recall have been significantly improved.

TABLE III. PERFORMANCE OF QUESTION DETECTION

Method	*Precision (%)*	*Recall (%)*
Question Words	81.4	96.3
Question Mark	95.7	76.0
Our	96.2	95.8

Detecting answers in a given thread is much more difficult than question detection. Table IV gives the evaluation of our answer detection strategy, comparing with two baseline methods. Cosine similarity method aims to compute the cosine value of the angle between two sentence vectors built by word frequency. From the results we can see the performance of this method is very poor, the main reason of which is the lack of shared words between questions and their answers. Besides, the low frequency of words in the contents of posts also makes the sentence vectors less meaningful. Since a direct idea to improve the performance of similarity computing is introducing semantic information to overcome the gap between questions and answers, we have designed a semantic similarity based method, including semantic similarity computed with HowNet and the count of shared words as its parameters. Unfortunately, the result of this approach is not as good as expected, the possible reason could be that post contents are usually much shorter than common texts. By adopting non-textual features, our classification based strategy has outperformed the former two methods significantly, with the precision and the recall both over 60%.

TABLE IV. PERFORMANCE OF ANSWER DETECTION

Method	*Precision (%)*	*Recall (%)*
Cosine similarity	10.2	15.9
Semantic similarity	24.7	36.1
Our	61.5	65.8

The research on English QA pair detection in forums has been done in [11] and [12]. [11] considers answer detection as a sequential labeling problem and uses conditional random fields to find answers. Among the results of [11], the precision of answer detection of their 2D CRF + Skip chain CRF approach has reached 66.90%, and their recall is over 70%. [12] takes a different way from [11] and our work to express their evaluation results, but their results are comparable with [11]. In this paper, we obtain a 61.5% precision and a 65.8% recall, which are a little lower than those of [11] and [12]. The reason is the contents in Chinese forum posts are more difficult to handle with more variable and informal expressing ways than English. Besides, our corpus comes from the IT forum where people mainly discuss programming with large amounts programming codes embedded in the contents of posts, which brings challenges to information extraction. Our model is simpler than those in the previous two papers, but this strategy has better efficiency.

V. CONCLUSIONS

In this paper we present an approach to detect Chinese question-answer pairs from online forums. To the best of our knowledge, our work is the first one that aims to extract Chinese question-answer pairs from forums. Our work can be applied to enrich the knowledge base of online QA service and provide corpus for other automatic QA researches. To find questions in forum posts, we propose a set of sequential rules which covers most of the questions. For answer detecting, non-textual features based on virtual community's characteristics are mixed with semantic textual features to improve the

performance of classifier. Experimental results on real forum data show that our strategy is effective.

This is our preliminary work on detecting Chinese QA pairs from online forums, so there is still space for improvement. In the future, we will focus on the following problems: 1) to adopt more features for answer detection, including textual features based on deep semantic mining and forum structure based non-textual features; 2) to model QA pair detection in a specific way and build a suitable framework for finding QA pairs in forums; 3) to summarize the multiple answers to the same question by introducing text summarization techniques; and 4) to evaluate our techniques in various domains.

ACKNOWLEDGMENT

The authors would like to thank Ke Sun and Deyuan Zhang for their valuable suggestions in preparing this paper.

REFERENCES

[1] J. Huang, M. Zhou, and D. Yang, "Extracting chatbot knowledge from online discussion forums", In Proceedings of IJCAI, pp.423-428, 2007.

[2] D. Feng, E. Shaw, J. Kim, and E. Hovy, "An intelligent discussion-bot for answering student queries in threaded discussions", In Proceedings of Intelligent user interfaces, pp. 171-177, 2006.

[3] J. Jeon, W. B. Croft, and J. H. Lee, "Finding similar questions in large question and answer archives", In Proceedings of CIKM, pp. 84-90, 2005.

[4] A. L. Berger, R. Caruana, D. Cohn, D. Freitag, and V. O. Mittal, "Bridging the lexical chasm: statistical approaches to answer-finding", In Proceedings of SIGIR, pp.192-199, 2000.

[5] S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu, "Statistical machine translation for query expansion in answer retrieval", In Proceedings of ACL, pp. 464-471, 2007.

[6] V. Jijkoun and M. de Rijke, "Retrieving answers from frequently asked questions pages on the web", In Proceedings of CIKM, pp. 76-83, 2005.

[7] A. Nenkova and A. Bagga, "Facilitating email thread access by extractive summary generation", In Proceedings of RANLP, pp. 287-296, 2003.

[8] O. Rambow, L. Shrestha, J. Chen, and C. Lauridsen, "Summarizing email threads", In Proceedings of HLT-NAACL, 2004.

[9] L. Shrestha and K. McKeown, "Detection of question-answer pairs in email conversations", In Proceedings of COLING'04, pp. 889-895, 2004.

[10] G. Carenini, R. Ng, and X. Zhou, "Summarizing email conversations with clue words", In Proceedings of WWW, pp.91-100, 2007.

[11] S. Ding, G. Cong, C.-Y. Lin, and X. Zhu, "Using conditional random fields to extract contexts and answers of questions from online forums", In Proceedings of ACL, pp. 710-718, 2008.

[12] G. Cong, L. Wang, C.Y. Lin, Y.I. Song and Y Sun, "Finding Question-Answer Pairs from Online Forums", In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 467-474, 2008.

[13] J. Jeon et al, "A Framework to Predict the Quality of Answers with NonTextual Features", In proceeding of SIGIR, pp. 228-235, 2006.

[14] C. Cortes and V. Vapnik, "Support-Vector Networks", Machine Learning, vol. 20, pp. 273-297, 1995.

[15] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines", technical report, 2001.