Journal of Southeast University (English Edition) Vol. 24, No. 3, pp. 272 – 275 Sept. 2008 ISSN 1003—7985

Question classification in question answering based on real-world web data sets

Yuan Xiaojie Yu Shitao Shi Jianxing Chen Qiushuang

(College of Information Technical Science, Nankai University, Tianjin 300071, China)

Abstract: To improve question answering (QA) performance based on real-world web data sets, a new set of question classes and a general answer re-ranking model are defined. With predefined dictionary and grammatical analysis, the question classifier draws both semantic and grammatical information into information retrieval and machine learning methods in the form of various training features, including the question word, the main verb of the question, the dependency structure, the position of the main auxiliary verb, the main noun of the question, the top hypernym of the main noun, etc. Then the QA query results are re-ranked by question class information. Experiments show that the questions in real-world web data sets can be accurately classified by the classifier, and the QA results after re-ranking can be obviously improved. It is proved that with both semantic and grammatical information, applications such as QA, built upon real-world web data sets, can be improved, thus showing better performance.

Key words: question classification; question answering; real-world web data sets; question and answer web forums; re-ranking model

1 Related Work

Question answering (QA) has become one of the most popular issues in the web IR field in recent years. There has been a great deal of academic results up to now[1-2], while much of the work has been accomplished on well-formed testing data sets, and not adaptable in real-world dialogue contexts. Therefore, most QA systems cannot be easily transferred to industrial production.

This paper pays more attention to QA based on real-world web data sets, such as question and answer (QnA) web forums. The QnA web forum is an important type of Internet service, where users can help each other by asking a question, answering a question or voting for a best answer. Live QnA, Yahoo Answers, Baidu Knows and Wondir are examples of this type. Popular QnA web forums usually keep large quantities of QnA threads data with manually labelled best answers. These data sets cover almost every field in daily life, on which many valuable applications can be built.

However, in a QnA thread, the meaning of a word or a sentence may depend on the thread context, and the content may be filled with colloquial words, spelling mistakes and grammar mistakes. The QnA data sets are much noisier than other well-formed data sets, and the QA results based on them are not good enough when dealt with by the same method.

Some former studies have tried to impose a hierarchical classification regarding questions, or to extract head chunks and related words from sentences[3], or to extract word dependency in a sentence[4], aiming at obtaining a more exact description of the questions. In Ref. [5], bringing semantic relations into a QA system is proposed, but it mainly focuses on well-formed factoids and list questions. It is discussed in Ref. [6] to do QA using the web, not limited to simple factoid questions. That is a good change. In Ref. [7], a structured retrieval method for QA is proposed, with linguistic and semantic information. As a summary, in modern QA systems, question class information, semantic and grammatical information are of great importance.

In this paper, a QA system in the English language based on the Yahoo Answers forum data set is built, with a vector space model (VSM). It looks like a system to retrieve answers from frequently asked questions, like the work in Ref. [8]. But the results can be improved. Then another question classification (QC) model is trained with the Naïve Bayes model for result re-ranking purposes.

The main contributions of this paper include:

1) Paying more attention to QA based on real-world web data set using both information retrieval (IR) and machine learning (ML) approaches;

2) Defining a new set of question classes for real-world web data sets using QC as a re-ranking method;

3) Drawing both semantic and grammatical information as various training features;

4) Defining a general QA re-ranking model and improving the QA results with it.

2 Question Answering

First, a QA system based on real-world web data is built without question class information. The experimental result are treated as a baseline for comparison and evaluation purposes later.

2.1 Yahoo Answers data set

A real-world QnA data set from Yahoo Answers is used as the whole data set, which is part of the Yahoo Answers threads in the Birds category, containing 18 000 data items. Each data item is in the form of a triplet as 〈question title, question body, best answer body〉.

2.2 Vector space model

The QA system is implemented with a VSM model in the following steps:

Step 1 Divide the whole data set into two parts: 17 900

Received 2008-04-15.

Biography: Yuan Xiaojie (1963—), female, doctor, professor, yuanxj@nankai.edu.cn.

Foundation items: Microsoft Research Asia Internet Services in Academic Research Fund (No. FY07-RES-OPP-116), the Science and Technology Development Program of Tianjin (No. 06YFGZGX05900).

Citation: Yuan Xiaojie, Yu Shitao, Shi Jianxing, et al. Question classification in question answering based on real-world web data sets[J]. Journal of Southeast University (English Edition), 2008, 24(3): 272 – 275.

data items as the training set and the remaining 100 data items as the testing set;

Step 2 All the terms in the training set are used to calculate tf and idf values, but a higher weight is given to terms in the question title and body.

Step 3 In the testing set, only terms in the question title and the body are used to calculate tf values, idf values, and the similarities to data items in the training set. The best answer bodies in some similar data items in the training set are returned as the result answers.

2.3 Experimental result

For each test question, the earlier 20 result answers are returned. The validity of each answer to the question is manually labelled. Then the mean reciprocal rank (MRR) result can be calculated for the total 100 testing questions for evaluation.

Tab. 1 shows the MRR result. It is not good enough, but can be improved by QC as a re-ranking method.

Tab. 1 The mean reciprocal ranking result

Question count	Mean reciprocal ranking
100	0.299 926

3 Question Classification

QC means putting questions into several semantic categories. It has already been used for improving the results of the QA system in previous works. As in Refs. [9 − 10], a two-layered taxonomy is defined to represent a natural semantic classification for typical answers. The hierarchy contains 6 coarse classes and 50 fine classes. These classes help further processing to precisely locate and verify the answer, or provide answer type specific information.

3.1 Question class definition

Definition 1 (question class) Question class is a set of semantic categories into which all questions can be put, according to the sentence structure of the question. It implies the asker's intention.

In this paper, a set of 10 classes is defined for the real-world questions, as shown in Tab. 2.

Tab. 2 The question class definition

Question class	Question sample
Entity	What is your favourite kind of bird?
Human	Who was the first man keeping a bird?
Location	Where can I find swan eggs?
Numeric	What is the size of an owl nest?
Time	How long does a humming bird live?
Definition	What is bird flu?
Description	How do birds react to music?
Manner	How can I stop parakeets from biting me?
Reason	Why do roosters crow in the morning?
Yes/No	Does a bird make a good pet?

In real-world data sets, it is difficult to classify each question into a two-layered taxonomy, due to the data noise. There are only coarse classes here. Thus the question class information cannot help to precisely locate and verify answers. It will be used to re-rank the VSM results later.

3.2 Naïve Bayes model

The QC model is implemented with the Naïve Bayes model in the following steps:

Step 1 Select 900 data items from the VSM training set as the training set for Naïve Bayes. The VSM testing set is used for Naïve Bayes as well. A question class property is added to each data item. All the property values are manually labelled. The count of data items in each question class is shown in Tab. 3.

Tab. 3 The count of data items in each question class

Question class	Count of training data	Count of testing data
Entity	124	16
Human	0	0
Location	57	10
Numeric	67	2
Time	109	7
Definition	4	1
Description	24	3
Manner	205	18
Reason	141	17
Yes/No	169	26
Total	900	100

Step 2 Select proper features to train the QC model on the training set. The feature details are discussed later.

Step 3 Using the QC model to predict the question class for each data item in the testing set. Comparing the predicted class with a manually labelled one, the precision can be calculated.

3.3 Feature selection

In all, six features from both semantic and grammatical information in a question are adopted in this paper, for the Naïve Bayes training. All the features are listed below:

1) The question word: an interrogative pronoun such as what, how, why, when, where, etc.

2) The main verb of the question: a notional verb, usually the predicate in the sentence.

3) The dependency structure: the dependency structure between the former two features: the question word and the main verb of the question, in the form of a triplet as ⟨question word, main verb of the question, dependency⟩. The dependency is a grammar relation such as sub, obj, det, etc, which can be analyzed by the Minipar English parser.

4) The position of the main auxiliary verb: the word index of the main auxiliary verb in the sentence. The main auxiliary verb is the one that modifies the main verb of the question such as do, have, can, could, may, must, etc.

5) The main noun of the question: the first noun following the question word.

6) The top hypernym of the noun: the top hypernym of the main noun of the question, which means the top category label of the noun. According to WordNet, all nouns are divided into 25 categories, and the nouns in the same category are treated as more similar than others.

3.4 Experimental result

With the features selected above, the QC model is trained and tested. Five groups of experiments are performed, using

500 to 900 data items to train the QC model, and testing the model with the 100 data items.

The precisions of the QC model in all experiment groups are shown in Fig. 1.

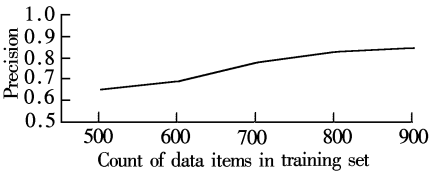

Fig. 1 Precisions of question classification

With the growth of the training sets, the precision of the QC model is improved. The best precision, from the 900 data items as a training set, is 0.85. That is good enough for the re-ranking task later.

4 Re-Rank Using Question Classification

4.1 General re-ranking model

To re-rank the QA results, a comparable *R*-value is assigned to each answer. The *R*-value should be calculated from both the original rank of certain answer and the class matching degree between the answer and the question. Then all the answers for one question are sorted by the *R*-value, and assigned a new rank.

Definition 2 (answer re-ranking model) An answer re-ranking model is a re-ranking model, merging the original answer rank and other question information, to calculate a new, more exact rank. A general re-ranking model for QA is in the following forms:

$$R\text{-value}_a = \text{merge}(\text{Rank}_a, \text{Match}_a) \tag{1}$$

$$\text{Match}_a = \text{match}(\text{QC}_a, \text{QC}_q) \tag{2}$$

Given different definitions to the merge and match functions, the *R*-value of any given answer will be different.

4.2 Re-ranking model in this paper

In this paper, the merge and match functions are defined as follows:

$$\text{merge}(\text{Rank}_a, \text{Match}_a) = \frac{1-d}{\text{Rank}_a} + d\text{Match}_a \qquad d \in [0,1] \tag{3}$$

$$\text{match}(\text{QC}_a, \text{QC}_q) = \begin{cases} 1 & \text{QC}_a = \text{QC}_q \\ 0 & \text{QC}_a \neq \text{QC}_q \end{cases} \tag{4}$$

Then the *R*-value of any given answer can now be calculated. The parameter d in Eq. (3) is a variable. Its proper value can be obtained from the following experiments.

4.3 Experimental result

Based on the re-ranking model above, the *R*-values of all answers are calculated, and a new ranking is assigned to each answer. Then the MRR result after re-ranking can be calculated, shown in Fig. 2.

It is obvious in Fig. 2 that the QA result has been improved after the QC-based re-ranking. What's more, when the parameter d is larger than 0.5, the re-ranking results reach the optimum and become steady. We take d as 0.5. It

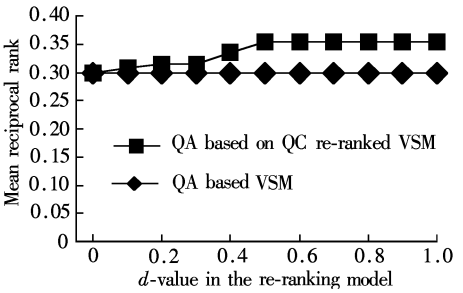

Fig. 2 Mean reciprocal ranks before and after re-rank

is proved that the re-ranking model adopted in this paper is feasible and acceptable.

5 Conclusion

A QC-based re-ranking method to improve the QA system, which is built upon real-world web data sets, is proposed. More information about question classes, semantics and grammar are drawn into both IR and ML approaches in this paper. Experiments show that the QA results can be improved by these approaches, thus providing better performance.

References

[1] Sun R, Jiang J, Tan Y F, et al. Using syntactic and semantic relation analysis in question answering [C]//*Proc of the* 14*th Text REtrieval Conference*. Gaithersburg, Maryland, 2005.

[2] Harabagiu S, Moldovan D, Clark C, et al. Employing two question answering systems in TREC 2005 [C]//*Proc of the* 14*th Text REtrieval Conference*. Gaithersburg, Maryland, 2005.

[3] Wen Xu, Zhang Yu, Liu Ting, et al. Syntactic structure parsing based Chinese question classification [J]. *Journal of Chinese Information Processing*, 2006, **20**(2): 33 – 39. (in Chinese)

[4] Lin Xudong, Peng Hong, Lin Piyuan, et al. Question interpretation and question classification based on dependency relations [J]. *Computer Science*, 2007, **34**(7): 208 – 210. (in Chinese)

[5] Lo Ka Kan, Lam Wai. Using semantic relations with world knowledge for question answering [C]//*Proc of the* 15*th Text REtrieval Conference*. Gaithersburg, Maryland, 2006.

[6] Soricut R, Brill E. Automatic question answering using the web: beyond the factoid [J]. *Information Retrieval*, 2006, **9**(2): 191 – 206.

[7] Bilotti Matthew W, Ogilvie Paul, Callan Jamie, et al. Structured retrieval for question answering [C]//*Proc of the* 30*th ACM SIGIR Conference*. Amsterdam, the Netherlands, 2007: 351 – 358.

[8] Jijkoun V, de Rijke M. Retrieving answers from frequently asked questions pages on the web [C]//*Proc of the* 14*th ACM CIKM Conference*. Bremen, Germany, 2005: 76 – 83.

[9] Li Xin, Roth Dan. Learning question classifiers [C]//*Proc of the* 19*th International Conference on Computational Linguistics*. Taipei, China, 2002: 556 – 562.

[10] Li Xin, Roth Dan, Small Kevin. The role of semantic information in learning question classifiers [C]//*Proc of the* 1*st International Joint Conference on Natural Language Processing*. Cambridge University Press, 2006: 229 – 249.

真实网络数据集自动问答系统中的问题分类

袁晓洁　于士涛　师建兴　陈秋双

（南开大学信息技术科学学院，天津 300071）

摘要：为了改善真实网络数据集上自动问答系统的性能，定义出新的问题类别集合和通用的答案重新排序模型. 问题分类器借助先验词典和语法分析，将语义和语法信息引入信息检索和机器学习方法，呈现为多种多样的训练属性，包括疑问词、中心动词、疑问词与中心动词依赖关系、中心助动词位置、中心名词、中心名词顶级上位词等. 进而通过问题类别信息，对问答查询结果重新排序. 实验表明：分类器能够精确实现真实网络数据集的问题分类，重新排序后的自动问答结果也能得到明显改善. 这说明借助语义和语法信息，真实网络数据集上的自动问答系统等应用可以得到改善，显示出更好的性能.

关键词：问题分类；自动问答系统；真实网络数据集；问答网络论坛；重新排序模型

中图分类号：TP391