Pattern Recognition Letters 31 (2010) 1935–1943

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A semantic approach for question classification using WordNet and Wikipedia

Santosh Kumar Ray [a,*], Shailendra Singh [b], B.P. Joshi [c]

[a] Department of Computer Science, Birla Institute of Technology, Mesra, International Centre, Muscat, Oman
[b] Samsung India Software Center, Noida, India
[c] Birla Institute of Technology, Noida, India

ARTICLE INFO

Article history:
Received 27 June 2009
Available online 1 July 2010
Communicated by C.L. Tan

Keywords:
Question Answering System
Question classification
Answer validation
WordNet
Google
Wikipedia

ABSTRACT

Question Answering Systems, unlike search engines, are providing answers to the users' questions in succinct form which requires the prior knowledge of the expectation of the user. Question classification module of a Question Answering System plays a very important role in determining the expectations of the user. In the literature, incorrect question classification has been cited as one of the major factors for the poor performance of the Question Answering Systems and this emphasizes on the importance of question classification module designing. In this article, we have proposed a question classification method that exploits the powerful semantic features of the WordNet and the vast knowledge repository of the Wikipedia to describe informative terms explicitly. We have trained our system over a standard set of 5500 questions (by UIUC) and then tested it over five TREC question collections. We have compared our results with some standard results reported in the literature and observed a significant improvement in the accuracy of question classification. The question classification accuracy suggests the effectiveness of the method which is promising in the field of open-domain question classification.

Judging the correctness of the answer is an important issue in the field of question answering. In this article, we are extending question classification as one of the heuristics for answer validation. We are proposing a World Wide Web based solution for answer validation where answers returned by open-domain Question Answering Systems can be validated using online resources such as Wikipedia and Google. We have applied several heuristics for answer validation task and tested them against some popular web based open-domain Question Answering Systems over a collection of 500 questions collected from standard sources such as TREC, the Worldbook, and the Worldfactbook. The proposed method seems to be promising for automatic answer validation task.

1. Introduction

Today, the World Wide Web has become the major source of information for everyone, from a general user to the researchers for fulfilling their information needs. Virtually all kinds of information are available on the World Wide Web in one or another form. Recently published article (Alpert and Hajaj, 2008) says that number of web pages on the Internet increased tremendously and crossed 1 trillion landmark in 2008 which was only 200 billion in 2006 as reported in (Wirken, 2006). Therefore, managing such a huge volume of data is not an easy task. Search engines like Google and Yahoo return links (along with snippets) to the documents for the user query. Most often, web pages retrieved by these search engines do not provide precise information and may contain irrelevant information in even top ranked results. This prompts researchers to look for an alternate information retrieval system that can provide answers of the user queries in succinct form. Question Answering System is one of such prominent information retrieval system that is getting popular among all type of users ranging from occasional surfers to specialist information seekers. Question Answering Systems, unlike other information retrieval systems, combine information retrieval, and information extraction techniques to present precise answers to user questions posed in a natural language.

A typical pipeline Question Answering System consists of three distinct phases: Question Processing, Document Processing, and Answer Processing. Question Processing phase classifies user questions (also termed as question classification), derives expected answer types, extracts keywords, and reformulates a question into semantically equivalent multiple questions. Reformulation of a query into similar meaning queries is also known as query expansion and it boosts up the recall of the information retrieval system. The Document Processing phase retrieves documents containing keywords in the original as well as expanded questions, applies ranking algorithms on the retrieved document set and returns the top ranked documents. In Answer Processing phase, the system

* Corresponding author. Tel.: +968 95716749; fax: +968 24449197.
E-mail addresses: santosh@waljatcolleges.edu.om, sapin2@gmail.com (S.K. Ray), shailendra.s@samsung.com (S. Singh), bp_joshi@yahoo.com (B.P. Joshi).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.06.012

identifies the candidate answer sentences, validates the correctness of the answers, ranks them and finally presents the answers to the user using information extraction techniques.

A Question Answering System returns answer of a user question in succinct form. In order to provide the precise answer, the system must know what exactly a user wants. The prior knowledge of the expected answer type helps the Question Answering System to extract correct and precise answers from the document collection. The question classification techniques help a Question Answering System to meet this requirement. This is the reason that almost all Question Answering Systems include question classification module. The question classification module parses the user question to derive the expected answer type. The error analysis of an open-domain Question Answering System by Moldovan et al. (2003) showed that 36.4% of the errors were generated due to incorrect question classification. This creates the need of an efficient method for question classification.

As discussed before, a Question Answering System identifies the candidate answer sentences, performs validation of the correctness of the answers, ranks them and finally presents the answers to the user using information extraction techniques. This is very easy to realize that providing incorrect answer is worse than providing no answer. To emphasize the importance of answer validation, TREC (Text REtrieval Conference) has introduced the notion of confidence since its 2002 edition. Output of the question classification phase can play vital role in the validation of the candidate answers identified by the system. If the candidate answer sentence contains the entity predicted by question classification phase in it, then it is more likely to be a valid answer. Therefore, in Section 5 of this paper, we have proposed an answer validation method that combines output of question classification process with some other heuristics to validate candidate answer sentences. The proposed method also exploits the fact that correct answers have higher frequency of occurrence on the World Wide Web as compared to incorrect answers.

The rest of the paper is organized as follows: Section 2 provides the details of previous work in the field of question classification. In Section 3 we have given detailed description of the proposed method along with multiple illustrations. Section 4 presents the results of the experiments conducted over the standard sets of question collections. Section 5 describes a unique application of question classification in the form of answer validation where we have proposed a World Wide Web based answer validation method. Finally, we have concluded our work in Section 6.

2. Related work

There are two main approaches for question classification: manual and automatic. Question Answering Systems using manual classifications (Hermjakob, 2001) apply hand-crafted rules to identify expected answer types. These rules may be very accurate but these are time consuming, tedious, and non-extendible in nature. The automatic classifications, on the other hand, are extendible to new question types and classify questions with a reasonably good accuracy. Automatic question classification is further divided into two main approaches known as machine learning and language modeling.

The primary machine learning algorithm used for question classification is Support Vector Machine (Hacioglu and Ward, 2003). Zhang and Lee (2003) employed tree kernel with a SVM classifier for question classification and reported 80.2% accuracy without the use of syntactic or semantic features. Li and Roth (2002) reported a hierarchical approach for question classification based on the SNoW learning architecture. They used a two stage classification process. The first stage returned the five most probable coarse-grained question types, and then the second stage classified the question into one of the child classes of the five coarse-grained question types with accuracy of 84.2%.

Language modeling based question classification approaches (Pinto et al., 2002) try to compute the probability of the question for a given question class. A language model is constructed for each question class, built up from all the questions from that class. Given this question class language model and a question, it is proved that the question was generated using the given language model.

In (Nguyen et al., 2007), a subtree mining method has been used for syntactic and semantic parsing problems for question classification using maximum entropy. Solorio et al. (2004) proposed an approach to question classification with only surface text and simple retrieval results from Google search engine. David et al. (2006) proposed an automatic feature extraction approach to question classification, which uses only statistical information from unlabeled corpus to extract features without the help of natural language processing techniques.

Use of online semantic resources such as WordNet for question answering has been reported in literature (Li and Roth, 2006). The experiments with WordNet indicate that use of semantic information for question classification greatly improves the performance of Question Answering Systems. All of the works discussed above reported very high precision for question classification. However, we are not aware of any research that uses online collaborative resources such as Wikipedia (Wikipedia) for question classification problem. In this paper, we are proposing a question classification method that efficiently exploits the features of both WordNet and Wikipedia. Use of collaborative systems such as Wikipedia can help the Question Answering System to predict the correct entity type especially when compound NPs (such as movie names) are involved. In the next section, we will provide details of the proposed question classification method.

3. Question classification algorithm

In this section, we are explaining the question patterns found in the experimental questions database collected from UIUC (Hovy et al., 2001). These question patterns are crucial to identify the nature of the questions. Later, we are proposing question classification algorithm to classify questions using WordNet and Wikipedia.

3.1. Question database collection

We have used the question collection set compiled by Li and Roth (2002) for training of our question classification system. The question database consists of 5500 training and 500 test questions collected from english questions published by USC (Hovy et al., 2001), two Text REtrieval Conferences (Text REtrieval Conference) namely TREC 8, and TREC 9 questions, and about 500 manually constructed questions for a few rare classes. The test questions have been collected from TREC 10 question dataset. All questions of the dataset have been manually labeled by Li and Roth according to the coarse and fine grained categories shown in Table 1.

3.2. Identification of question patterns

We carefully surveyed the collection of questions and identified eight main patterns: seven for standard 7-Wh questions and eighth for other questions. Each of these patterns further consists of several sub-patterns. The descriptions of these patterns are given below.