Article can be accessed online at http://www.publishingindia.com

Improving the Performance of RDQA Using Lexical Based Inference Extraction

Renita Raymond*, Karnavel Kuppusamy**

Abstract

This paper presents an enhanced approach for Question Classification and Answer Extraction in Restricted Domain Question Answering (RDQA). Question Classification and Answer Extraction is the core problem of RDQA and determines the performance of the Question Answering in the Restricted Domain. The proposed approach improves the performance of RDQA by means of (1) Question type prediction model based on Bayesian classification (2) Lexicalized-Index based Passage Retrieval (3) Lexical-Semantic based Inference Extraction. This paper also describes user-centered task-based evaluations for Answer Validation. Further improvements are achieved by combining our model with the classic one to improve the performance of Restricted Domain Question Answering.

Keyword: Restricted Domain Question Answering (RDQA), Bayesian Classification, Passage Retrieval, Answer Extraction, Text Inference

Introduction

The question answering system is one of the emerging and useful areas of research in natural language processing applications of Artificial Intelligence and Information Retrieval,because there is a need to move from the traditional document retrieval to actual information retrieval by providing an answer to a question rather than a ranked list of relevant documents.

Nowadays the availability of comprehensive and reliable resources in complex domains enables interesting and fruitful research to be carried out in restricted-domain natural language processing. In short, research in restricted-domain question answering (RDQA) addresses problems related to the incorporation of domain-specific information into current state-of-the-art QA technology with the hope of achieving deep reasoning capabilities and reliable accuracy performance in real world applications

The core problems of Question answering are Question Classification and Answer Extraction. These two components mainly have the direct impact on the performance of RDQA. Recently, many RDQA systems have been developed by incorporating syntactic or semantic elements in the model. Also some systems are developed mainly to address Answer extraction problem. Our work is not novel, but the combination is such that it has not been studied before.

In our approach, the performance of RDAQ is improved by means of specialized index based retrieval. First the collected documents are preprocessed to incorporate lexical and sematic features. Then we build the lexicalized index for the passages in the documents. And also the answer type is predicted for the question from set of predefined classes by means of Bayesian classification. Finally, the inferences are created from the retrieved passages using lexical and semantic features to obtain the correct answer.

In this paper, we present the results of our work on the development of a Restricted Domain Question Answering.

* Computer Science and Engineering Department, Anand institute of Higher Technology, Chennai, Tamil Nadu, India.
E-mail: reniamala18@gmail.com
** Computer Science and Engineering Department, Anand institute of Higher Technology, Chennai, Tamil Nadu, India.
E-mail: treseofkarnavel@gmail.com

The rest of the paper is organized as follows. The second section provides an overview of previous work done in RDQA. The third section describes the baseline Question Answering in Restricted Domain. The proposed methodology is presented in the next section. The final section describes experimental results and discussions, followed by the conclusions.

Related Work

Recently, restricted-domain QA regained attention, as shown by a dedicated ACL workshop to be held in 2004 (In Proceedings ACL 2004 Workshop on Question Answering in Restricted Domains) and the AAAI-05 Workshop to be held in 2005 (The AAAI-05 Workshop on Question Answering in Restricted Domains).

The restricted-domain systems of today are different from the toy systems from the early years of QA (Voorhees & Tice, 2000), which might be what first comes to mind when reading the term "restricted-domain". Benamara (2004) showed an experiment to designing a logic based QA system (WEBCOOP) for the tourism domain, that integrates knowledge representation and advanced reasoning procedures to generate cooperative responses to natural language queries on the web. Nguyen & Kosseim (2004) mentioned a restricted domain QA system using semantic information, which consists in finding a set of special terms and building a concept hierarchy which can effectively characterize the relevance of a retrieved candidate to its corresponding question, and it aims at replying questions on services offered by a large company.

Niu & Hirst (2004) provided an QA research in clinical-evidence texts, which identifies occurrences of the semantic classes (disease, medication, patient outcome) and the relations between them, so that determine whether an outcome is positive or negative. Chunget al. (2004) described a research work for QA by restricting the question domains and extract. These early systems encoded large amounts of domain knowledge in databases. The restricted-domain systems of today are far less dependent on large knowledge bases and do not aim for language understanding per se. Rather, they use specialized extraction rules on a domain specific collection.

Overview of RDQA

The common levels that are used by different Restricted Domain Question Answering systems architectures are as follows:

Question Classification

This level provides correct answers by classifying the user query into one of the question type to which it belongs to. This module also processes the question, analyzes the question type, and produces a set of keywords for retrieval. Depending on the retrieval and answer extraction strategies, some question analysis module also performs syntactic and semantic analysis.

Answer Extraction

This level extracts the correct plausible answers for different classification of questions, given the top relevant passages from the retrieval module. The answer extraction module performs detailed analysis and pin-points the answer to the question.

Answer Selection

Among the plausible answers obtained, ranking approaches are used to mine the best accurate answers based on its weightage factor.

Each module can contain the sub modules as needed to improve the performance of RDQA.

Proposed Methodology

In our approach, we extend the basic standard Restricted Domain Question Answering System. This can be done by incorporating following model which consists of:

1. Bayesian classification-based Question Type Prediction
2. Lexicalized Index-based Passage Retrieval
3. Lexical-Semantic-based Inference Extraction

Bayesian Classification-Based Question Type Prediction Model

One of the core structures of a Restricted Domain Question Answering System is Question Classification module. However, things are different when we deal with the question type prediction in Question Classification from set of pre-defined Classes leads to Classification Problem.

Approaches to question classification can be divided in two broad classes, namely, rule-based and machine learning methods. Most recent studies have been based on machine learning approaches. Li and Roth proposed 6 coarse classes and 50 fine classes for TREC factoid question answering. The UIUC QC dataset, which they developed, contains 5,500 training questions and 500 test Questions. Krishnan *et al*. used SVM with question bi-grams.

Naïve Bayesian Classification Model

This section describes a model for binary classification, Naive Bayes. Naive Bayes is a simple but important probabilistic model. In this model,after analyzing certain features F_i, the question type class C_i can be assigned to the question Q.

$$\text{Classify } (f_1, f_2\ldots) = \arg\max p(C = c)\prod_{i=1}^{n} p(Fi = fi|C = c) \quad (1)$$

This model includes feature selection and needs certain preprocessing to build a classifier.

Feature Selection

Feature selection is one of the most important preprocessing steps in Classification. It is an effective dimensionality reduction technique to remove noise feature. In general, the basic idea of feature selection algorithm to searches through all possible combinations of attributes in the data to find which sub-set of features works best for prediction. Thus, the attribute vectors can be reduced in number by which the most meaningful ones are kept and the irrelevant or redundant ones are removed and deleted. In our approach we consider features based the question type, Context and Lexical features.

Lexicalized Index-Based Passage Retrieval

The recent major approach we examine is the Index based retrieval approach for answer extraction. We can build our own type of index to represent the documents we collected.

A method for retrieving information from a document includes a process of grouping paragraphs in the document to form passages, and forming indexes relating to a number of words in the passages. The number of paragraphs in a passage is determined based on the number of paragraphs considered optimum for a writer to cover a particular topic. Passages are formed by merging each N consecutive paragraph in the document, where N is an integer greater than 1. Thus, individual passages may include paragraphs that are identical to other passages.

In our approach we preprocessed the document collection to incorporate lexical and sematic features. This can be done using POS Tagging, Named Entity Tagging and

Syntactic parsing. After the preprocessing we developed the lexicalized index for the passages.

At first, the method transforms passages into xml documents. Then the method scans the xml documents to create and maintain an inverted index with the tool of Lucene. In the last, the method searches the user's question and its extended questions in the inverted index database, and returns the passages. Because using B-trees to maintain the index, Lucene has well-behaved I/O characteristics (lookups and insertions are O(log n) operations). Therefore, the index-based method speeds up searching question-answering pairs.

Lexical and Semantic-Based Inference Extraction

An inference approach to QA is one in which there is given a question and a passage that contains the answer. For example, the following question

e.g What type of animal is Winnie the Pooh? and the answer passage is,

A Canadian town that claims to be the birthplace of Winnie the Pooh wants to erect a giant statue of the famous bear; but Walt Disney Studios will not permit it.

It is clear that there is a linkage between the question word animal and the answer word bear. That the word bear occurred in the answer, in the context of Winnie, means that there was a hidden "cause" for the occurrence of bear, and that was the concept of animal. In general, there could be multiple words in the question and answer that are connected by many hidden causes.

There has been a growing amount of work which focuses on inferring semantic relatedness of concepts from different information embedded in Wikipedia. Strube and Ponzetto (2006) infer semantic relatedness from the concept hierarchies in Wikipedia, but it cannot give adequately accurate results. ESA (Gabrilovich and Markovitch, 2007) infers semantic relatedness based on the textual similarity among Wikipedia articles, although achieving extremely accurate results.

Experimental Results and Discussions

Dataset Used for Experiments

Developing a set of test questions was easier said than done. Unlike the open domain evaluations, where test questions can be mined from question logs (Encarta, Excite, Ask Jeeves), no question sets are at the disposal of restricted domain evaluators. We manually developed a question test set for our experiments in computer topics domain. Based on the examples of TREC questions, also we developed the questions.

Evaluation Metrics

The quality of the response (generated by question answering systems) is evaluated with three metrics, i.e., Recall, Precision and F score. We define these metrics formally as follows. For query q, Ans = a_i} represents its answer collection where each answer i a is associated with an importance value $imp(a_i)$, and m denotes the number of answers contained by the response.

Recall measures how much the response covers the answers and is defined as

$$\text{Re } call = \frac{\sum_{ai \in response} imp(ai)}{\sum_{aj \in Ans} imp(aj)} \quad (2)$$

In Sub Encarta, the answers are given in the form of entire sentences. Therefore, Precision for this dataset is defined as

$$\text{Precision} = \frac{m}{M} \quad (3)$$

Each answer is given as a nugget which consists of several words. In this case, we adopt the Precision definition of TAC 2008 [5],

$$\text{Precision} = 1 - ((l - A)/A) \quad (4)$$

where l is the number of non-white space characters in the response, A is the *character allowance* for the response and $A = C.m$(C is the character allowance for each nugget and is set to100.

Fscore is the combination of Recall and Precision. In TAC 2008, Fscore is calculated with the official value $\beta = 3$, which means Recall is three times as important as Precision

$$\text{Fscore} = (\beta 2 + 1) \cdot \text{Re call} \cdot \text{Precision} \cdot \beta$$

$$= 2 \cdot (\beta 2 \cdot \text{Precision} + \text{Re call} \quad (5)$$

The overall performance of question answering on a whole dataset is evaluated by averaging the metric values of each question in it.

Performance Evaluation

We have implemented the standard model my making some modifications and gives the result which is not appreciable. We have also implemented our proposed system by combining it with the standard one. The performance of our system is nearly 10% better than the standard one, and our results shows the proof for that.

So it clearly indicates that modification of Question Classification and Answer Extraction Module will improve the overall performance of the RDQA systems.

Conclusion and Future Work

In this paper we presented an enhanced approach for Restricted Domain Question Answering. And in our approach the performance of RDAQ is improved by means of Question Type prediction model based on Bayesian classification, Lexicalized-Index-based Passage Retrieval, Lexical-Semantic-based Inference Extraction.

Our approach also shows some improvements over the standard one and indicates that specialized index-based retrieval plays important role in RDQA Systems.

There are still some problems occuring in the RDQA that need to be discussed. In our approach we have solved certain problems but some problems like domain coverage and accuracy of the answers still remain unquestioned. In future we have planned to address these issues. And also the inference model can only deal with the relations between two single sentences. For an inferable relation across the sentence, a transformation-based strategy or a multi-sentence word/event chain method should be considered.

References

Benamara, F. (2004). *Cooperative question answering in restricted domain : the WEBCOOP experiment*, In Proceedings ACL 2004 Workshop on Question Answering in Restricted Domains.

Chung, H., Song, Y., Han, K., Yoon, D., Lee, J., Rim, H. & Kim, S.(2004). *A Practical QA System in Restricted Domains*, In Proceedings ACL 2004Workshop onQuestion Answering in Restricted Domain.

Gabrilovich, E. & Markovitch, S. (2007). *Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis*. In Proceedings of the 20th International Joint Conference on Artificial Intelligence.(IJCAI'07), (pp. 1606-1611).

Strube, M. & Ponzetto, S. P. (2006). *Wiki Relate! Computing Semantic Relatedness using Wikipedia*. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (pp. 1419-1424).

Nguyen, H. D. & Kosseim, L. (2004). *Using Semantic Information to Improve the Performance of a Restricted-Domain Question-Answering System*, In Proceedings ofthe Question-Answering workshop of TALN (TraitementAutomatiquedela Langue Naturelle), Fes, Maroc, pp.441-450.

Niu, Y. & Hirst, G. (2004). *Analysis of Semantic Classes in Medical Text for Question Answering*, In ProceedingsACL 2004 Workshop on Question Answering in Restricted Domains.

Voorhees, E. M. & Tice, D. M. (2000). *Implementing a Question Answering Evaluation*. In Proceedings of LREC' 2000 Workshop on Using Evaluation within HLT Programs: Results and Trends. (pp. 130).

Liang, X., Wang, D. & Huang, M. (2010). "*Improved Sentence Similarity Algorithm based on VSM and its application in Question Answering System*," In Proceedings of the Intelligent Computing and Intelligent Systems (ICIS), Japan.

Diekema, A. R., Yilmazel, O., Chen, J., Harwell, S., He, L. & Liddy, E. D. Finding Answers to Complex Questions. To appaer. In Maybury, M.(Ed.) New *Directions in Question Answering*. AAAI-MIT Press.

Gaizauskas, R. & Humphreys, K. (1998). *A Combined IR/NLP Approach to Question Answering Against Large Text Collections*. University of Sheffield UK.

Li, X. & Roth, D. (2002). *Learning Question Classifiers*. In Proceedings of the 19th International Conference on Computational linguistics (pp. 1-7).