Design of the Effective Question Answering System by Performing Question Analysis using the Classifier

Gayatri Chavan
Department of Computer Engineering
Pimpri ChinchwadCollege of Engineering
Pune-411044

Sonal Gore
Department of Computer Engineering
Pimpri ChinchwadCollege of Engineering
Pune-411044

ABSTRACT

Search engines have played a very important role in helping the users to search the necessary information from the huge information. By displaying the list of links to documents.The Question-Answering systems are gaining popularity. Because The main benefit of such QA systems is that the user can ask the query (question) in natural language and he /she get a precise and appropriate answer instead of just displaying a list of links to documents.

The main advantage of the proposed Question answering system, which is not restricted to a specific domain. This approach is related to a natural language interface to the database (NLIDB), which takes a natural language query as input and giving the appropriate answer from the manually created knowledge base(structured database). There are two main steps of implementation of the proposed question answering system. The first step is to use a classifier to identify appropriate tables and columns in a structured database for an incoming question, and the second step is to perform the free text retrieval to lookup answer. The system uses named entity normalization, part-of-speech tagging, and a statistical classifier trained on data from the TREC QA task.

Keywords

Part of speech tagging, named entity normalization, statistical classifier, TREC QA data.

1. INTRODUCTION

Automatic Question Answering (QA) systems return Answers not displaying a link to a list of documents—in response to a user’s query. One special type of QA systems extensively studied during the 1970s–1990s comprised Natural Language Interfaces to Databases (NLIDB), which used to structure databases (DBs) as the information source and aimed at hiding complex database query languages from the user; see [5] for an overview. Find the answer to a question in a large collection of documents questions (in place of keyword-based query) answers (in place of documents).

The basic architecture of many open domain textual QA systems consists of a three-stage pipeline: question analysis, document retrieval, and answer extraction [10] is shown in fig 1. There are several approaches to implementing these three stages of the basic architecture of the QA system. The main focus of a proposed QA system in question analysis phase, where information extraction (answer) done from structured database.

The structured database is used as a knowledge base for this QA system. The knowledge base contains facts corresponding to commonly occurring question types are extracted and stored in a structured database format for lookup at question time. The main advantage is to avoid the expensive text analysis, and system can achieve good run time behavior because the information extraction at the table creation phase only.

The question analysis stage, i. e lookup stage of this approach, where facts corresponding to commonly occurring question types are extracted and stored in a database for lookup at question time. For this approach, use a set of databases in the form of table those are manually created as the source of answers. Then map an incoming question to a knowledge base query consisting of several tables. The SQL like query,” Select Answer Factor from table tablename where similarities between Question Factor and Question”, where table that contains the answer candidates in the field answer factor and its other field question factor has a high similarity with the input user question. In this query formalism, the main task is mapping an incoming question to a tuple < tablename, question factor, answer factor> (a table lookup label) and efficiently implementing the similarity function in question and question factor. The generation of table lookup labels as a classification task and apply a standard machine learning approach to it.

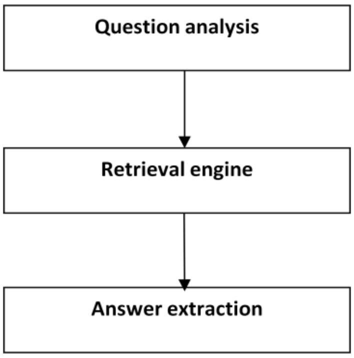

Fig. 1

2. RELATED WORK

Many question answering system uses a machine learning approach by extending the canonical pipeline architecture in different ways . This paper provides an implementation approaches for various categories of question answering system such as Closed domain based QAS, Open domain based QAS, BBASED QAS, Information Retrieval or Information Extraction(IR/IE) based QAS, and rule based QAS which will be helpful for new directions for work in this area [2]. E.g. [13] describes a QA system where passages identified by an information retrieval engine are re-ranked by a machine learning component trained on a corpus of questions and answers to classify passages for “answerhood.” Machine learning is also applied to question classification, often understood as identification of the expected answer type for a Question. E.g., [14] represents a the semantic and syntactic features of question are understand and then applies the a supervised SVM classifier to performing the question classification task which improve the system performance; [15] presents a cascaded classifier and describes a training corpus of 5,500 questions manually annotated with expected answer types. There has also been much research in defining the similarity or relevance functions for short text segments (e.g.,Questions and answer sentences). E.g., [16] presents an in depth Exploration of the TREC Novelty task [17], i.e., identifying relevant and novel sentences in a ranked list of relevant documents. In [6], describes the comparism for many different query expansion methods; [17] uses WordNet to disambiguate between different word senses in order to assign an appropriate sense too Each query term for query expansion.The group of words are gathered under one sense in wordnet. They described an open domain QA system that answers to natural language questions using tabular data that has been automatically extracted from a newspaper collection by using Information Extraction tool[1].

3. PROPOSED SYSTEM

The following are the main module of the poaposed Question Answering system.

1. Retrieval module
   - Text Representation
   - Retrieval query formulation
2. Classifier module
   - Table lookup label

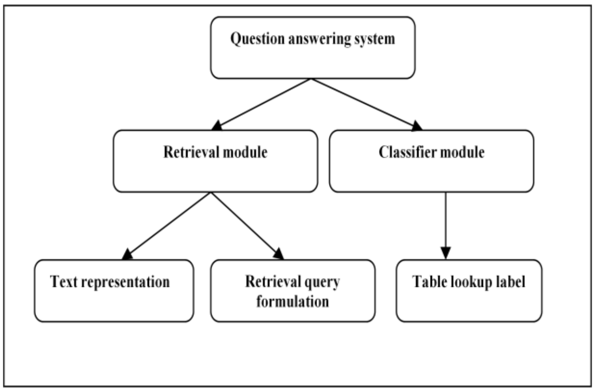

Fig. 2

The architecture depends upon two modules in fig 2: The classifier that predicts table lookup label and the retrieval model along with the text representation and the retrieval query formulation.

For proposed question answering system, use the standard set of question/answer pairs from the TREC QA tasks of 2001-2003 and a knowledge base with tables created manually referring the question answer pairs of TREC dataset. The tables contain the information such as question factor, answer factor stored in a tabular format. For a source of answers, the QA system uses a set of databases (tables). These tables contain text, for example, the Roles (role, name) table contains the role of George Bush as United States President. This system does not have any prior knowledge of the semantics of the table and field names, views the content of the database as a black box.

The input to the proposed QA system is user question which is in natural language. When a user post a question, the first to extract the feature of the question usin POs tagger. The POs tagger assigns the part of speech tag to given question by using statistical part of speech tagger TNT [7]. Then named entity tagger based on TNT is used to generate a question feature which in vector format in feature extraction step. On the other side, in the retrieval module use the vector space model to get the relevant information for the given question by using the knowledge base. Then we form the retrieval query by translating the user question which is run against an index that contains the values of all fields of all rows in our database as separate documents. The vector space model to generate training data for the classifier. For the retrieval method, as a collection of questions with known correct answers. There are various several choices for selecting retrieval method those are first how to represent a text of documents, i.e. field values, second query formulation for the retrieval and third which text retrieval model to use. Here we consider the standard stemming and named entity normalization. Retrieval using vector space modeling with standard stemming and normalization.

The text representation, in which provide the abbreviations for some words of the knowledge base for better performance. For example E.g., “U.S.” and “United States” belong to the same WordNet synset and thus would become identical after normalization. For this use a WordNet synsets2 as canonical forms of NEs. After that find the similarity between question factor and question using retrieval models to generate a ranked list of field values from our database. Which is input to the statistical classifier. Then apply a statistical classifier that assigns a table-lookup labels, i.e. a tuple < tablename, question factor, answer factor>. The k-NN is a type of instance based learning or lazy learning, This algorithm is among simplest of all machine learning algorithm.

Algorithm: KNN

Input :

- A test question Q in < T, QF> and the associated Table rows to be ranked.
- Training data <Ti, Qfi, Afi> Where i= 1,2,3,...n
- Number of nearest neighbors k.

Output:

Candidate answer for Question Q.

- For training:

(1) For each training Question Q, use Vector space model to retrieve its top tables and compute features <T, QF,AF>.

(2) For each training Question Q, find k nearest neighbors Of Q

- For testing question:

(1) Generate the question feature < T, QF>

(2) Find k nearest neighbors of Q, in the training data in the <T, QF, AF> feature space.

(3) Find the most similar training

(4) Generate the candidate answers.

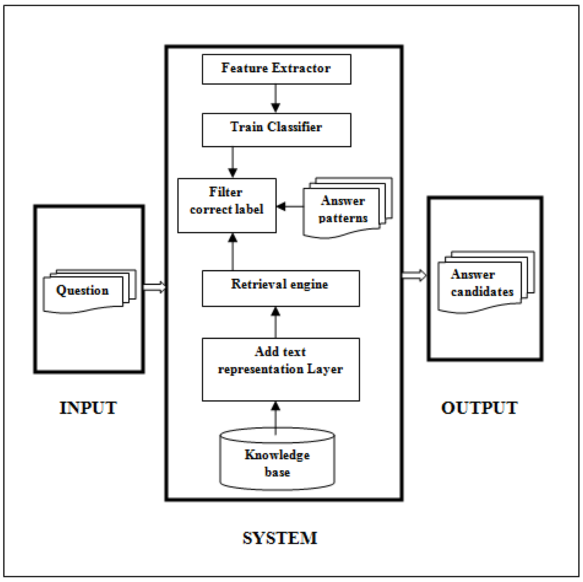

Fig. 3

4. CONCLUSION

Design an effective Question Answering system for open domain that answers the natural language question using tabular data. The main advantage of this approach is hat by moving the expensive text analysis and information Extraction stages offline (to the table creation phase), systems Can achieve good run-time behavior, even with a large Amount of information. The system retrieves as answers the content of the field answer factor of the row of the table of which the similarity between question factor and question is the highest by mapping an incoming question to a table lookup label. This question answering system treats the data (Knowledge base) as a black box; therefore it does not require any semantic information of the knowledge base.

5. REFERENCES

[1] Mahboob Alam Khalid, Valentin Jijkoun, Maarten de Rijke, “Machine Learning for Question Answering from Tabular Data”, 18th International Workshop on Database and Expert Systems Applications, 2007 IEEE.

[2] Ms. Pooja P. Walke, Mr. Shivkumar Karale, “Implementation Approaches for Various Categories of Question Answering System”, Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013), 2013 IEEE.

[3] Rana Forsati, Mehrnoush Shamsfard, Pouyan Mojtahedpour, “An Efficient Meta Heuristic Algorithm for POS-Tagging”, 2010 Fifth International Multi-conference on Computing in the Global Information Technology, 2010 IEEE.

[4] Varsha Bhoir, M. A. Potey, “Question Answering System : A Heuristic Approach”, 978-1-4799-2259-14/$31.00©2014.

[5] Muthukrsihanan Ramprasath and Shanmugasundaram Hariharan, “Improving QA performance through semantic reformulation”, 2012 Nirma University International Conference on Engineering, NUiCONE-2012, 06-08DECEMBER, 2012 IEEE.

[6] V. Jijkoun, M. de Rijke, and J. Mur. Information extraction for question answering: Improving recall through syntactic patterns. In Proc. COLING 2004, 2004.

[7] T. Brants. TnT – A Statistical Part-Of-Speech tagger. In Proc. of the 6th Applied NLP Conference, 2000.

[8] D. Ahn, V. Jijkoun, K. M¨uller, M. de Rijke, and E. Tjong Kim Sang. Towards an offline XML-based strategy for answering questions. In Accessing Multilingual Information Repositories, pages 449–456, 2006.

[9] M. Fleischman, E. Hovy, and A. Echihabi. Offline strategies for online question answering: answering questions before they are asked. In Proc. ACL ’03, pages 1–7, 2003.

[10] D. Harman. Overview of the TREC 2002 novelty track. In Proc. TREC 2002, pages 17–28, 2002.

[11] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In Proc. ECDL 1998, pages 569–584, 1998.

[12] E. Voorhees and D. Tice. The TREC-8 question answering track evaluation. In Proc. TREC-8, 1999.

[13] G. Ramakrishnan, S. Chakrabarti, D. Paranjpe, and P. Bhattacharyya.Is question answering an acquired skill? In Proc.WWW, pages 111–120, 2004.

[14] D. Metzler and W. Croft. Analysis of statistical question classification for fact-based questions. Journal of Information Retrieval, 8:481–504, 2005.

[15] X. Li and D. Roth. Learning question classifiers. In Proc. COLING 2002, 2002.

[16] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In Proc. SIGIR 2003, pages 314–321, 2003

[17] D. Harman. Overview of the TREC 2002 novelty track. In Proc. TREC 2002, pages 17–28, 2002