Encyclopedia of Business Analytics and Optimization

John Wang
Montclair State University, USA

Volume IV
Question Answering Systems for Managing Big Data

Sparsh Mittal
Iowa State University, USA

INTRODUCTION

Recent advancements in the fields of biomedicine and bioinformatics have resulted in exponential growth in the amount of data (Agrawal et al., 2008; Mittal, 2014). For example, the PUBMED database offers more than 18 million articles and hundreds of thousands more are being added every year (Howe et al., 2008). However, such vast amount of data is meaningful only when effective techniques for retrieving the data are available. Modern search engines such as Google and Bing enable users to search the Web; however the search engines return “documents” and “answers” and hence, they require the users to manually search the vast number of documents to obtain the desired answer. Further, search engines retrieve documents based on “keywords,” but ignore the structure and intent of the query. For example, the following three questions, “how is snake poison employed in counteracting neurotoxic venom?,” “when is snake poison employed in counteracting neurotoxic venom?” and “why is snake poison employed in counteracting neurotoxic venom?” all have different meanings. However, search engines typically ignore the WH-words and hence, they cannot differentiate between these questions.

To address these limitations, question answering systems (QASs) are of vital importance (Cao et al. 2011, Cairns et al. 2011, Toba et al. 2014). QASs use information retrieval (IR) and natural language processing (NLP) techniques to answer the questions posed by humans in the natural language. The examples of existing QAS include START (Katz, 1997), AskMSR (Bril et al., 2002), EureQA (Gupta et al., 2008) and BioinQA (Mittal et al., 2008a, 2008b).

QASs have several important applications. In medical domain, QASs are extremely important for improving the health care by assisting the physicians in gaining latest information on the field. They quickly answer the questions that arise during their meetings with patients. In e-learning, QASs are important in assisting novice learners. In this chapter, we discuss the working of QASs and also discuss recent trends in development of QASs.

BACKGROUND

Zweigenbaum (2003) discuss the role and importance of QASs in biomedicine. In the literature, several techniques have been proposed for answering biomedical questions, such as answering by role identification (Niu et al., 2003) and document structure (Sang et al., 2005). In a study conducted with a test set of 100 medical questions collected from medical students, a thorough search in Google failed to obtain relevant documents within top five hits for 40% of the questions (Jacquemart & Zweigenbaum, 2003). Moreover, due to need of answering the question swiftly and the busy practice schedules, doctors spend less than two minutes on average for searching an answer to a question. Hence, search engines fail to fully answer most of the clinical questions (Ely et al., 1999). These research studies further confirm the importance of question answering systems.

Some QASs are closed-domain, which implies that they deal with a specific domain or accept only a restrict kinds of questions. Other QASs are open-domain which can answer multiple kinds of questions from different fields. An example of
closed-domain QAS is biomedical QAS. Sondhi et al. (2007) discuss a biomedical QAS named Internet doctor (INDOC). Their system works by indexing the entire document set. The system processes the user-question to recognize the difference in significance of different parts of the query. The answers are ranked by measuring the relevance of the documents to the query.

**MAIN FOCUS**

**Working of a QAS**

The Question Answering system is based on searching the entities of the corpus, for effective extraction of answers. The system recognizes the keywords of the corpus material using Link parser. This is especially useful in technical domains (e.g. biomedical domain) where extended terms (e.g. nucleocapsid, immunoglobulin, ultrasonography, etc.) of the lexicon are classified as entities. The question is parsed during Question analysis step. The question is then translated into a set of queries which are used to access the corpus. In a QAS, a grammar parser decides the syntactic structure of the question and also extracts part of speech information. Question classifier then uses pattern matching based on wh-words (such as when – refers to an event, why – reasoning type, etc.) and simple part-of-speech information to determine question types.

Then, question focus is identified by finding the object of the verb. More importance is given to the question focus. The contribution of each occurrence of each query term is summed to obtain a similarity score for a specific location in any document. After phrase matching, system processes the passages according to the classification done in question classification. Based on the similarity score, the returned documents are re-ranked and finally, they are presented to the user.

Figure 1 shows the block diagram of a QAS. A QAS first parses the question to understand its intent and find the query terms. It then searches the corpus (dataset) for selected passages or sentences for effective extraction/construction of answers. The corpus could be either unstructured (such as the Web), semi structured (such as WordNet or the CIA World Fact Book), or structured (such as geography databases). The object of the verb of the question generally determines its focus. Based on the occurrence of different query terms at any particular location in the document, its relevance is determined. In this manner, multiple answer passages are selected. To present focused and limited number of answers to the user, the QAS may further re-rank or filter the answers, based on additional requirements such as the limit on the number of answers, background of user or other metadata. Finally a single answer or a ranked list of answers is presented to the user.

*Figure 1. Block diagram of a question answering system (overview)*
Performance Evaluation Metrics for QAS

In literature, several metrics have been used for analyzing the performance of a QAS. A widely used metric is referred to as Mean Reciprocal Answer Rank (MRAR) suggested in TREC (Voorhees 1999). It is defined as follows.

\[
RR = \frac{1}{\text{rank}[i]}, \quad MRAR = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(\text{rank}[i])}
\]

Here \( n \) is the number of questions, \( i \) is the question number. \( RR \) is the Reciprocal Rank. \( \text{rank}[i] \) shows the rank (number) of the document/answer where the first right answer for that question was found.

Benefits and Challenges from use of Natural Language

For human-computer interaction, natural language (NL) provides the most convenient information access mechanism, since it is intuitive, easy to use, rapidly deployable, and requires no specialized training. Use of NL enables user-centric approach in question answering, where the system tries to adapt to the needs of the user and not vice versa. However, use of NL also poses challenges for the QAS developers since different human-users have different needs and ways of expression. This makes the task of developing QAS inherently challenging.

Use of Multiple Resources

Several existing QAS only provide limited search capabilities or cover a limited number of corpuses. Thus, these systems cannot properly answer a variety of complex questions posed by different users. To address this, systems which use multiple techniques or resources for answering are required. Such QAS are referred to as versatile QAS (Mittal & Mittal, 2011). For example, multimedia QAS provide both textual and multimedia contents (e.g. images, sounds, videos) as the answer to user questions. The VideoQA (Yang et al., 2003) is a multimedia QAS which provides video answers (news summaries) to simple factoid questions posed over the news video collection. However, there are several challenges in the use of multimedia QAS. Accessing heterogeneous data, comprising of text, images and videos, from a text query brings issues of lack of coherent and contextual presentation. Further, the ability to meaningfully respond to natural language questions with textual and multimedia content crucially depends on natural language annotations. Thus, the knowledge coverage is dependent on the amount of annotated material. Due to the dynamic nature of Web, any annotation becomes redundant in small period only. To address this, dynamic or adaptive annotations are required.

Multi-modal QAS employ non-text input mode (such as audio-video) for entering the question, apart from/in place of usual text input. Schofield et al. (2003) demonstrate spoken question answering using a commercial dictation engine with language models customized to questions, a Web-based text-prediction interface allowing quick correction of errors, and an open-domain question-answering system, AnswerBus available on the Web. The challenge in using multi-modal QAS is that the input itself is likely to have errors, since additional errors are introduced due to speech recognition, apart from the limitations of QAS.

Similarly, a multi-document comparison QAS generates answers of comparison seeking questions, by searching the answer of each component from possibly a different document (Mittal et al., 2013). For example, in the question, “What is the difference between OSI model and TCP model?” the components being compared are OSI model and TCP model. Thus, by searching relevant descriptions of OSI and TCP model, the QAS can generate meaningful comparison answers.

Multiple-choice question answering systems (MCQAS) seek to answer multiple-choice questions, where a set of possible answers is provided together with the question (Awadallah et al. 2006).
Since the question and four answer choices are already provided, the approach used in these systems to focus on answer selection and validation. The systems can use either a local dataset or, the Web as a corpus.

A multi-stream QAS employs multiple streams for finding the answer of the question, where a stream is a small question answering system on its own (Jijkoun, 2004). Every stream generates a ranked list of answer candidates. As an example, in the Quartz question answering system, different streams are table lookup, collection patterns, Web patterns, collection ngrams and Web ngrams. Note that, in the field of computational linguistics, an n-gram is a contiguous sequence of n items from a given sequence of text or speech.

Multi-Perspective question answering (MPQA) system target answering opinion based questions of the following sort:

1. How is the chairman’s decision of not to pass the upcoming bill looked upon by other members?
2. How do the people from other department regard the admission rules of computer science department?

Stoyanov et al. (2005) explore the use of machine learning and rule-based subjectivity and opinion source filters to guide MPQA systems and argue that the use of these tools may substantially improve the performance of an end-to-end MPQA system. They use special corpus, which is manually annotated with phrase-level opinion information. An example of such a corpus is OpQA which is a corpus of opinion questions and answers. However, using manual annotations does not scale well since the information content available on Web is very large. Thus, given the challenges presenting in understanding the underlying opinion, these systems are still far from a fully functional question answering system.

**FUTURE TRENDS**

A limitation of existing question answering systems is that they are generally not flexible enough to adapt themselves to the knowledge level of a user. Different user-groups (such as novice, researchers, etc.) use different terminologies for understanding the concepts or data relationships. This problem is known as the heterogeneity problem and it arises from using varied vocabulary to describe similar concepts or from using the same metadata to describe different concepts. The future QA systems must resolve the heterogeneity which will enable them to become more user-friendly.

In several technical fields like biomedicine, acronyms are of great importance and are widely used. Moreover, they also reduce the burden of typing long names which may lead to mistake even due to error in a single character. Thus, an intelligent QAS must identify acronyms and facilitate quick matching of them with their expanded (full) forms. However, a challenge in the use of acronyms is that, in several fields, new acronyms are periodically added and hence, maintaining an updated set of acronyms would require significant efforts on maintenance.

It is well-known that in a personal conversation, the question of a user rarely contains the full information required to answer the question. This is because the speaker assumes that listener knows current time, date, etc., and hence, does not include these obvious facts in the question. However, this information may be required to fully answer the question. Thus, the question may contain many unstated assumptions and also requires extending or narrowing the meaning of the question to either broaden or shorten the search process. The question answering systems need to devise ways to deal with such implicit assumptions of the user.

Multi-lingual QAS refers to the QASs that work on and interact with user in multiple languages. Cross-lingual question answering deals with providing an answer in one language (called the target language) to a question posed in a different language (called the source language). Several
existing question answering systems allow capabilities of cross-lingual and multi-lingual question answering (Forner et al., 2009). However, to take the benefits of Big Data to a wide variety of users from different backgrounds, further developments in these question answering systems is required.

CONCLUSION

Question answering systems are inevitable tools in addressing the challenges posed by Big Data. In recent years, developments in the field of QAS have been motivated by various research forums and workshops such as TREC, CLEF, and NTCIR. To realize the goal of fully-automatic question answer systems, several research directions are being pursued, which include automatically built annotations, utilizing advanced NLP techniques and facilitating data integration process through well-designed authoring tools.

It is envisioned that future QAS would provide the ability to easily access large data sets and derive specific, relevant answers. Further, with the advances made in related fields such as image processing (Pande et al., 2009) and high-performance computing (Raju et al., 2009; Raju et al. 2012; Khaitan et al., 2012a, 2012b), multimedia content-rich answers can be provided to the user to further increase his understanding. This will greatly enhance the awareness of the people and enable them to meaningfully use and manage the Big Data.

REFERENCES


Question Answering Systems for Managing Big Data


**KEY TERMS AND DEFINITIONS**

**Closed-Domain QAS:** A closed-domain QAS answers questions on a specific field, such as a Biomedical QAS.

**Cross-Lingual QAS:** Cross-lingual QAS offers the capability to provide answer, to a question posed in one language, in another language.

**Information Retrieval (IR):** IR refers to the task of retrieving useful information from a data set.

**Multi-Document Comparison QAS:** A QAS which can generate a single answer by searching relevant data from multiple documents. This QAS can answer comparison type questions.

**Multi-Lingual QAS:** Multi-lingual QASs interact with the user in multiple languages.

**Multi-Media QAS:** A QAS which returns both text and multi-media information as the answer to the user question.

**Multi-Modal QAS:** A multi-modal QAS employs non-text input mode (such as audio-video) for entering the question, in addition to or in place of usual text input.

**Multi-Perspective QAS:** A Multi-perspective QAS answers opinion-seeking questions.

**Multi-Stream QAS:** A multi stream QAS employs multiple streams for finding the answer of the question, where a stream is a small question answering system on its own.

**Natural Language Processing (NLP):** NLP refers to analyzing, understanding, and generating languages that humans use naturally, without requiring modification to suit computer syntax.

**Open-Domain QAS:** An open-domain QAS can answer questions from different fields which include history, science and geography.

**Question Answering System (QAS):** QAS aims to answer the questions posed by humans in the natural language. Moreover, it provides “answers” and not merely documents.

**Search Engines:** Search engines search the Web resources and provide documents in the response to the question need of the user.