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 (Brill 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

DOI: 10.4018/978-1-4666-5202-6.ch175

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)