Question Answering: A Survey of Research, Techniques and Issues

Vaishali Singh, Department of Computer Science, B.B. Ambedkar University, Lucknow, India
Sanjay K. Dwivedi, Department of Computer Science, B.B. Ambedkar University, Lucknow, India

ABSTRACT

With the huge amount of data available on web, it has turned out to be a fertile area for Question Answering (QA) research. Question answering, an instance of information retrieval research is at the crossroad from several research communities such as, machine learning, statistical learning, natural language processing and pattern learning. In this paper, the authors survey the research in area of question answering with respect to different prospects of NLP, machine learning, statistical learning and pattern learning. Then they situate some of the prominent QA systems concerning these prospects and present a comparative study on the basis of question types.

Keywords: NLP, QA System, Statistical Machine Learning, Statistical Models, Surface Patterns, Templates

1. INTRODUCTION

The World Wide Web has become a popular place and information retrieval systems have become popular medium to access information throughout the world today. However, most of the modern IR systems only help in locating the relevant document and user himself has to investigate further within the ordered list of documents to extract the useful information. For e.g., the question “Who is the Chief Justice of India?” should get back the response “Rajendra Mal Lodha” but instead the user is presented with a list of relevant documents to explore the accurate answer. Thus, QA systems are recognized as a capability with great potential to present user with exact string of response instead of list of documents. A QA interface simulates man machine interaction and provides answers to user question in natural way.

Though automatic question answering will definitely be a significant advance in the state-of-art information retrieval technology in forthcoming years but still there are many challenging issues that are yet to be resolved. The task of understanding the natural language question and deducing its precise meaning for appropriate answer retrieval faces following issues:

DOI: 10.4018/ijirr.2014070102
A. **Question Classification:** Since the advent of QA systems, question classification has remained a separate research problem in itself. Though, lot of efforts has been made for accurate question classification but its success is restricted mainly to factoid questions. When a user inputs a question to its interface, it has become necessary for the QA system to assess the question class accurately to further extract accurate answer, for instance the question, “Who has won Cricket World Cup in 1984 and 2011?” requires name of a country instead of a person name as an answer.

B. **Formulation of Right Query from Question:** Actually this problem could be regarded as a common name to the issues like ambiguity resolution, semantic symmetry detection, identification of temporal relationships and co reference resolution. For example, the question “Who has won maximum gold in Olympics 2012?” has most obvious reformulation as the person who won maximum individual gold medals in Olympics 2012. But the question has one more interpretation as the country who won maximum gold in Olympics 2012.

C. **Knowledge Reuse:** This problem could be regarded as a sub-problem of the above problem as it is related with issue of data (information resource) and the reuse of the knowledge that has already been extracted from the web. A QA system must have capability to identify that the questions “When did India attain independence?” and “In which year India got independence?” is associated with identical set of information.

D. **Answer Granularity:** This is an issue related to the assessment of relevance for correct answer. For instance, consider a question “How many students qualify NET every year from BBAU?” The answer extraction module finds two answer segments, one as ‘More than 50 students qualify NET every year form BBAU’ and other as ‘around 100 students qualify NET every year from different departments of BBAU’, then how does the module be able to select one segment as correct answer.

E. **Answer Validation:** This is a problem that specifically deals with authenticity of the information resource. E.g., for the question, “How many NASA scientists are Indians?” if answer is fetched from a web resource talking about persistent myths about Indians, then how QA system is going to predict the answer as incorrect.

F. **Personalization of Extracted Response:** This problem is often associated with the type and presentation of information, since it is likely that people differ in the contents and presentation they prefer while interacting with the QA system. E.g., the answer of the question “How Columbia University could be reached?” expected by a local New York resident is merely different from the user in abroad.

Different question answering techniques are so far used to deal with these issues directly or indirectly. By the direct approach we mean that the application of the techniques directly addresses the problems. For example, SVM classifier used in question classification identifies question class. By indirect approach we mean that the techniques are used as a part of one of the QA modules that addresses the above problems. For example, NLP could be used to frame heuristic rules for question classification. In the proposed paper we are trying to put forward the prominent QA researches so far done in more structured way from the NLP, statistical learning, machine learning and pattern learning point of view. However, the QA systems that we survey do not necessarily use same technique throughout the whole processing and may adapt different techniques at different phases of question answering.

Earlier, there have been several surveys on the question answering technologies. Simmons (1965) reviewed the very first approaches to question answering in English. Hirschman and Gaizauskas (2001) discussed different phases and evolution of question answering system through the years. Work done by the Lapshin