ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2017

Question Analysis and Classification for Question Answering System

Madhu Siddhartha[1], Ashish Kumar Singh[2*], Sanjay K Dwivedi[3]

[1]Department of Computer Science, Banasthali University, India

[2]Department of Computer Science, Kamla Nehru Institute of Technology, Sultanpur, UP, India

[3]Department of Computer Science, Central University, Lucknow, India

Abstract: Question answering(Q/A) are part of information retrieval (IR) research in which users instead of providing a few keywords to retrieve a set of documents, actually provide complete questions and expect from the QA system to get the most relevant answers(s). This way an efficient Q/A system will be more helpful in providing an accurate answer to a question instead of a query based information retrieval system. No one could have predicted that question answering system would become an indispensable technology for Information Retrieval that would enable the creation of new technologies for information retrieval. The Q/A systems have now been recognized as one of the challenging problem in IR. The task of question classification is one of the most important steps for the efficiency and relevancy of any QA system design. After going through the available literature on the QA systems, we found that there could be different classification of questions. This paper first discusses the general architecture of a QA system and then proposes a more elaborative classification of questions along with some major approaches used for classification in order to design a suitable question answering (QA) system.

Keywords: Q/A system; Question classification; Q/A architecture; Rule based approach; Learning based approach; Hybrid approach

I. INTRODUCTION

The main focus on question answering research came into existence when the Text Retrieval conference began a QA track in 1999. From a collection of documents, a question answering system retrieve relevant information related to questions asked by user. Most of the documents can be searched from the web also, but it is not always possible that web will give relevant answer [1]. Question Answering is a specialized form of Information Retrieval (IR) and the automatic question-answering has become an important research field that may result in a significant improvement in the performance of IR systems and search engines. A Q/A systems may be Open-domain that requires question answering systems to be able to answer questions about any conceivable topic. Such systems cannot, therefore, rely on hand crafted domain specific knowledge to find and extract the correct answers. Open domain QA systems are mainly based on IR techniques [2]. The IR based question answering systems try to find out the answers of a question by processing a corpus of related documents either from web or from any database and then, finding out the relevant answer of that questions. In closed- domain question answering, question are asked under a specific domain (for example, medicine or tourism), and can be seen as an easy task because it refers only those situation where only limited type of questions are accepted. While the TREC-8 Question Answering track involves questions about open domain topics, the question types themselves are a mostly closed set. Pointed out that conception and design of Q/A Systems is the most challenging task. On the other hand David Azari emphasise that general Question Answering System depends on techniques for analysing questions and for composing answers from some corpus of knowledge. It is a challenging problem because the corpus may not contain an explicit matching answer or may contain multiple relevant answers [3].
In Q/A system, a user instead of providing a few keywords to retrieve a set of documents, actually provides complete question and expects from the IR system to get the most current answers(s). This way an efficient Q/A system will be more helpful in providing an accurate answer to a question instead of a query based information retrieval system.
Question in a QA system may be asked in a variety of ways. The WH questions are most common type of question patterns. There may be some other types of questions as well such as list question. When dealing with list questions, we

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2017

need to find all distinct instances and hence one cannot ignore the less frequent answer candidates. Factoid or fact questions are based on the idea that they require only one instance [4]. Definition/descriptive questions can be asked with the motto that the whole information regarding the query can be solved. This paper mainly focuses on different types of Questions or the different ways by which Questions can be asked by the user. Later, the paper provides a brief overview of the three common approaches of question classifications.

II. ARCHITECTURE OF QUESTION ANSWERING SYSTEM

There are three major obstacles to upgrading a search engine to a question-answering system proposed these obstacles as:

- The problem of world knowledge.
- The problem of relevance.
- The basic problem of precipitation of meaning.

The problem of word knowledge means that knowledge which humans can gain or receive through their experiences, education and simple communications such as, Paris is the capital of France. This can be identified from education which we have as this is a fact. If there is no relevancy in documents then would be easy to get the relevant information about the questions. It is difficult for machines to under-stand the natural language but humans can understand it easily. So for web it is important to give the precise answer and for that precision of meaning should be understood by web. These problems may be ad-dressed by the question answering system by enabling us to ask questions in natural language. It is just not possible for the person to search billions of matches by the web, it is however possible in Q/A system wherein a person may ask the queries in a simple question form and expect the relevant answers form the system in the simple natural language without any ambiguity. The architecture of question answering system is shown in (Figure 1). It has similarity with standard crawlers. The question answering system consists of three distinct phases: Question classification, document processing or query reformulation and finally answer ex-traction. In question classification phase, questions asked by the user are processed and detailed information is retrieved rather than a single document. The query is retrieved by the query interface and then query analyzer divides the question into subject, verb and object [5]. By using different approaches (such as support vector machine), to identifies the question type and maximize the performance level of the sys-tem. Second phase is query reformulation or document processing in which documents and sentences are segmented into sentences and entities. The patterns are used to identify the nature of questions asked by the user. Last phase is answer extraction which plays an important role in giving the final answer. The feature based approaches such as support vector machines maximum entropy and naïve based approach are commonly used answer extraction techniques. As the number of matching document returned by the system may be large (documents are retrieved using web search), the answer filtering helps in retrieving the most appropriate paragraph with the help of keywords related to questions. If quality of paragraph is inadequate then again it will returned to the question keyword extraction module and the whole process is again repeated. Answer validation process determines if each candidate answer is correct or incorrect, and also estimates a classification [6].

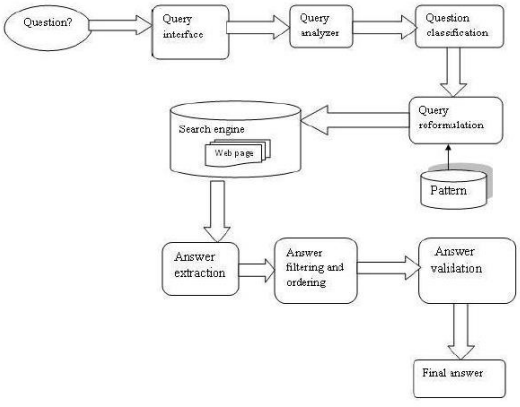

Figure 1: Question answering system architecture.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2017

It is very difficult for a Q/A system to match the right paragraph in order to display the most appropriate answer unless the question type as asked by the user has been classified to know the important keywords and answer types [7]. In the next section we focus on question type's classification.

III. QUESTION TYPE CLASSIFICATION

In order to find out the correct answers, understanding the question type classification is very important. The Question Classification can be done in many ways. Many researchers have proposed various different taxonomies for question classification, proposed a conceptual taxonomy with 13 conceptual classes, propose a multi- layered taxonomy, which has 6 coarse classes (ABBREVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION and NUMERIC VALUE). Moldovan provides another set of question classes and subclasses along with corresponding answer types, based on the 200 question used in TREC 8. We now look at some different Question Types [8]. Although different types of QA systems have different architectures, most of them follow different framework and in some of them question classification plays an important role. Here we present some different taxonomy of Question types which can be as follows;

WH Questions
True or False/yes no
Fact/factoids
List
Definition/descriptive

WH questions -Many researchers have worked on WH questions, they have given many things but still there are so many confusions regarding the probability of asking the same question in how many ways because one question can be asked in many ways and in many forms so here limitations must be defined. Question set contained 500 questions drawn from the logs, puts an additional 193 questions that are syntactically different from the original questions. Some examples of WH questions are:

- What is the capital of up?
- Why does a rainbow form?
- Who is the prime minister of India?

Under this category questions related to the why, when, where may can be asked. True or False/yes or no: To give the answers just in yes-no/true-false can fall in this category, proposed that answering of true false questions can be given in extended form and the whole process can be known as 'over-answering 'means extended responses to yes-no questions are those answers that contain more than a 'yes' or 'no'. But still if we talked about the questions that are only asking the questions in true or false forms can be only answered in one word either true or false. For example:

- Are you studying in this college?
- Delhi is the capital of India?

Fact (factoid) questions: A fact based question only has one correct answer. An answer to a fact based question asks the reader to simply recall the text and point to one specific passage. So here the answers can be given in a factual way such as name and the capital of Brazil [9]. Only one specific answer could be there no other. For factoids we would say that factoids are not suited for web-based because in web search lots of searches come whereas in natural language only one instance comes for the factoids. For example:

- What is your name?
- What is the capital of Brazil?

List based questions: When dealing with list questions, we need to find all distinct instances and hence we cannot ignore the less frequent answer candidates. Based on the observations that different answers can appear in the form of list or tables and from them the relevant form of answers can be retrieved.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2017

Definition /descriptive based questions: These questions can be asked with the motto that the whole information regarding the query can be solved. With not only the basic information but the whole description should be there about the person, places or anything which exists must have full information. In Q/A system this is the basic need that the information which is retrieved should be complete. For ex-ample:

- Describe the Taj Mahal?
- What is the appearance of thief?

These questions require descriptive answers.

IV. QUESTION PROCESSING

Questions processing involves some steps to be followed (Figure 1). Query interface is used to accept the question from the user. The query analyser then phrases the question into subject, verb, object etc. Question classification is used to identify type of the question as discussed before. Classification is an important step as it helps in getting the most appropriate answer type. Finally, Query reformulation is per-formed in order to search the search engine for the possible answer sets. The reformulation of query is carried with the help of important keywords extracted during the query analyser phase and by using the patterns corresponding to a question. Search engine send candidate answers collection to the answer ex-traction module which extracts candidate answers from retrieved documents.
After question classification, answer can be extracted from large amount of questions by finding out the question type or in other words we can first determine the questions category in which they fall and accordingly answers can be matched. The task of question classification is to predict the entity type or category of the answer.
The task of question processing is to analyse the question and create a proper IR query as well as detecting the entity type of the answer, a category name which specifies the type of answer [10,11].

4.1 Approaches to Question Classification

The task of question classification is to predict the entity type or category of the answer. A number of approaches may be used for classification. One of the known approaches is support vector machine for the question classification which identifies the question type and maximizes the performance level of a system. In this section we discuss following three popular approaches.

1. Rule based approach
2. Learning based approach
3. Hybrid approach

Rule based approach: Rule based approach may be applied on the particular dataset because they per-form well on them and if any new dataset is given; its performance does not remain the same. This approach tries to match the question with some manually handcrafted rules. So this approach suffers from the need to define too many rules. Li and Roth provided an example which shows the difficulty of rule-based approaches. The questions shown below are basically same but they are reformulated in different syntactical forms.

- What features of horoscope attracts people?
- Is there any real existence of horoscope?
- What is the main motive of reading the horoscope?
- Is really reading of horoscope is like making fool to the people?
- Who has given the idea of horoscope?
- Is there any genuine reason for the 12 zodiac sign in horoscope?

The above questions are of same class (as they try to get same type of answers) but they have different syntactical forms so they require different matching rules.
Learning based approach: In this approach, classification is performed by extracting some features from questions; it is carried by training a classifier and predicting the class label using the trained classifier. The construction of a manual classifier for questions is a difficult work that requires the analysis of a large number of questions. So, mapping of questions into fine classes requires the correct use of lexical items or specific words. It is very difficult to write a classifier that depends on thousands or more features. A learned classifier is more flexible and easy to reconstruct than

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2017

a manual one because it can be trained on a new taxonomy in a very short time. The difficulty in manually constructing a classifier is to consider reformulations of a question such as:

- What features of horoscope attracts people?
- Is there any real existence of horoscope?
- What is the main motive of reading the horoscope?
- Reading of horoscope is like making fool to the people?
- Who has given the idea of horoscope?
- Is there any genuine reason for the 12 zodiac sign in horoscope?

All these reformulations target the same answer that is location.

Hybrid approach: the hybrid approach combines both the rule based and learning based approaches. The most of the work of hybrid approach has been done by Silva, on question classification. In this approach, the question is first matched with some pre-defined rules and then matched rules are used as features in the learning-based classifier. Learning-based and hybrid methods are the most successful approaches on question classification.

V. CONCLUSION

The objective of question classification is to identify the important word in the question and the type of patterns required to be searched for providing the appropriate answer to the question. Apart from discussing the general classification of question in a Q/A system, we also discussed major approaches used in classification and further processing in QA systems.

VI. REFERENCES

1. T Arturo, Translation Engines: Techniques for Machine Translation. Springer, 2010.
2. V Ellen, H Donna, Overview of the Ninth Text Retrieval Conference (TREC-9). National Institute of Standard and technology.
3. ZA Lotf, From Search Engines to Question Answering Systems – The Problems of World Knowledge, Relevance, Deduction and Precisiation. Laboratory, Department of EECS, University of California, Berkeley, USA 2006; 163-210.
4. A David, HJ Eric, et al. Web-Based Question Answering: A Decision-Making Perspective. University of Washington and Microsoft Research, Washington, 2003; 11-19.
5. L Andrew, A Quick Introduction to Question Answering 2004.
6. M Ramprasath, S Hariharan, A Survey on Question Answering System. International Journal of Research and Reviews in Information Sciences. 2002; 2: 171-174.
7. L Babak, A Survey of State of the Art Methods on question classification. Delft University of Technology Netherlands 2011.
8. M Dan, P Marius, et al. Performance issues and error analysis in an open-domain question answering system 2003; 21:133-154.
9. W Wolfgang, M Heinz M, et al. over- answering yes-no questions: Extended Responses in a NL Interface to a Vision System 1983; 2:643-646.
10. Y Hui, CS Tat, Web-Based List Question Answering. National University of Singapore, 2004.
11. L Xin, R Dan, Learning Question Classifiers 2004; 1: 1-7.