A Survey of Question Answering Systems Using Techniques of Natural Language Processing

Rakesh Kumar

Department of Computer Science, University College Miranpur, Patiala, India
rakesh1404@gmail.com

Abstract

Question Answering (QA) is an ever-progressing research area that integrates the research from multiple streams like Information Retrieval (IR), Information Extraction (IE) and Natural Language Processing (NLP). Question Answering systems obtain answers from a pile of records written in natural language such as web pages of Wikipedia, WWW pages, summary reports of compiled newswire etc. in this survey paper, a literature survey of question answering systems with use of natural language processing is presented. Question Answering systems, is an emerging and exciting research area in natural language processing along with artificial intelligence and Information Retrieval. Generally, QA system (QAS) has three basic parts such as classification of questions, question-based template matching and extraction of answers.

Keywords: Natural Language Processing, Information Retrieval, Question Answering System (QAS), Wikipedia.

I. Introduction

Natural Language Processing (NLP) is a subfield of artificial intelligence that requires the expert skills in multiple inter-disciplinary domains, such as linguistics, statistical, expert systems etc. The term Natural language processing is often utilized to include a collection of methods and techniques involved in the processing of text of unstructured type, although the techniques and methods themselves covering a large range in their application of language knowledge. However, some techniques and methods utilize reduced semantic knowledge, and are based on the appearance of words in textual data. For these techniques and methods, the only semantic knowledge needed is the knowledge of what makes a word; these techniques and methods often count on the bag-of-words approach or a keyword approach. One example of a method is a search engine, which utilizes only words, retrieving all the records data containing the appearance of a combination of words in a cluster, although these words may not totally identical to each other in the documents that are extracted. Question Answering Systems provides a list of possible applicable summary reports in response to a question asked by user, question answering provides the user with either just the answer in the form of text itself or a pathway leading to answer [1]. Provided a question and a collection of documents, a QA system attempts to search the exact response or answer, or at slightest the exact part of text in which the answer is contained.

Question Answering (QA) is an ever-progressing research domain that integrates the research from multiple streams like Information Retrieval (IR), Information Extraction (IE) and Natural Language Processing (NLP). Question Answering systems obtain responses from a pile of records written in natural language such as web pages of Wikipedia, WWW pages, summary reports of compiled newswire etc.,[6]. It is very challenging area but also the methods and techniques developed from question answering systems motivate novel ideas and thoughts in many closely associated areas such as retrieval of documents, time and recognition of named-entity expression. The foremost type of queries that researchers focused on was factual queries, like “When was Mahatma Gandhi born?”, “In what year did world war- I take place?” However, the recent research trends are moving towards more complex types of questions such as definitional questions like “Who is Virat Kohli?”, and entity definitional queries such as “What is HIV?”, listing queries like “List the name of the players who have won Gold Medal in swimming at Olympics”, Question based on situation like “Given a short description of a scenario, answer questions about relations between entities mentioned in the scenario and why-type questions”. A QA system framework is not identical to a data extraction system in that the data that is to be retrieved is of unknown type. Generally, data extraction framework is beneficial for QA system as it provide a greater help by recognizing entities in the textual data. However, Natural Language Processing could prove more beneficial for a QA framework rather than a data extraction or a data recovery system [2].

II. Types of Question Answering Systems

QA systems are emerging as a very favorable research area in Natural Language Processing along with Artificial Intelligence and Information Retrieval [5]. QA systems are intended to automate computing machines for answering the questions of users queries in the language of the users normally in the similar manner human beings can respond to questions and communication purpose. To devise computers or computing machines to answer Natural Language such as English questions is an inspirational and arduous job. The automated question answering systems are categorized into under two categories:
1) Open Domain Question Answering Systems
2) Closed Domain Question Answering Systems.

Open domain QA systems are capable of answering all queries covering each and every public domain. These systems often use search engines to look for the best available answer for the asked question.

Closed domain QA systems generate the answers of the queries under specified domain which search engines are unable in finding and answering to specific queries that are not available in the concerned domain publicly. So, the responses of such queries are maintained by the knowledgebase in a data repository. During retrieving answers, the best available response found from data repository is provided to the client. Closed-domain QA systems requires a technique of matching the templates to perform searching process [4].

Question answering is the area of computer science handling with retrieval of information and NLP. The basic aim of Information Retrieval is to look out for the text in the data repository that looks similar according to client's particular requirement, and the aim of Natural Language Processing is to develop an environment for the dialogue between the client and the machines in human language.

III. Classifications of Questions Levels

User can ask different type of possible questions such as casual questions, template-based question, complex-level questions and professional-related informational analytic questions. As the variations could present in types of questions but the main motive remains same that is to retrieve exact response from the Question Answering systems. Questions can be classified into different levels is discussed as follow

i. Informal Type Questions

Informal type or Casual type questions are termed as those questions which are asked from the system by clients normally. It focuses in normal "perspective" for handling the queries like "When the Statue of liberty was constructed?" and "which leader built the Statue of Liberty?", "When he was born?" and "who invented Helicopter?" All these type of questions are asked from QA systems in the normal context.

ii. Template-Based Questions

Template-based questions are those queries in which templates are constructed for the asked query, which is more associated on the queries based on "linguistic knowledge", like "how to manage time for play" and "does any planet other than earth has life?", and "How Komal manage to complete a task?" and "Does any specific reason to invent Helicopter?" All these type of questions are template-based questions asked from QA systems [3].

iii. Complexity Based Questions

Complexity based questions are those queries which are partitioned into small fragments of queries. These questions are mostly equipped with contextual and specified relations for answering the queries of this kind. The Question Answering system requires to look out for answers from several origins which do not come under the scope of the database searching area [7]. It has the ability for answering the queries like "Does any species of insect have wings?" Cube reporter tries for generating small fragments of queries which are closely related to the original major query that is “When did Rajiv Gandhi died?”, “What was the reason behind his death?” and “What was revolutionary about the Green revolution?” and “When did Ravan died?”, “What was the reason behind his death?” and “What was released by Indian government after Gandhi’s death?”.

iv. Professional Informational Questions

Professional Informational Questions are those queries that are closely associated with "future perspectives". These are used to find diverse type of taxonomies and several factoids which are incorporated in the queries, but it needs a large extent of reasoning methods/techniques to answer these type of questions. For example, “What are the actions taken by Indian government to honor Atal Bihari Vajpayee?”

IV. Architecture of the QA System

The general Framework of the QA system can be partitioned into three basic parts [3], as demonstrated in Figure.1.

1. Pre-Processing Module

Pre-processing module is primarily responsible for preprocessing of the inputted sentences in natural language for cleaning and making sentences more effective. It comprises of three parts, first one is able to transform SMS short forms or abbreviations into simple words in English Language, second one is able to remove stop words and last one is able to remove vowels in sentences. Since the QA system is assumed to compile textual data with both Short Message Service and natural languages it is needed to change the Short Message Service acronyms with the associated words of English language before the process work of user queries further.

This is carried out by linking up with a pre-stored common abbreviations of Short Message Service. Stop words are termed as those words that do not influence the correct interpretation of sentences even if they are ignored or eradicated. Removal of stop words is responsible for increasing the efficiency of the QA system by saving memory and time. Finally, last module get rid of vowels from the sentences for handling mistakes presents in spellings. This process is termed as disemboweling.

2. Question-Template Matching Module

This module compares pre-processed sentence with all templates which are pre-stored until it obtain the exact answer for user question. These pre-stored templates are constructed adhering to a specified syntax and semantics. Moreover, synonyms of words are handled using a separated file of synonyms. It is easily modifiable by user and constantly updatable from WordNet [6].These pre-stored templates are constructed for user queries only. The main aim of this module is to find the best matched template according to the query which was asked by the user.

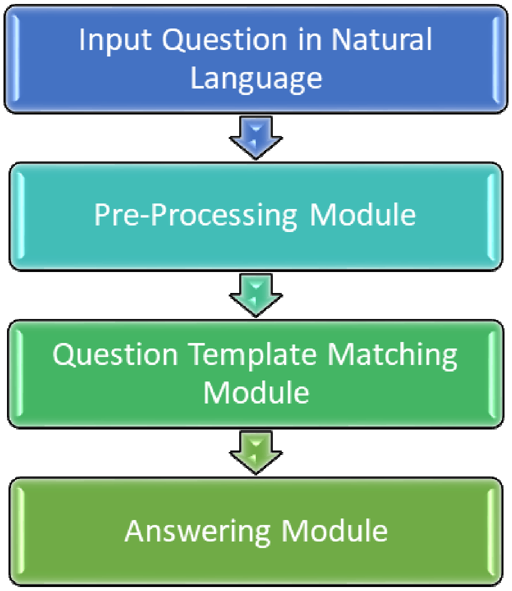
Input Question in Natural Language
Pre-Processing Module
Question Template Matching Module
Answering Module

Figure 1. Architecture of QA System

3. Answering Module

All pre-stored templates associated with a query are maintained in a data repository along with corresponding response. The possible closest template for the query is identified, and the associated response to that template will be extracted and sent to the client.

V. Related Work

Question Answering is the process of constructing designated systems in which it delivers the precise responses of the queries that is raised or searched by an individual in a natural language form. In the task of query execution, a programming code usually obtain and build its own responses by questioning or extracting a structured informational or knowledge database also known as knowledge-base. However, it is also capable of extracting out some answers/responses from some unstructured database set of informational type knowledge. Some of document samples gathered in natural language are as follows [3]:

- A bunch of local collection of reference texts that is so important
- Internal documenting reports and related documents retrieved from WWW.
- Well-documented and summarized newswire documents.
- A collection of Webpages retrieved from Wikipedia.
- A sub collection of WWW pages.

The Text Retrieval Conference (TREC) QA was the foremost examination of Question Answering projects in the year 1999. The basic objective was to generate or extract short responses to a specified query. During surveying of TREC Question Answering systems highlighted that the most of the people like to receive a straightforward answer rather than a record in which they should find out the response taking a lot of time. Earlier, only two Question Answering systems named “BASEBALL and LUNAR” existed. The “BASEBALL“ QA system was able to answer more questions for an comparative study of rocks that was resulted out of “Apollo Moon Missions”. Both the Question Answering systems proved very productive and contributed effectively. In 1971, “LUNAR” was showcased with a lifetime of one year up to 1972. Thus, “LUNAR” was able to answer the geological nature of the lunar science and it succeeded in producing a success rate of 90% in answering the questions that were thrown upon it. Afterwards, the linguistic capabilities of “BASEBALL” and “LUNAR” utilized in similar manner like “ELIZA” and “DOCTOR”. These were the first-ever Chatterbot programs. “SHRDLU” was succeeded working very effectively in the familiar manner of QA systems which was coined by “terry wino guard” in the late 60’s and early 70s and it was primarily associated operating the functionalities of Robots in the toys world.

In 1970s, with the idea of streaming knowledge in several domains knowledge bases were constructed. QA systems are successful in developing a user interface aligned with these developed expert systems. These systems provided the resemblance of the modernized Question Answering systems but were not identical in the internal framework of QA systems. In 1970s and 1980s, types of linguistic computation were developed leading to more productive and influential projects in the area of QA and better understanding of textual data. “EAGLI” is the latest QA system constructed to cater healthcare and life beneficial needs of the people in recent times.

VI. Conclusion

In this paper, a literature survey of Question Answering systems using Natural Language Processing is presented. In this review paper, the types of QA systems and diverse types of question levels are described. The general architecture of question answering system is also presented in this paper. The different research findings of the questions answering system in this domain is discussed. The future direction is Opened domain QA systems; these handles systems with queries about everything with exact answer with minimal answer time which can believe on ontologies and world-level knowledge.

References

[1] Hirschman, Lynette, and Robert Gaizauskas. "Natural language question answering: the view from here." natural language engi-neering 7.4 (2001): 275-300.

[2] Chowdhury, Gobinda G. "Natural language processing." Annual review of information science and technology 37.1 (2003): 51-89.

[3] Tilani Gunawardena, Medhavi Lokuhetti, Nishara Pathirana, Roshan Ragel and Sampath Deegalla," An Automatic Answering System With Template Matching For Natural Language Questions ",IEEE 2010,971-1-4244-8551-2/10.

[4] Prof. Dhanshri Patil,, Abhijeet Chopade, Pankaj Bhambure, ,Sanket Deshmukh, Aniket Tetame," A Proposed Automatic Answering System For Natural Language Questions", International Journal Of Engineering And Computer Science ISSN:2319-7242 , Volume 4 Issue 4 April 2015, Page No. 11310-11312.

[5] Friedman C, Rindflesch TC, Corn M. (2013) Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. J Biomed Inform 46(5), 765–773.

[6] Saranya R, Christopher Augustine, " Schemes and Approaches in Question Answering System", International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) ,Vol. II, Special Issue X, March 2015.

[7] Yogish, D., N, P. M. T., & Hegadi, P. R. S. (2016). A Survey of Intelligent Question Answering System Using NLP and Information Retrieval Techniques. International Journal, 5(5), 536–540. https://doi.org/10.17148/IJARCCE.2016.55134