IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

A Survey of Intelligent Question Answering System Using NLP and Information Retrieval Techniques

Deepa Yogish[1], Prof. Manjunath T N[2], Prof. Ravindra S Hegadi[3]

Department of Information Science and Engineering, Don Bosco Institute of Technology, Bangalore, Karnataka, India[1]

Department of Information Science and Engineering, BMS Institute of Technology, Bangalore, Karnataka, India[2]

Director, School of computational Sciences, Solapur University, Solapur, Maharashtra, India[3]

Abstract : Day by day users of Internet is increasing for information extraction with precise answer. As Users struggle to get exact information, the demand for automated question answering becomes so necessary to search the answer of their questions meaningfully. This is a ambitious work to provide the accurate answer to the user query. When a user asks any question then user finds many of immaterial and unwanted answers. From these answers users get annoy in searching require answer.(QA) system is an intelligent system, which takes user's question in natural language (like English), process the question and find the proper result to the user according to question. The QA systems provide a semantic search on the web, which helps the user to get exact result instead of the uncertain and irrelevant data.QA systems, is a promising research area in NLP along with AI and Information Retrieval. In general, question answering system (QAS) has three modules such as question classification, question template matching and answer extraction.

Keywords: Information Retrieval (IR), Natural Language Processing (NLP), Question Answering (QA) System.

I. INTRODUCTION

Question Answering (QA) is a fast-growing research area that combines the research from Information Retrieval, Information Extraction and NLP. QA systems generate answers from a lump of natural language documents such as Wikipedia web pages, World Wide Web pages, Compile newswire reports etc.,[1]. It is difficult area but also the techniques and methods developed from question answering inspire new ideas in many closely related areas such as document retrieval, time and named-entity expression recognition. The first type of questions that researcher focused on was factoid questions, like "When was Mother Theresa born?", "In what year did world war-II take place?" The recent research trend is shifting toward more complicated types of questions such as definitional questions such as "Who is Sachin Tendulkar?", and entity definition questions such as "What is AIDS?", list questions such as "List the countries that have won the World Cup", scenario-based Question such as Given a short description of a scenario, answer questions about relations between entities mentioned in the scenario and why-type questions. In 1999, an annual evaluation track of question answering systems has been held at the Text retrieval Conference (TREC) (Voorhees 2001, 2003b). Following the success of TREC, in 2002 both CLEF and NTCIR workshops started multilingual and cross-lingual Question Answering tracks, focusing on European languages and Asian languages respectively [2].The TREC has sponsored a question answering track which evaluates systems that answers factual questions by consulting the documents of the TREC corpus. Natural Language processing is an area of artificial intelligence which requires the expertise in multi-disciplinary areas, such as statistics, linguistics, expert systems etc. NLP used by many domains to arrive at a solution using text analytics, where computer interacts text phrases for automation and to structure text to mine and analyze.

II. RELATED WORK

QA systems are a promising research area in NLP along with AI and Information Retrieval. Computers are machines and they are expected to automate such machines to answer to users query in normal English language in the same way human beings can answer and communicate. To develop a computer to answer English language questions is an interesting and challenging task. The automatic answering systems are classified under two categories- open domain systems and close domain systems. Closed domain systems answer the questions under specific domain which cannot be searched using a search engine and answers to questions are not present in the public area. Hence, the answers to such questions are preserved in a database by the knowledge base. During retrieval of answers, the best matched answer searched from database is returned to the user. To perform closed domain systems a template matching technique is used. Open domain question answering system answers to all questions covering all domains. These systems take help of search engines to find the required answer [2].

Question answering is the field of science dealing with IR and natural language processing. The main objective of IR is to search for the data in the resource that match with

DOI 10.17148/IJARCCE.2016.55134

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

user's specified need, and the objective of NLP is to create an environment for the dialog between the user and the system in natural language [10].The TREC Question Answering (QA) was the first evaluation of QA systems. The goal was to return short answers to a particular question. In a survey of TREC QA system turned out that the majority prefer a direct answer rather than a document in which they must search for the answer so it takes long time. As the process of understanding language requires spelling, grammar correction, text classification, sentiment analysis, parsing, meaning extraction and machine translation.

III. OVERVIEW OF QAS

There are many search engines, which are used by the users to get meaningful answer. Evi, wolframa Alpha, Allexperts, Answerbag, Blurtit ,Ask.com,Answers.com, Swoogle, SenseBot, Sindice, START are some search engines .More research work are going on semantic question answering system to provide an exact and precise answer to the user query in a perfect way. The biggest problem of question answering system is that the user not getting the accurate answer they want. Old QA system provides answer to the user with enormous data and irrelevant data. Users have to find answers from large document and even not sure about the answer, because they do not guarantee that the answers provided by systems are correct. Those traditional systems provide bulk of links and data and the user has to search among them to get correct answer.. Another problem occurs in question answering system is that, they use keyword based search. Keyword based search means that search based on the exact word matching. If the exact word not found, then no answer will be shown. Another terminology, which is used in Semantic search engines, is Resource Description Framework (RDF). Resource Description Framework (RDF) is a directed graph, which represent the information on the web. It defines the term metadata, which is "Data about the Data".

The RDF is similar to class diagrams. It uses RDF triples as subject-predicates-objects. It divides the question in subject, predicates and object form to search the answer from the database. Ontology is another important term in Semantic Web to search more appropriate answer. It makes communication easy with the database. It uses the Knowledge Base (KB) for communicating with the system. Ontology defines a conceptual representation of concepts and their relationships within a specific domain. Ontology play an essential role in the semantic web by enabling knowledge sharing and exchange, ontological knowledge structures play an important role in the utility of background knowledge for question answering and information retrieval. Question classification systems is the task of returning documents that are relevant to a particular natural language query [4].A standard example is a query submitted by users through Google search and the query needs to be classified or labeled accurately in order to find relevant documents [5]. The main difference between an IR system and a QA system is that while the former returns to the user the documents that are more likely to be of interest to the query, the latter aims at producing a concise answer extracted from the document(s), not the list of documents.

IV. CLASSIFICATIONS OF QUESTIONERS LEVELS

The possible questions from the users may be casual questioners, template questioners, complex questioners and professional information analyst questioners. The types of questions may vary but the objective is to obtain precise answer from the QA system. Classification of different levels of Questioners is given as follow

A.QUESTIONS BASED ON INFORMAL

Informal or Casual questioners are normal questions to the system from users. It focus in normal "perspective" to handle the questions such as **"When the Great Wall of China built?"** and **"which leader built the great wall of china?", When he was born?"** and **"who invented ELECTRIC BULB?"**All these type of questions are having normal context.

B. QUESTIONS BASED ON TEMPLATE

Template questioners are questions in which templates are prepared for the given question, which focuses on the questions based on "linguistic" knowledge, such as **"how to manage time for study**" and "**does any planet has life?",** and **How Karan manage to complete a task**?" and "**Does any specific reason to invent bulb?"**

C. QUESTIONS BASED ON COMPLEX

Cube questioners are complex questions which are broken down into small set of questions. It majorly consists of context and specific relations to answer the questions of this type. The QA system needs to search answers from multiple sources which lies beyond the database search. It **can answer the questions like "Does any species of insect have wings?" Cube reporter generates small set of questions which are associated to the chief question that is "When did Mahatma Gandhi died?" "What was the reason behind his death?" and "What was revolutionary about the agricultural revolution?" and "When did Ram died?" "What was the reason behind his death?" and "What was released by Indian government after Gandhi's death?".**

D.QUESTIONS BASED ON PROFESSIONAL INFORMATION

These questions are having future perspectives. It is used to identify different taxonomies and multiple facts which are involved in the questions, but it requires much reasoning techniques for answering. Ex **"What are the actions taken by Indian government to honour Mahatma Gandhi?"**-panel menu-selection schemes

V.TYPES OF QA SYSTEMS

Different types of QA systems which are divided based on techniques used and the question that deals [6].

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

A.QA System Based On Web Source

Web is the best source provider to obtain information with the wide spread usage of internet, where user get a enormous data. Web based question answering systems is using the search engines Like Google, Yahoo, Alto Vista etc., to get request webpage's that containing answers to the questions. The majority of these Web based QA systems works for open domain and some of them works for closed domain [16]. The data that is available on web has the characteristics of semi structure, heterogeneity and distributivity. The Web Based QA systems mostly handles wh-type of questions such as **"who killed Indira Gandhi"? Or "Which of the following is correct".** This QA system provides answers in various forms like text documents, Xml documents or Wikipedia.

B. QA System Based On IR / IE

IR based QA systems are providing a set of top ranked documents as responses to the user query. Information Extraction(IE) system is using the natural language processing(NLP) systems to parse the question or documents returned by IR systems, yielding the "meaning of each word". IE systems need several resources like Named Entity Tagging (NE), Template Element(TE), Template relation (TR), Correlated Element (CE), and General Element (GE). Open domain QA systems provide a concise answer to a question, addressed in natural language that is not restricted to any specific field. The knowledge base of a QA system is usually a large collection of documents in natural language. Depending on the size of the information involved, many QA systems use IR modules in their architecture, because of their techniques to process and store the information in a way that enables a query over a large amount of data to be retrieved in a reasonably short time. IR systems process and store large quantities of unstructured information, so that it is able to quickly return the information that is relevant to a given request. Information is input into the IR system through the document concept.

A document is a block of text that will be returned as a whole, by the IR system, as a match to a query to the system. The returned documents of the IR called as hits which are ordered by a scoring function that tries to determine the relevance of the document to the query. The main difference between an IR system and a QA system is that while the former returns to the user the documents that are more likely to be of interest to the query, the latter aims at producing a concise answer extracted from the document(s), not the list of documents. In a QA system, the IR component is generally used to filter out documents that have nothing to do with the question, retaining only the documents that are related, for further processing [6].

C. QA System Based On Restricted Domain

This type of Question answering system required a linguistic back up to understand the natural language text in order to answer the questions accurately. An efficient method for improving the accuracy of QA system was done by restricting the domain of questions and the size of knowledge base which resulted in the development of restricted domain question answering system (RDQA). These systems have special feature like "System must be Accurate" and "Reducing the level of Redundancy". RDQA overcomes the problems find in open domain by attaining better precision.

D.QA System Based On Rule Based

The rule based QA system is an extended form for IR based QA system. Rule Based QA doesn't use deep language understanding or specific advanced techniques. A wide coverage of NLP techniques are used in order to achieve accuracy of the answers retrieved.

Rule a based QA system generates heuristic rules with the help of lexical and semantic features in the questions. For each type of questions it generates rules for the questions such as **who, when, what, where** and **Why** type questions. **"W**ho" questions generates rules which contains Names that are mainly Nouns of persons or things, such as **Who is the chief minister of India?**. **"W**hat" questions generate rules based on generic word matching function shared by all question types it consists of DATE expression or nouns such as **What are the actions taken by government for women's welfare?**. "**W**hen" questions generates rules mainly consists of time expressions only such as When **was television invited?.** "**W**here" questions generate rules are mostly consisting of matching locations such as "in", "at', "near" and inside such as **Where is Taj Mahal ?** "**W**hy" rules are based upon observations, that are nearly matching to the question such as **Why did Dhavan has not attented college?** .These Rule Based QA systems first establish parse notations and generate training cases and test cases through the semantic model. This system consists of some common modules like IR module and Answer identifier or Ranker Module.

VI. GENERALARCHITECTURE

The overall architecture of the system can be subdivided into three main modules [8], as shown in Fig.1.

A.Pre-Processing Module

Pre-processing module mainly consists of three methods, first converting SMS abbreviations into general English words, secondly removing stop words and lastly removing vowels. Since the system is expected to process texts with both natural and SMS languages it is necessary to replace the SMS abbreviations with the corresponding English words before processing user questions further.

This is done by referring to pre-stored frequently used SMS abbreviations. Stop words are the words that add no effect to the meaning of a sentence even if they are removed. Removing stop words is done to increase the effectiveness of the system by saving time and disk space. Examples of stop words are the, a, and, etc. Next step in this module is to remove vowels from the text to handle spelling mistakes. This process is called disemboweling.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

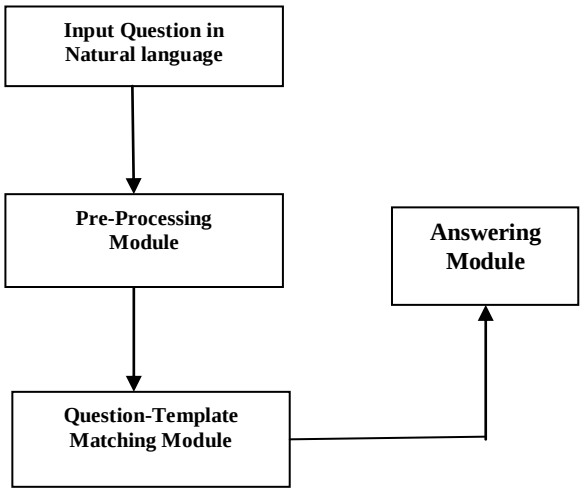

Fig .1 General Architecture of Question Answering System

B. Question-Template Matching Module
The pre-processed text is compared with each and every pre-stored template until it finds the correct answer for user query. In this module, templates are created according to a particular syntax and the information. Further words that are considered to have synonyms are referred in a synonym file. This synonym file can be modified according to the relevant domain and are updated from a standard database such as WordNet [6].The templates are created for questions and not for answers. The objective of this unit is to identify the closest template that matches the question which has received from the user.

C. Answering Module
Every template representing a question are pre stored in a database with its answer. The best matched template for the question is found, the corresponding answer will be returned to the end user.

VII. APPROACHES IN QA SYSTEMS

A. Linguistic approach
Linguistic methods such as tokenization, POS tagging and parsing were applied to develop questions into a correct query that purely extracts the relevant answers from the structured database. The type of Questions Handled in this approach are Factoid questions and there is a deep Semantic understanding .It is quite difficult as knowledge base are generally designed only to handle their pre-stored questions. It is most reliable as answers are extracted from self maintained knowledge base. It is scalable and quite complex as new rules have to be introduced in the knowledge base for every new concept [1].

B. Statistical approach
These approaches propose techniques that not only deal with the huge amount of data but also with their heterogeneity too. The major restriction is that they consider each word independently and fail to find the linguistic properties of a combination of terms. They Support vector machine (SVM) classifiers, Bayesian classifiers, Maximum entropy models are some methods that are used for question categorization purpose. This approach is quite suitable in handling large volume.

C. Pattern Matching approach
This approach uses the communicative power of text patterns to substitute the advanced processing concerned in other competing approaches At present, many of the QA systems automatically study such text patterns from text passages rather than making use of complicated linguistic knowledge or tools viz., parser, named-entity recognizer, ontology, WordNet, etc. It requires much time and rare human skills to install and maintain the system. Various patterns matching QA systems are: Surface Pattern based and Templates based response generation. This approach best suits to small and medium size websites, Semantic web. Various types of questions which are handled are Factoids, definition, acronym, birth date and is having a semantic understanding less than all other competing approaches. It depends on the validity of knowledge resource and scalability is less as new patterns have to be learned for each new concept. The Evaluation Technique/Test Collections used in this approach are domain specific manually developed test collections.

VIII CONCLUSION

In this paper we described about the survey of a QA system for an English language. It receives natural language questions from the user and selects most appropriate answer. This survey paper also describes the different question answering approaches and different types of question answering system. Question Answering (QA) Systems is an automated approach to retrieve correct responses to the questions asked by human in natural language. The concepts behind QA system are to help and improve user-system interaction. Closed-domain question answering deals with questions under a specific domain such as medicine or automotive maintenance and can be seen as an easier task because NLP systems can exploit domain-specific knowledge frequently formalized in Ontologies. The future scope is Open-domain question answering; it deals with questions about anything with accurate answer with less response time which can rely on general ontologies and world knowledge.

REFERENCES

[1]. 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.
[2]. 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.
[3]. Text Retrieval Conference. Trec. http://trec.nist.gov/, 2012.
[4]. Mengqiu Wang,, "A Survey of Answer Extraction Techniques in Factoid Question Answering" , Association for Computational Linguistics, Volume 1, Number 1.
[5] Anbuselvan Sangodiah,, Manoranjitham Muniandy ,Lim Ean Heng ,"Question Classification Using Statistical

Approach: A Complete Review", journal of Theoretical and Applied Information Technology, 31st January 2015. Vol.71 No.3.

[6]. Unmesh Sasikumar, , Sindhu L ," A Survey of Natural Language Question Answering System ",International Journal of Computer Applications (0975 – 8887) ,Volume 108 – No 15, December 2014

[7]. Sanjay K Dwivedi, Vaishali Singh, "Research and reviews in question answering system ",International Conference on Computational Intelligence: Modeling Techniques and Applications(CIMTA),2013.

[8]. 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.

[9]. Sunil A. Khillare,Bharat A. Shelke, C. Namrata Mahender ," Comparative Study On Question Answering Systems And Techniques , International Journal Of Advanced Research In Computer Science And Software Engineering , Volume 4, Issue 11, November 2014.

[10]. M. S. Patil , M. S. Bewoor, S. H. Patil ," A Hybrid Approach for Extractive Document Summarization Using Machine Learning and Clustering Technique", International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1584-1586.

[11]. Tilani Gunawardena, Nishara Pathirana, Medhavi Lokuhetti, Roshan Ragel, and Sampath Deegalla," Performance Evaluation Techniques For An Automatic Question Answering System" International Journal of Machine Learning and Computing, Vol. 5, No. 4, August 2015,DOI: 10.7763/IJMLC.2015.V5.523,pp 294-300.

[12]. S.Babanne,Dr.S.T.Patil,D.J.Joshi," Intelligent Question answering System", International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013 , ISSN 2229-5518 .

[13]. Junichi Fukumoto, Noriaki Aburai, Ryosuke Yamanishi," Interactive Document Expansion for Answer Extraction of Question Answering System", 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013, Procedia Computer Science 22 (2013) PP 991 – 1000.

[14]. Mukul Aggarwal," Information Retrieval And Question Answering Nlp Approach: An Artificial Intelligence Application ",International Journal of Soft Computing and Engineering (IJSCE),ISSN: 2231-2307, Volume-1, Issue-NCAI2011, June 2011

[15]. Gunawardena T, Lokuhetti M, Pathirana N, Ragel R, Deegalla S,"An automatic answering system with template matching for natural language questions", In Proceedings of 5th IEEE International Conference on Information and Automation for Sustainability (ICIAFs), 2010,pp. 353-358.