An Overview of Question Answering System

Article · October 2013

CITATIONS
4

READS
9,402

1 author:

R. Mervin
SAVEETHA ENGINEERING COLLEGE
11 PUBLICATIONS **25** CITATIONS

SEE PROFILE

International Journal Of Research In Advance Technology In Engineering (IJRATE)
Volume 1, Special Issue, October 2013
Proceedings of National Conference on Recent Trends In Web Technologies- Rtwt 2013

An Overview of Question Answering System

R.Mervin
Department of Computer Science and Engineering
Saveetha Engineering College
Chennai
mervinvijay@gmail.com

***Abstract*--Question Answering (QA) system is an information retrieval system in which a direct answer is expected in response to a submitted query, rather than a set of references that may contain the answers. It is a man machine communication device. The basic idea of QA systems in Natural Language Processing (NLP) is to provide correct answers to the questions for the learners. This paper presents a survey of various types of QA systems. These QA systems are classified as Text based QA systems, Factoid QA systems, Web based QA systems, Information Retrieval or Information Extraction based QA systems, Restricted Domain QA systems and Rule based QA systems. The paper further investigates a comparative study of these models for different type of questioners which led to a breakthrough for new directions of research in this area.**

Keywords: Information retrieval, Natural Language processing, Question Answering System.

I.INTRODUCTION

QA systems aim to retrieve point-to-point answers rather than flooding with documents or even matching passages as most of the information retrieval systems do. For E.g. "who is the first prime minister of India?" the exact answer expected by the user for this question is (Pandit Jawaharlal Nehru),but not intends to read through the passages or documents that match with the words like first, prime minister, India etc.,. The major challenging issues in Question answering system is to provide accurate answers from tremendous data available on the web. It also aims to recognize the cross linguistic questions which allow the users to ask the questions and obtain the answers in their native language. The processing of time based information to answer temporal queries still remains as a challenge. In this paper, we focus on different types of QA systems.

QA research attempts to deal with a wide range of question types including: fact, list, definition, *How*, *Why*, hypothetical, semantically constrained, and cross-lingual questions. Generally, the question answering system can be classified into closed domain question answering system and open domain question answering system. Closed domain question answering deals with questions under a specific domain and can be seen as an easier task because NLP systems can exploit domain-specific knowledge frequently formalized in ontologies. Alternatively, closed-domain might refer to a situation where only a limited type of questions are accepted, such as questions asking for descriptive rather than procedural information.Open-domain question answering deals with questions about nearly anything, and can only rely on general ontologies and world knowledge. On the other hand, these systems usually have much more data available from which to extract the answer.

The remaining part of the paper is organized as follows, **Section 2** deals factoid QA systems , **Section 3** describes Web Based QA Systems, **Section 4** discusses Information Retrieval or Information Extraction based QA Systems, **Section 5** deals with Restricted domain QA Systems, **Section 6** deals with Rule Based QA Systems. Finally **Section 7** concludes the paper by giving a brief glimpses into the future directions of research in this area.

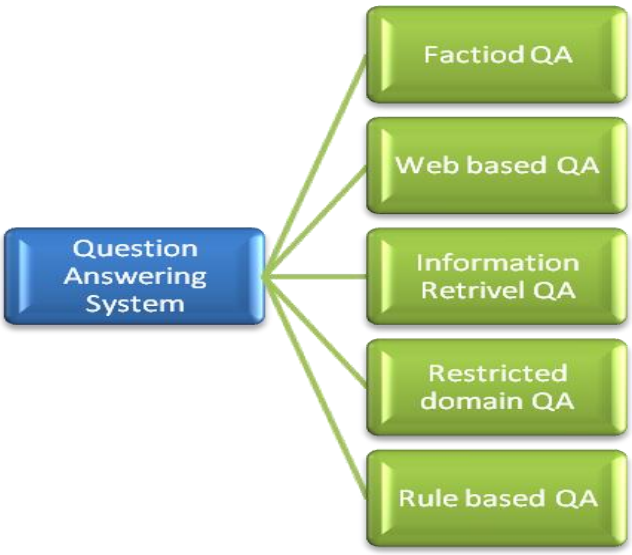

Fig 1 : Types of Question Answering System

II. FACTOID QUESTION ANSWERING SYSTEM

Question Answering (QA) is the task of extracting short, relevant textual answers in response to natural language questions. As a subset of QA, factoid QA focuses on questions whose answers are syntactic and/or semantic entities, e.g. organization or person names. QA systems in general have many important real-world applications, such as search engine enhancements or automated customer service. As of today, the state of the art in QA technology combines machine learning (ML) with linguistic information encoded by human experts in the form of rules or heuristics in most of the QA systems' components (Pa¸sca, 2003). The main drawback of these approaches is that such systems have a extremely high cost of development or customization, especially when high coverage is desired.

The QA system introduced uses a typical architecture consisting of three components linked sequentially: Question Processing (QP), which identifies the type of the put question, followed by Passage Retrieval (PR), which extracts a small number of relevant passages from the underlying speech transcriptions, and finally Answer Extraction (AE), which extracts and ranks exact answers from the previously retrieved passages.

- **Question Processing (QP):** The QP component detects the type of the input questions by mapping them into a two level taxonomy consisting of 6 question types and 53 subtypes.
- **Passage Retrieval (PR):** The PR consists of two main steps: (a) in the first step all non-stop question words are sorted in descending order of their priority, and (b) in the second step, the set of keywords used for retrieval and their proximity is dynamically adjusted until the number of retrieved passages is sufficient.
- **Answer Extraction (AE):** The Answer Extraction (AE) component identifies candidate answers from the relevant passage set and extracts the answer(s) most likely to respond to the user question.

III. WEB BASED QUESTION ANSWERING SYSTEM

With the wide spread usage of internet a tremendous use of data is available, web is one of the ideal source to obtain the information. Web based question answering systems uses search engines (Like Google, Yahoo, Alto Vista etc.,) to retrieve webpage's that potentially contains answers to the questions. Most of these Web based QA systems works for open domain while some of them works for domain oriented also. [7]The web based QA systems such as MULDER [C.Kwok,(2001)], NSIR [D.R.Radev,(2002)] ANSWERBUS [ZhipingZheng,(2002)] falls into the category of domain independent QA systems, while START [Katz,B,(2002) and Katz,B,(1997)] is referred as domain specific QA systems.

The Web Based QA systems mostly handles wh-type of questions such as ***"who killed Mahatma Gandhi"? Or "Which is the longest river in the World".*** This QA system provides answers in various forms like text documents, Xml documents or Wikipedia. The common levels that are used by different web based Question Answering systems architectures are as follows:

- **Question Classification:** This level provides correct answers by classifying the user query into one of the question type to which it belongs to. The question classification is made to provide better accuracy in the results.
- **Answer Extraction:** This level extracts the correct plausible answers for different classification of questions.
- **Answer Selection:** Among the plausible answers obtained, ranking approaches are used to mine the best accurate answers based on its weightage factor.

One of the popular open domain web based QA systems such as LAMP [DellZhang,(2002)] make use of snippet tolerant property to calculate ranking percentages based on which responses are produced.Dialogue based QA systems [RamiReddyNandiReddy,(2004)],WEBCOOP[Benamara.F,(2004)]and[Saint- Dizier,(2004)], is a cooperative type of web based QA systems which integrates knowledge representation and advanced reasoning procedures to generate cooperative responses.

Most of the Web based QA systems performs a better effort to produce correct answers. Though these systems are mostly capable to handle wh-type of questions, but lacks to produce accurate answers, instead these systems retrieves the relevant passages that contain keywords to extract answer from the knowledge base which require an additional process to obtain exact answer. This QA systems fails to provide answers to temporal based queries like for instance "When did the X died?" etc.

IV.IR / IE BASED QUESTION ANSWERING SYSTEMS

Most of the IR based QA systems returns a set of top ranked documents or passages as responses to the query. [4]Information Extraction (IE) system uses 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). IE systems architecture is built into different levels like

- Level 1 NE tagger is used to handle named entity elements in the text(who, when, where, what etc..,) .
- Level 2 handles NE tagging +adj like(how far, how long ,how often etc..,).
- Level 3 builds correlated entities by using the major entity in the question and prepares General Element(GE)which consists of asking point of view.

For *Eg: "Who won the first Nobel Prize in India?*". By passing this question onto the levels mentioned above ASKING POINT is Person (Noun) KEY WORDS such as won, noble, prize etc.., are retrieved. The architecture of IE systems consists of two common modules, they are

- ***Question processor*** which takes the question as input and generates asking point for the question which in turn helps to match for the answer in the text.
- ***Text Processor*** retrieves named entities keywords from the text to generate accurate answers.

Some of the IR Based systems like AskJeeves, LaSiE system performs text analysis which uses some basic modules like Tokenizer, Sentence splitter, Parse process, Name matcher, Discourse Interpreter [Robert Gaizauskas,(1998)].

The IR/IE based QA systems depends on knowledge base which requires an extension to CE and GE components to handle yes/no types of questions in the text. This systems can answer only wh-type of questions but other than wh-type of questions such as *"How can I assemble a computer?"* remains unanswered.

V.RESTRICTED DOMAIN QUESTION ANSWERING SYSTEMS

This type of Question answering system requires a linguistic support to understand the natural language text in order to answer the questions correctly. [1]An efficient approach of improving the accuracy of QA system is 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). RDQA have specific characteristics like "System must be Accurate" and "Reducing the level of Redundancy". RDQA over comes the difficulties occurred in open domain by achieving better accuracy. Early RDQA systems like LUNAR[Woods.W,(1972)] allows to ask geologist questions about rocks. BASEBALL[Green W,(1961)] is another restricted domain QA system, which can only answer about one season's Baseball data. These early systems has encoded large amount of domain knowledge in data bases.

RDQAS doesn't focus on language understanding it focuses on specific set of domain rules. Current RDQA systems are restricted to specific domains like Railways, Medicine, Weather Forecast[Diekema A,(2004)] and Geographic Systems [Chung,H,(2004)] etc,. The RDQA systems uses IE engines which consist of web crawlers and Wrappers, WebCrawler is used to select set of extraction rules that can extract domain information, while Wrappers used to retrieve relevant domain oriented WebPages which contains answers. RDQA first analyzes questions and translates into Structured Query Language (SQL) statements which are further processed to obtain results by retrieving data from database.

Domain Oriented systems make use of domain oriented knowledge bases, domain servers, information systems, parsers etc. Knowledge Acquisition and Access Systems such as KAAS[Anne R,(1999)] uses IR systems to retrieve relevant passages that are processed by NLP. RDQA mostly handles specific questions like for instance Eg**:** "what is the recently invented medicine for cancer?". Domain oriented QA systems contribute great effort to achieve accuracy from the data which is retrieved by the information retrieval. The system requires "Situated evaluations" while comparing with that of open domain QA systems. This type of QA systems needs a domain classifier to handle different domains which is one of the difficult tasks to achieve.

VI.RULE BASED QUESTION ANSWERING SYSTEMS

The rule based QA system is an extension for IR based QA system. Rule Based QA doesn't use deep language understanding or specific sophisticated approaches. A broad coverage of NLP techniques are used in order to achieve accuracy of the answers retrieved. Some popular rule based QA systems such as Quarc [Ellen Riloff,(2003)] and Noisy

channel [Abdessamad Echihabi,(2000)] generates heuristic rules with the help of lexical and semantic features in the questions. For each type of questions it generates rules for semantic classes like *who, when, what, where* and *Why* type questions. "Who" rules looks for Names that are mostly Nouns of persons or things. "What" rules focuses on generic word matching function shared by all question types it consists of DATE expression or nouns. "When" rules mostly consists of time expressions only. "Where" rules are mostly consists of matching locations such as "in", "at', "near" and inside. "Why" rules are based on observations that are closely matched to the question.

These Rule Based QA systems first establish parse notations and generate training cases and test cases through the semantic model. [10]This system consists of some common modules like IR module and Answer identifier or Ranker Module.

- ***IR module*** **:** It retrieves set of documents or set of sentences that contain answers to a given question and returns the results ranker module.
- ***Ranker Module*** **:** Assigns ranks or scores to the sentences which are retrieved from IR module.
- ***Answer identifier*** **:** This module identifies the answer substrings from sentences depending on the score or rank.

Rule Based QA system approach is a wonderful test-bed for NLP to provide accurate answers. The coreference resolution requires automatic extraction of semantic knowledge [AbdessamadEchihabi,(2000)] which is a difficult task to achieve.

VII. CONCLUSION

Question answering system is one of the emerging areas of research in natural language processing applications of Artificial Intelligence. QA systems aim to produce accurate answers, but the current QA systems are succeeded only to some extent. The survey of different QA systems has shown that there are different aspects in answering the questions, while temporal aspect has got little attention. As the data changes in the dynamic world it requires accessing the information and facts which are true at current point of time. Hence forth, it shows that answering temporal based questions is essential in the present QA systems, answering the questions like for Eg "*Who is the president of India during 1950-52?" as* remained unsolved in the present QA systems. The survey of various QA system models has given a novel idea to develop temporal Based Question Answering system that can answer different types of temporal queries.

REFERENCES

[1] Anne R. Diekema, Ozgur Yilmazel, and Elizabeth D. Liddy,"Evaluation of Restricted Domain Question-Answering Systems".(1999) Center for Natural Language Processing.

[2] Abdessamad Echihabi and Daniel Marcu "A Noisy-Channel Approach to Question Answering", (2000).

[3] Benamara.F.,"Cooperative Question Answering in Restricted Domains: the WEBCOOP Experiment",2004. In Proceedings of the ACL Workshop on Question Answering in Restricted Domains.

[4] Benamara,F.and Saint-Dizier,P.,Advanced Relaxation for CooperativeQuestion Answering,(2004).In New Directions in Question Answering. MIT Press.

[5] C. Kwok, O.Etzioni, D. Weld."Scaling Question Answering to the Web". In Proceedings of WWW10, Hong Kong, 2001.

[6] Chung,H.,Han, K., Rim, H., Kim, S., Lee, J., Song, Y. & Yoon, D. "A Practical QA System in Restricted Domains" (2004). In Proceedings of the ACL Workshop on Question Answering in Restricted Domains.

[7] Dell Zhang and Wee Sun Lee."A Web-based Question Answering System" (2002) ,October 31.

[8] D.R.Radev,W.Fan,H. Qi, H. Wu and A. Grewal."Probabilistic Question Answering from the Web",(2002), In Proceedings of the 11th World Wide Web Conference,Hawaii.

[9] Diekema A.R, Yilmazel Ozgur, and Liddy E.D. " Evaluation of Restricted Domain Question-Answering Systems" (2004).In Proceedings of the ACL2004 Workshop on Question Answering in Restricted Domain ,p.p 2-7,.

[10] Ellen Riloff and MichaelThelen."A Rule Based Question Answering System for Reading Comprehension Tests",(2003){rioloff,thelen}.

[11] Green W, Chomsky C, and Laugherty K..BASEBALL: An automatic question answerer.(1961). Proceedings of the Western Joint ComputerConference, p.p. 219-224.

[12] Hermjakob.U."Parsing and Question Classification for Question Answering" (2001). In Proceedings of the ACL Workshop on Open-Domain Question Answering.