A Survey of Natural Language Question Answering System

Unmesh Sasikumar
Department of Computer Science
College of Engineering, Poonjar, Kerala,India.

Sindhu L
Department of Computer Science
College of Engineering, Poonjar, Kerala,India.

ABSTRACT
In Question answering (QA) system retrieves the precise information from large documents according to user query. This survey paper describe the different methods of natural language question answering system (NLQA) for general languages.QA System automatically retrieve correct responses to the question asked by the human in natural languages**.**

General Terms
Question answering system (QAS), Natural language question answering system (NLQA).

Keywords
Factual questions, Question Answering System, closed domain system, Information Retrieval, Natural Language Processing and Information Extraction.

1. INTRODUCTION
Question Answering (QA) is a fast-growing research area that brings together research from Information Retrieval (IR), Information Extraction (IE) and Natural Language Processing (NLP). The techniques and methods developed from question answering inspire new ideas in many closely related areas such as document retrieval, time and named-entity recognition (NER), etc.

The Question Answering system takes questions from natural languages as input and searches matching answer in set of documents and extracts the precise answer to natural language questions. It is different from information retrieval (IR) or information extraction (IE). IR system present the users with a set of documents that related to user questions, but do not exactly indicate the correct answers.

The Question Answering technology takes both IR and IE steps further and provide precise answer to open domain questions formulated naturally [1] IR system locate the documents contain important Information but most of them leave it to the user to extract the useful information from a ranked list [2].The main three components of QA System are first information retrieval engine that sites the top of the document collection and handles retrieval request. The second component is a query interpretation system that decipher natural language question in to keywords or queries for the search engine for fetching the significant documents from database the document can potentially answer the question[3].the third component is answer extraction, it evaluate documents and extract answer snippets from them[4].The taxonomy presented here in this paper for categorizing various QA systems based on following approaches such as Linguistic approach, Statistical approach and Pattern Matching approach. The paper is contains section 2 discuss about QA Approaches. In Section 3, we discuss classification of QA types. Section 4 contains related works Section 5 covers conclusion.

2. QUESTION ANSWERING APPROACHES

2.1 Linguistic approach
A question answering logic contains AI based methods that integrate Natural Language processing (NLP) technique and knowledge base. The knowledge information is organized in the form of production rule, logic frames, templates, ontology and semantic networks; it is used during the analysis of QA pair. Parsing, Tokenization, and POS tagging are linguistic techniques, it implemented to users question for formulating it into a precise query that specified extract the respective response from structural database. In recent work the limitation of knowledge base is accepted as the capability to provide a situation specific are Clark et al [5] presented an approaches for augmenting online text with knowledge base question answering ability. Existing question answering START [6], QA system by chang Et.al [7] and mishra Et.al [8] have acquired web as their knowledge resource.

2.2 Statistical Approach
Importance of statistical approach is increased by the sudden growth of available online text repositories. Statistical approaches are independent of SQL and can formulate queries in natural language form. One disadvantage of statistical approach is it treats each term independently and fails to identify linguistic features for a combination words or phrase. Statistical techniques successfully applied to the different stages of the QA system. The technique used for classification purpose is Maximum entropy models, support vector machine (SVM) classifiers, Bayesian classifiers. The important work based on the statistical method was IBM's statistical QA [9] system. It used maximum entropy model for question/answer based various N-gram features.

2.3 Pattern Matching Approach
The pattern matching approach uses the expressive power of text patterns. It replaces the sophisticated processing involved in other competing approaches. **World Cup 2014 held?"** follows the pattern **"Where was <Event Name> held?"** and its answer pattern will be **"<Event Name> was held at <Location>".** There are two approaches: Surface Pattern based and Template based. Most of the patterns matching QA systems use the surface text patterns while some of them also rely on templates for response generation.

2.3.1 Surface Pattern based
It is either human crafted or automatically learned patterns through examples. Answer sentences for example, the question **"Where was Football"** is extracted using statistical techniques or data mining measures. Pattern learned by in semi automatic and the most compatible application area is small and medium size website.

2.3.2 Template based

This approach makes use of preformatted patterns for questions. The main focus of this approach is more on demonstration rather than explanation of questions and answers. The templates set is built in order to contain the optimum number of templates protect that it sufficiently cover the space of problem, and each of its members represents a wide range of questions of their own type. The entity slots of Templates, which are missing elements bound to the concept of the question that has to be filled to generate the query template to retrieve the corresponding response from the database. The response returned by query will be a raw data; it is going back to the user.

Table 1. Characterization of QA systems

Dimensions	**QA system based on NLP and IR**	**QA systems Reasoning with NLP**
Technique	Syntax processing, Named Entity tagging and IR	Semantic Analysis or Reasoning
Data Resource	Document Free text	Knowledge Base
Domain	Domain Independent	Domain Dependent
Responses	Snippet Extraction	Synthesized Responses
Questions Deals with	Mostly wh- type of Questions	Beyond of wh- type of questions
Evaluatios	Exisisting Information Retrieval	Not Applicable

3. TYPES OF QA SYSTEMS

QA systems are divided into two major groups based on the methods used by them.

3.1 Web based Question Answering System

With the sudden increase in the use of internet, web is the most important source to obtain the information. It depends on the search engine like Google, Yahoo etc. The web based Question answering mainly handles wh-type of questions for example:"who killed Mahatma Gandhi?"

3.2 IR / IE Based Question Answering Systems

Most of the IR based QA systems is returning a set of top ranked documents or passages as responses to the 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". Information Extraction systems have the resources like Named Entity Tagging (NE), Template Element (TE), Template relation (TR), Correlated Element (CE), and General Element (GE). IE systems architecture is build into distinct levels:

Level 1 NE tagger is use 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 the correlated entities by using the most important entity in the question and prepares General Element (GE) which consists of asking point of view. For ***E.g.:*** **"who won the golden boot award in FIFA 2014?"** The answer of the natural languages question is a person (Noun). Then question is passed in to separate levels which mentioned above

3.3 Restricted Domain Question Answering systems

Restricted Domain Question Answering systems requiring a linguistic support to understand the natural language text in order to answer the questions accurately. An efficient approach 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). This system has particular characteristics like "System must be Accurate" and "Reducing the level of Redundancy". RDQA overcomes the difficulties incurred in open domain by achieving better accuracy.

3.4 Rule based Question Answering Systems

The rule based QA system is an extended form for IR based QA system. Each type of questions it generate rules for class like who, when, what, where and why. "**W**hen" rules mainly consists of time expressions only."**W**here" rules are mostly consisting of matching locations such as "in", "at', "near" and inside. "**W**hy" rules are based upon observations that are nearly matching to the question. These Rule Based QA systems first establish parse notations and generate training cases and test cases through the semantic model. Some common modules in the systems are like IR module and Answer identifier or Ranker Module.

IR module*:* It gives the set of documents or sentences that includes the answers to the given question and returns the results back to the ranker module.

Ranker Module*:* Assigning ranks or scores to the sentences which are retrieved from IR module.

Answer Identifier*:* It identifies the answer substrings from the sentences based upon their score or rank.

3.5 Classification of Questioners Levels

The questions may be instructive, investigative, shared or confident in normal context. The methods of these types of questions may vary but the common goal is to obtain precise answer from the system. Classification of different levels of Questioners is given as follows

Table 2. Comparative study of questioner levels with QA systems

Types of Questioner	**Different methods**			
Casual Questioner	Focus: on normal context	Focus: on normal context	Focus: on normal context	Focus: on normal context
	Property: Snippet based	Property: IR from Passage	Property: Domain Service	Property: Rule based
	Reply: Possible Answer from Knowledge Base	Reply: Possible Answer from Data Base	Reply: Named entity Tagging	Reply: from different IR
Template Questioner	Focus: Linguistics concerned templates	Focus: Linguistics concerned entities	Focus: linking generated template with domain	Focus: Rule for each templates
	Property: Snippet based	Property: NLP Technique	Property: NE tagging	Property: Parse notation
	Reply: Possible Answer from Knowledge Base	Reply: Possible Answer from Knowledge Base	Reply: Answer in the asked domain	Reply: from different IR
Cube Reporter	Focus: Resolve ambiguities	Focus: Ambiguities Resolution	Focus: Linking of Generated templates	Focus: Rule based
	Property: Snippet based	Property: Named entity Tagging	Property: Named entity Tagging	Property: Heuristic
	Reply: Possible Answer from ontology	Reply: Large number of files	Reply: Answer in the asked domain	
Professional Information analyst	Focus: Question of Future constraints	Focus: Future perspective	This type of question are not handle	Very difficult to handle
	Reply: Fails to handle, need of temporal knowledge	Property: Named entity Tagging		
		Reply: Very less response or fail		
Precision	High	Low	High	N/A

3.5.1 Casual Questioners

In this type of questioners normal questions are posing to the system. Majorly it focus in normal "perspective" to handle the questions such as E.g.**:"when the Great Wall of China built?"** and **"which leader built the great wall of china?"** All these type of questions are having normal context.

3.5.2 Template Questioners

In this type of questioners, templates are generated for the given question, which focuses on the questions "linguistic" knowledge, For E.g.**: "how to manage time for study**" and "**does any planet has life?"**

3.5.3 Cube Reporter

In this type of questioners the complex questions 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

E.g.**: "does any species of spider have wings?"** Cube reporter generates small set of questions which are associated to the chief question that is E.g.:

"**When did raja Ravi Varma died?" "What was the reason behind his death?" and**

"What was revolutionary about the agricultural revolution?"

3.5.4 Professional Information Analyst

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.

3.6 Question Answering System based on Information Retrieval

Currently, the accessible information, predominantly obtained through the Internet is gradually increasing. The most significant way to access the information is through information retrieval (IR) systems. IR system takes a user's query as input and returns a set of documents sorted by their relevance to the query. Some standard technologies are used to perform the IR task such as existing web search engine like (Google, Alta vista etc...).Question answering is an information retrieval task constrained by an expression of all or a part of the information need as a set of natural language questions or statements. IR systems are usually based on the segmentation of documents and queries into index terms, and their relevance is computed according to the index terms they have in common, as well as according to other information such as the characteristics of the documents, for instance number of words, hyperlink between papers.

4. RELATED WORK

In a system developed Athira P. M, Et.al [10], presented an architecture of ontology-based domain-specific natural language question answering that applies semantics and domain knowledge to improve both query construction and answer extraction. The web as a broad scope, auto-updating knowledge store, answer is mined automatically with a wide range of quires with much less work than is required by modern search engines. The system is able to filter semantically matching sentences and their relations effectively, it ranked the correct answers higher in the result list.

Another system developed by Pragisha K. Et.al [11], described about the. It receives Malayalam natural language questions from the user and extracts most appropriate response by analyzing a collection of Malayalam documents. The system handles four each question. The main answer extraction module is NER in Malayalam. The proposed system design and implementation of a QA system in Malayalam also covered the implementation of some linguistic resources classes of factual questions what, which, Where and which, it extracts precise answer and short answer for user queries in Malayalam.

Research and reviews in question answering system developed by Sanjay K Dwivedi Et.al[12] propose taxonomy for characterizing Question Answer (QA) systems, survey of major QA systems described in literature and provide a qualitative analysis of them. It includes the QA system like Linguistic Approach, Statistical approach, pattern matching approach, Surface Pattern based, Template based etc, They observed that the choice of a technique is highly problem specific. Often a hybrid approach, blending evidently different techniques, provides improved results in the form of high speed, increased relevancy, and higher accuracy and recall measures. QA techniques based on linguistic approach, statistical approach and pattern based approach will continue to remain in sharp focus.

In a System developed by Poonam Gupta Et.al [13] A Survey of Text Question Answering Techniques. Question answering is a difficult form of information retrieval characterized by information needs that are at least somewhat expressed as natural language statements or questions, and was used as one of the most natural type of human computer communication. In comparison with classical IR, where complete documents are considered similar to the information request, in question answering, and specific pieces of information are come back as an answer. The user is interested in a precise, understandable and correct answer, which may consult to a word, sentence, paragraph, image, audio fragment, or an entire document [14]. The main purpose of a QA system is to find out 'HOW, WHY, WHEN, WHERE, HOW, WHAT, WHOM and WHO?"[15].QA systems combines the concepts of information retrieval (IR) with information extraction (IE) methods to identify a set of likely set of candidates and then to produce the final answers using some ranking scheme [16].Types of QA systems are Web Based Question Answering Systems.IR / IE Based Question Answering Systems. Restricted Domain Question Answering systems. Rule Based Question Answering Systems.

Template Matching Automatic Answering System For natural languages questions proposed by Pachpind Priyanka Et.al [17], Frequently Asked QA System that replies with pre-stored answers to user questions asked in regular English, rather than keyword or sentence structure based retrieval mechanisms. Techniques: pattern matching technique Types of QA Systems are, closed-domain QA that deals with questions under a specific domain. Open domain QA that deals with questions about almost everything, and can rely only on general ontology and world knowledge. Main modules are: *Preprocessing*: (a) converting SMS abbreviations into common English words (b) removing stop words, and (c)removing vowels. *Question template matching: The* pre-processed text is coordinated against each and every pre stored template awaiting it finds the best template. *Answering* the matching answer will be returned to the end user.

5. CONCLUSION

In this paper we described about the survey of a QA system for a general languages. It receives natural language questions from the user and extracts most appropriate response. This survey paper also describes the different question answering approach 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 (for example, medicine or automotive maintenance), 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. QA systems in the context of machine reading applications have also been constructed in the medical domain, for instance related to Alzheimers disease. The future scope is Open-domain question answering; it 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. Survey of major QA systems described in literature and provides a qualitative analysis of them.

6. REFERENCES

[1] Surdeanu M, Moldovan D, (2003) "On the role of Information Retrieval and Information Extraction in Question Answering Systems, " Information Extraction in Web Era - Springer.

[2] Tuffis D, (2011) "Natural Language Question Answering in Open Domains, " Computer Science Journal of Moldova.

[3] Kwok C, Etzioni O, Weld D S, (2001) "Scaling Question Answering to the Web," ACM Transactions on Information Systems.

[4] Sucunuta M E, Riofrio G E, (2010) "Architecture of a Question-Answering System for a Specific Repository of Documents, " In 2nd International Conference on Software Technology and Engineering.

[5] Clark P, Thompson J, and Porter B. A knowledge-based approach to question answering. In Proceedings of AAAI'99 Fall Symposium on Question-Answering Systems, 1999, pp. 43-51

[6] Katz B. Annotating the World Wide Web using natural language. In Proceedings of the 5th RIAO conference on Computer Assisted Information Searching on the Internet, 1997, pp. 136-159.

[7] Chung H, Song YI, Han KS, Yoon DS, Lee JY, and Rim HC. A practical QA System in Restricted Domains. In Workshop on Question Answering in Restricted Domains. 42nd Annual Meeting of the Association for Computational Linguistics (ACL), 2004, pp. 39-45.

[8] Cai D, Dong Y, Lv D, Zhang G, Miao X. A Web-based Chinese question answering with answer validation. In Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp. 499-502, 2005.

[9] Ittycheriah A, Franz M, Zhu WJ, Ratnaparkhi A and Mammone RJ. IBM's statistical question answering system. In Proceedings of the Text Retrieval Conference TREC-9, 2000.

[10] Athira P. M., Sreeja M. and P. C. Reghuraj

Department of Computer Science and Engineering, Government Engineering College, Sreekrishnapuram, Palakkad, Kerala, India, 678633. Architecture of an Ontology-Based Domain-Specific Natural Language Question Answering System.

[11] Pragisha K. "design and implementation of a QA system in Malayalam".

[12] Sanjay K Dwivedi, Vaishali Singh. Research and reviews in question answering system Department of Computer Science, B. B. A. University (A Central University) Luck now, Uttar Pradesh, 226025, India.

[13] Poonam Gupta, Vishal Gupta

Assistant Professor, Computer Science & Engineering Department University Institute of Engineering & Technology Panjab University, Chandigarh.

[14] Kolomiyets, Oleksander. And Moens, Marie-Francine. "A survey on question answering technology from an information retrieval perspective". Journal of Information Sciences 181, 2011.5412-5434. DOI: 10.1016/j.ins.2011.07.047. Elsevier.

[15] Moreda, Paloma., Llorens Hector., Saquete, Estela. And Palomar, Manuel. "Combining semantic information in question answering systems" Journal of Information Processing and Management 47, 2011. 870- 885. DOI: 10.1016/j.ipm.2010.03.008. Elsevier.

[16] Ko, Jeongwoo., Si, Luo., and Nyberg Eric. "Combining evidence with a probabilistic framework for answer ranking and answer merging in question answering" Journal: Information Processing and Management 46, 2010 541-554. DOI: 10.1016/j.ipm.2009.11.004. Elsevier.

[17] Pachpind Priyanka P, Bornare Harshita N, Kshirsagar Rutumbhara B, Malve Ashish D *BE Comp S.N.D COE & RC, YEOLA,"* An Automatic Answering System Using Template Matching For Natural Language Questions".