Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
ISSN NO. 2320-5407
Article DOI: 10.21474/IJAR01/1303
DOI URL: http://dx.doi.org/10.21474/IJAR01/1303

*RESEARCH ARTICLE*

# SURVEY ON QUESTION ANSWERING SYSTEM.

***Maria Vijoy and Sangeetha Jamal.**
Department of Computer Science Rajagiri School of Engineering and Technology Kochi, India.

***Manuscript Info***

*Manuscript History*

Received: 19 June 2016
Final Accepted: 16 July 2016
Published: August 2016

***Key words:-***
atural language processing, question answering system.

## *Abstract*

Question answering is an important technique in this new era of information. Question answering system comes under the branch of natural language processing. It works with full potential of information extraction and information retrieval in the area of text mining. Automatically answering repeated question will help us to answer repeated questions exactly ones and it will help to save wastage of resource. This method is very well suited for web based QA systems.This paper mainly aims different technique in answering repeated questions.

*Copy Right, IJAR, 2016,. All rights reserved.*

## Introduction:-

Question answering comes under on wide area of computer science named natural language processing. QA is a special case ofinformation retrieval where text documents are been processed and we will get the relevant and exact answers to the question. Natural language processing are sub groups of artificial intelligence. QA system has two major goals, First is to identify the problems with natural language representation and understanding, second is to built an interface to computers for natural language.

Question answering system is itself a intersection of natural language processing, information retrieval, Machine learning, knowledge representation logic and inference, semantic search[2].QA systems are basically classified into two types 1. Closed domain QA system, 2. Open domain QA system[2]. Closed domain QA systems will be dealing with a specific domain like sports, music etc. Closed domain QA systems are much easy toimplement. Open domain QA system consists all data under. Open domain QA systems are comparatively hard to implement because it will be handling large amount of data. Other types of QA system are, 1. Web based question answering system, 2. IR/IE based question answering system, 3. Restricted domain question answering system, 4.Rule based question answering system.

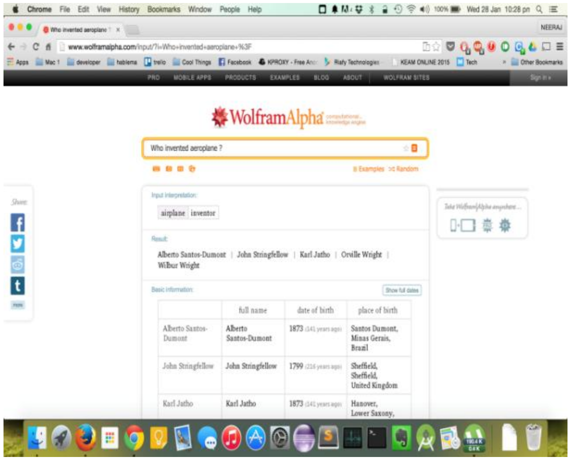
Many web based QA systems are being used now a days like like the different search engines yahoo, Google, Ask etc. Web based QA systems handle Wh questions like who, where, how etc. Example “When was Mahatma Ghandi born?”. IR/IE use information retrieval technology to retrieve exact answer after processing the document. IE system need several resources like Named Entity Tagging(NE), Template Element(TE), Template Relation(TR), Correlated Element(CE) and General Element(GE)[3]. Restricted domain question answering system were built to improve the accuracy of QA systems. Rule based QA systems will be providing rules for each type of questions like

**Corresponding Author:- Maria Vijoy**
Address:- Department of Computer Science Rajagiri School of Engineering and Technology Kochi, India.

who, why, where, where, when etc. For example when is used to retrieve an answer which has time tag associated . Rule Based QA system will improve the accuracy of the system.

When we consider an user interactive question answering system(UIQA) or an collaborative question answering system(CQA)[1] like yahoo answers, stack flow etc, there are millions of questions beenasked by users and another millions of answers produced. And in these CQA systems many questions are asked in different ways which have same answers. Here in this scenario semantic question pattern will help, where it can give a common answer to all similar questions been asked. On an survey taken on 2008 it has been recorded that Yahoo answers alone have got 40 million questions and 500 million answers[4]. This will be a huge wastage of recourse. Semantic search can retrieve information source called ontology[ontology and semantic web]. An ontology compartmentalises the variables needed for some det of computations and establishes the relationship between them[8]. Or simply it can be explained as they reduce the query processing time in a QA system.

WolframAlpha is one of the best QA system that is available now. WolframAlpha is a computational answering engine developed by Wolfram Research. It started of as mathematical QA system, later evolved the domains to larger spectrums and has became one of the most powerful QA System in the world. Wolfram powers Apple’s popular AI assistant SIRIin its Question Answering, also used by Microsoft in BING. Figure 1 shows the user interface of WolframAlpha.



Who invented aeroplane ?

Input interpretation:

airplane | inventor

Result:

Alberto Santos-Dumont | John Stringfellow | Karl Jatho | Orville Wright | Wilbur Wright

Basic information:

| | full name | date of birth | place of birth |
|---|---|---|---|
| Alberto Santos-Dumont | Alberto Santos-Dumont | 1873 (141 years ago) | Santos Dumont, Minas Gerais, Brazil |
| John Stringfellow | John Stringfellow | 1799 (216 years ago) | Sheffield, Sheffield, United Kingdom |
| Karl Jatho | Karl Jatho | 1873 (141 years ago) | Hanover, Lower Saxony, |

**Figure 1:-** WolframAlpha Interphase.

**General qa architecture:-**

To a QA system users can provide queries, system will process the query and will extract the exact answer from the database. In short it is said that an QA system will contain three main module 1. Query Processing 2. Document Processing 3. Answer Processing module[2]. Figure 1 shows the general architecture of QA system.

**Query processing:-**

When the system receives a question the first task assigned is to understand what the actual question is. Query processing will analyze what is the question. In Web based QA systems[ 5] the user will be having a field to input the query. The first step include in identifying category of questions. Here in web based QA questions it consider only factoid questions. Factoid questions are actually one word questions. Therefore query processing initially group the question todifferent types of questions like What, Where, how, who, when etc. Secondly they will classify queries into different groups. Now the question type is been identified. To extract the answer these questions are been passed to information retrieval part. So for this process we identify keywords from the questions. This process is done by different techniques like Named Entity Recognition, Part Of Speech Taggers, Stop Word List etc[5]. By this process a new query is been generated than the initial query user provided. In query processing system will be checking whether the given query is valid or not. If the query is not valid then the question will not be processed further.

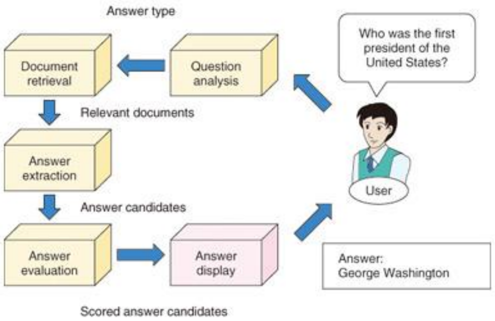


**Figure 2:-** QA architecture [5]

**Document processing:-**

In document processing the new question pattern formed by query processing module is been passed, and this will return set of matched documents. The main technique that is been used in document processing is information retrieval. Simply using pattern matching[ use any] from the given corpus. The document returned by the process willbe very large and they may be large paragraphs. These paragraphs are needed to be filtered [4]. Using paragraph filtering the number of candidate document can be reduced. The main steps that should be included in document any [3]. To retrieve the desired information keywords from the query is been taken. Matching of related documents of information retrieval is done by two matrices precision and recall [5]. Precision refers to the ratio of relevant documents returned to the total number of documents returned. Recall refers to the number of relevant documents returned out of the total number of relevant documents available in the document collection[ 3]. The use of paragraph filtering is same as mentioned above that is to reduce the number of candidate documents. The keywords should be present in nearby paragraphs than in scattered way. If the keywords can be found from successive paragraphs then they can be taken else those can be neglected. Now next step is to Oder the paragraphs according to priority. The ordering of paragraphs can be done by radix sort algorithm[ 7].

**Answer processing:-**

Answer processing will be the final phase of question answering. Here in this phase the final answer will be displayed . The result is stored as a document which is in wx format it will be converted to the text that user required [ 13]. Answer processing willgathering and validating answers. It has three main parts 1. Identify 2. Extract 3. Validate. First the answer should be identified. Answer identification is simply identifying the type of the answer. For identifying type of the answer Named Entity Taggers can be used which will identify name of persons, organizations, date units etc. Or Part of Speech taggers can be used to identify the answer candidates from paragraphs. Once the answers candidates are got some heuristics can be applied and the required answer is been extracted. Simply pattern matching is done, that is all the matched patterns will be returned. If no matched pattern is found QA system will deliver best ranked paragraphs. The final step is to validate the answer got. The answer can be checked with help of different domains. Several people have investigated using the redundancy of the web to validate answers based on frequency counts of question answer collocation and found to be surprisingly effective[9].

## Related works:-

The main task in UIQA system is to answer repeated questions. This is somewhat similar to FAQ answer processing. A structural question pattern can be used for similar questions. Here the importance of semantic question patterns comes. Without using semantic question patterns user will not get the exact and precise answer required. To overcome these shortcomings structural patterns can be used. Semantic questions generalize a class of questions with samesentence structure and relevant semantics. A question pattern is a generalize question with one or several slots termed as variable components termed as semantic labels. Semantic labels are used to remind users fill correct words and also let machines to know the semantics of the filled in words. The question corresponding to semantic question pattern is given .

**Question:-**Who is the president of Russia?
**Semantic Pattern:-** <target:Human/Individual><Q> who</Q> is [Human/Title] of [Location/Country]
TOOLS

The different tools that are been used in semantic question answering systems are Named Entity recognition, Part of speech Tagging and Stemming. Named Entity Recognition will classify elements into predefined groups like persons, organization, location etc. Part Of Speech Tagging is a process marking up a word in text as corresponding to a particular Part Of Speech. For example nouns, verbs, adjectives etc. Here in this process using Part Of Speech taggers all the key nouns are been extracted. Stemming or stemmer's algorithm will reduce its stem or root form. That is stemming helps in reducing the similar or derived word to its unified root. For example "fishing", "fisher", "fished" are reduced to the root word fish.

**Working:-**

Users can ask questions in semantic patterns or in free text or natural language. With these text question answering system first tries to match questions with suitable pattern in pattern database. If there is no patterns matched, the system automatically create new patterns using a pattern generation method and let the user confirm. Each question and its acquired answer is stored in Database as [Pattern-ID/variable components]. Given a new question, system finds the matched question and retrieves the answer if previously answered. This QA system uses both structured and semantic matching [ 7] between the new questions accumulated questions. While implementing semantic QA system, patterns are generated using Automatic Pattern Generation Method [ 10] to have a primary set of patterns. This system mainly have three modules. 1. Structure processing, 2. Pattern Matching And Filtering, 3. Question Similarity Evaluation and Answer Retrieval. Figure number shows detailed flow chart of semantic Question Answering system.

**Structure processing:-**

New free text question is given, the structure processing is necessary to extract the question's main structure and key nouns. The main structure is a simplified representation of the original question or as a question template with key nouns replaced by slots. All nouns are then extracted based on based on Part Of Speech Tagger. Named Entity Recognition (NER) is used to identify certain atomic elements of information in text like persons names, company/organization names, location, sates/time etc. For example " who is the president of Russia?". First acquire the question type "who" by checking question with type list ("what", "who", "when", "where", "how", "how much" etc ). After that label it as <Q> who </Q>. Based on Part Of Speech, "president" is a noun and according to NER, "Russia" is extracted as a named entity of location. Resulting representation will be "<Q> Who </Q> is [key noun] of [key noun]".

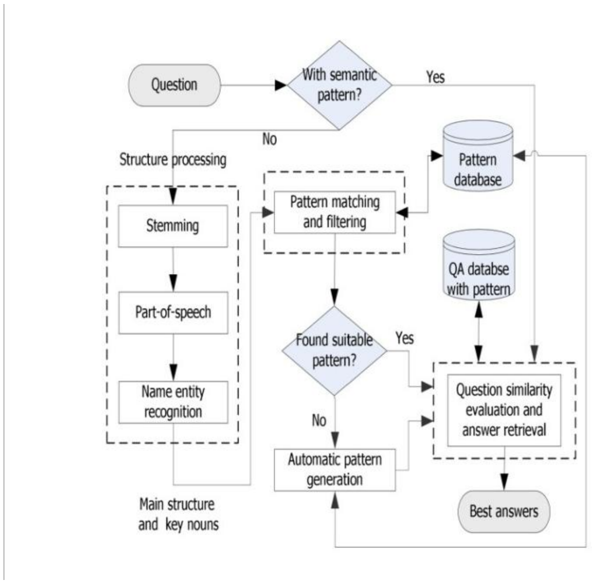


**Fig 4:-** Semantic based QA system.

**Pattern matching and filtering:-**

The pattern matching procedure is used to find the question patterns which best match with the obtained Main Structure (MS). Each pattern will have a unique pattern-ID which will be assigned to the matched question. With this best matched pattern we can retrieve best answer for the question from QA database. This procedure consists of 3 main steps. 1. Retrieve Patterns from the pattern database. Here the result is referred to as Initial Pattern Set (IPS). The MS is split into set of words and match them with each pattern in the pattern database. 2. Find the best label for each key noun. There will br tagging nouns in the main structures with suitable semantic label. For each key noun in the question similarity of all labels is calculated in the label list. For better understanding by common users, it includes only two- level concepts such as “location/city”. We use both the relative depth of the nouns/label as well as their distance to matching labels. 3. Acquiring matched pattern set (MPS). IPS is filtered with the semantic labels, obtained from step 2 and assigned for key nouns. MPS consists of all patterns with final matching scores system.

**Question similarity evaluation and answer retrivel:-**

Each Question is assigned a unique Pattern-ID in pattern database as stated earlier. Related questions and answers are retrieved by querying their pattern IDsin the QA database with pattern. Each question retrieved is a Question Candidate (QC), similarity between Question and each QC is calculated. The Question Candidate with the highest similarity is selectedas the
best matched question.

Semantic based Question answering systems are said to have a precision of 90 %. Precision can be calculated with the equation below.

$$\text{Precision}=\frac{|\text{correctly retrieved answers}|\text{ X100}}{|\text{total retrieved answers}|}$$

This system is best for one word answers. When it comes to paragraph answers accuracy of the system will be failed.

## Result and Discussion:-

This paper give a detailed study of general QA architecture and semantic based QA architecture. The difference in the two architecture is been clearly explained. From this paper it is clearly explained the advantage of using semantic based Question answering system. Using Semantic based QA system will reduce answering repeated questions and it will helps from huge resource wastage.

## References:-

1. Tianyong Hao, Liu Wenyin, Automatically Answering Repeated Questons based on semantic patterns, proc. 10 th IEEE, .c on congnitive informatics & congnitive computing, 2011
2. Ali Mohamed Nabil Allam, and Mohamed Hassan Haggag, ”The Question Answering Systems: A Survey”, International Journal of Research and Reviews in Information Sciences (IJRRIS), September 2012 Science Academy Publisher, United Kingdom
3. Poonam Gupta, Vishal Gupta, A survey of text question answering systems, International Journal of computer applications, September 2012.
4. Text Retrieval Conference (TREC), http://trec.nist.gov.
5. B.F. Green, A.K. Wolf, C. Chomsky, and K. Laughery ,”Baseball: An automatic question answerer”, In Proceedings Western Computing Conference, 1961
6. L. Hirschman, R. Gaizauskas,”Natural language question answering: the view from here”, Natural Language Engineering 7, 2001 Cambridge University Press
7. T.Y. Hao,, D.W. Hu,, W.Y. Liu and Q.T. Zeng. Semantic patterns for user-interactive question answering, Journal of Concurrency and Computation-practice & Experience, 2007, vol. 20, pp. 1-17
8. T.Y. Hao, W.P. Song and W.Y. Liu. Automatic generation of semantic patterns for user-interactive question answering, In Proceedings of Asia Information Retrieval Symposium 2008, Harbin, Jan. 16-18, 2008.
9. M. Ion. Extraction patterns for information extraction tasks: a survey, In Workshop on Machine Learning for Information Extraction, Orlando, 1999.
10. Hovy E., Gerber L., Hermjakob U., Junk M., and Lin C.Y.,”Question answering in webclopedia”, In NIST Special Publication 500-249:The Ninth Text REtrieval Conference (TREC 9)
11. Muthukrishnan Ramprasath, Shanmugasundaram Hariharan , ”Using Ontology for Measuring Semantic Similarity for Question Answering System”, IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT) , 2012.
12. Suarez, O. S., Riudavets, F. J. C., Figueroa, Z. H., and Cabrera, A. C. G. “Integration of an XML electronic dictionary with linguistic tools for natural language processing” Journal of Information Processing & Management, vol. 43, 2007, 946-957
13. Metais, E. “Enhancing information systems management with natural language processing techniques,” Journal of Data & Knowledge Engineering, vol. 41, 2002, 247-272.