Question Answering System based on Question Classification and Sentential Level Ranking

Shruti Gupta
Department of CSIT
Moradabad Institute of Technology, Moradabad
U.P. India

Shilpi Malhotra
Department of CSIT Moradabad Institute of Technology, Moradabad,
U.P. India

ABSTRACT
Question answering system provides the way which helps us in reducing the time of search for the useful information from huge amount of data. An extensive work has been done in the field of Question Answering systems but there exists a scope of further improvement in this field. In the proposed architecture the answers are extracted by correctly classifying the questions. Paragraph ranking is used to reduce the text that reduces the memory as well as processing requirement. Named Entity Recognition technique helps to increase the accuracy of the answer returned.

The proposed implementation takes question as input from user, classifies the question and then attempts to find answer that will be based on corresponding answer type. The techniques like paragraph ranking, preprocessing, indexing etc are used to improve the efficiency and accuracy of the system. Thus the system provides accurate and relevant answer of the query without making so much effort and also helps in reducing overall time in searching for answers. Moreover the result returned will be very concise.

Keywords
Indexing, Named Entity Recognition, Question Classification, Ranking, Summarization

1. INTRODUCTION
The amount of data is increasing exponentially with time but the information is lacking. There is a crucial need to mine the useful information from the available pool of data. Question Answering System Based on Question Classification and Sentential Ranking is the next step in information retrieval. The main motive of the question answering system is to provide a short answer to the user so that user does not need to search in the whole data. In QAS, the main issue for the researchers is to provide accurate answer from huge collection of information on the Web. A Question Answering System (QAS) allows the user to ask questions in natural language and to obtain one or several answers. If compared with a classical search engine, QAS retrieves exact answer in few or less lines to user's question rather than retrieving the whole document. Question Classification is a technique used to extract useful information from the question by identifying its class. Then, to provide user with relevant answer, the appropriate answer type needs to be identified on the basis of user's question. If the user asks "Who invented the first computer?" the user expects "Charles Babbage" as the answer which is the name of a person. In this, the question class "Who" is mapped to the expected answer type i.e. Person. The next important thing is data. For the success of Question Answering System relevant data is the main component. We can collect the data from the web. The problem with Web Content is that the data retrieved from web is very frequently unorganized; therefore there is a need to identify useful data and then using the portion of data that is significant for the user's question. For this purpose we need to summarize the data. Summarization is a technique used to extract sentences from a text document that best represent its content. Text summarization (TS) is the process of identifying the most salient information in a document or set of related documents and conveying it in less space (typically by a factor of five to ten) than the original text [1]. There are two types of summarization that is used; content based and context based summarization. The content-based summarization utilizes textual content of the web document while context-based makes use of the hypertext structure of the web and there are basically two known techniques for the summarization. Extractive summaries (extracts) are produced by concatenating several sentences taken exactly as they appear in the main document being summarized. Abstractive summaries (abstracts) are expressed in the words of the summary author. [2] In abstractive summaries the sentences can be rephrased. The proposed implementation uses content based extractive summarization. The techniques like stop word listing and stemming are used to increase the accuracy of the system. The proposed architecture focuses on factoid questions inducing entities and provides the short and relevant answer by using efficient searching and ranking techniques.

2. RELATED WORK
This paper tried to retain the features of other findings in this area and few more important ones are added to enhance the performance further. Huang et al. in [3] discussed a classification using Head Words and their Hypernyms where two models of classifier are used namely Support vector machine and maximum entropy model. Support Vector Machine is a useful technique for data classification. It uses Kernel function for problem solution. In his technique, each question is represented as a bag of features like feature sets, namely question wh-word, head word, Word Net semantic features for head word, word grams and word shape feature.

Similarly, Bu et al. in[4] proposed a function-based question classification technique in which question classifier based on MLN is included. A function-based question classification category tailored to general question answer. The category contains six types namely Fact, List, Reason, Solution, Definition and Navigation. Each question is split into functional words and soft patterns from content words is generated. The matching degree is either 0 or 1 for strict pattern.

Chang et al.in [5] Minimally Supervised Question Classification and Answering based on Word Net and Wikipedia, the question is classified into semantic categories in the lexical database. Without the help of external knowledge, surface pattern methods suffer from limited ability to exclude answers that are in irrelevant

semantic classes, especially when using smaller or heterogeneous corpora.

Zhang et al. [6] considered the machine learning approach for question classification. Support Vector Machine replaced the regular expression based classifier with the one that learns from a set of labeled questions. SVM is a binary classifier where the idea is to find a decision surface that separates the positive and negative examples called support vectors. Although SVMs are binary classifiers but they can be extended to solve multi-class classification problems, such as question-classification.

A QA System supported by Information Extraction discusses an information extraction system Text extract in (NL) and examines the role of IE in QA application. But this system also suffers from the problem of linguistic processing.

3. PROPOSED ARCHITECTURE

The proposed architecture uses the question class so that answers can be searched in that particular domain only. There by increasing the efficiency of the system. For example if the question is of "When" class it is better to search the document for date and time rather than searching the document and returning the sentences that contain information other than date. Therefore the relevant document is further summarized so as to reduce the text amount to only the sentences that may contain required information. The set of questions that are considered are factoid questions inducing entities. Such type of questions includes when, who, where, which etc. The high level architecture for Question Classification based Question Answering System is given in **[FIGURE-1]**.

Each of the components is discussed below:

A. User Interface

A user can enter the question using user interface that acts as input to Question Classifier. User interface passes the question to the next module Question Classifier.

B. Question Classifier

Since every question consists of two parts Question Class and the rest of the terms for which we are searching the information. Question class tells the type of class i.e. Who, When, Where, Which, Why and How category. The remaining terms of the question will form query keywords. Classification process consists of two steps:

I. First word of the question identifies the question class. For e.g. "Who invented the first computer?" has "Who" as question class and "invented, computer" as Term Set. The system maintains a table called QuesClass that stores the question class with expected Answer Types. For e.g. in the questions mentioned above user is expecting that the answer returned by the system consist of the names of the person or an organization. So it is fruitful and efficient to search only the sentences that consist of Person or Organization relevant to the question while discarding others.

II. Remaining part of the question is termed as the Query keywords. These query keywords are then used to collect the document from web. The process of converting the question into Question Class, Answer Type and Term Set is given in **[TABLE-1]**

TABLE 1: Processing of Question

Question	Question Class	Query	
		Answer Type	Term Set
Who invented the first computer?	Who	Person, Organization	Invented, first, computer
Where the Taj Mahal is located?	Where	Location	Taj Mahal, located

C. Crawler

The Crawler downloads web pages. Question Classifier module passes the Term Set to the crawler. These term sets are converted to query and passes to the search engine. The list of results returned from any search engine is already ranked. So it is sufficient to take first five results and convert them to text files (Page Repository) which are then passes to next module summarizer.

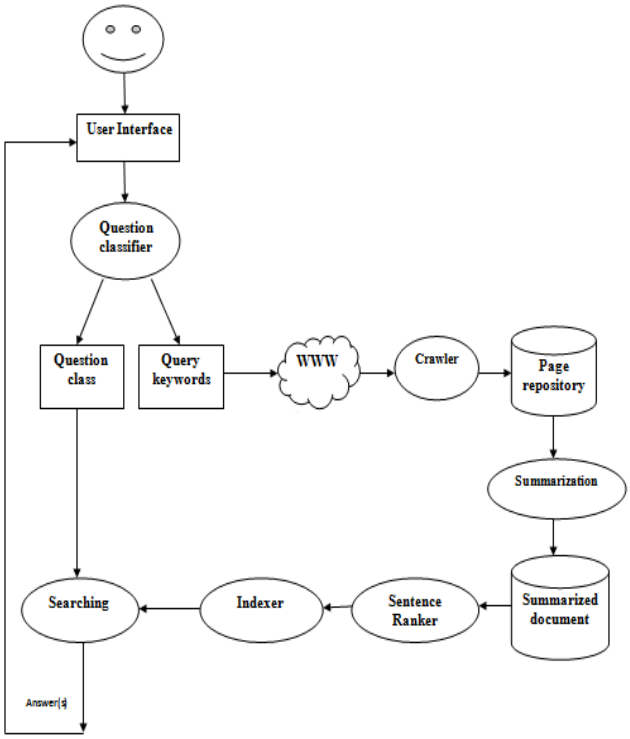

FIGURE 1: Proposed Architecture

D. Summarizer

Summary is a brief explanation of a larger body of work. To summarize means to take out main ideas and/or themes and explain them in shorter way. Summarization process filters out non-relevant content from the documents. The text produced specifies what the original document is about.

The proposed architecture summarizes the document using paragraph ranking and then extracting the paragraphs with

highest rank. The rank of the paragraph is done by giving the score to the paragraph on the basis of same word sequence score i.e. computes the number of words from the query that are recognized in the same sequence in the sentence.. Thereby the relationship among terms in query is used to search the information relevant to query and discarding others. For example for the question," Who invented the first computer?" Then we have terms such as invented, first and computer. So the system will check for the occurrence of these terms in the same order. The paragraph is assigned score every time the term set occur in the same order in sentences of the paragraph. This process continues for all the paragraphs and at the end of the process, each paragraph carries a score and highly ranked paragraph are extracted as summary of the document.

The process of summarization is described in **[FIGURE-2]**

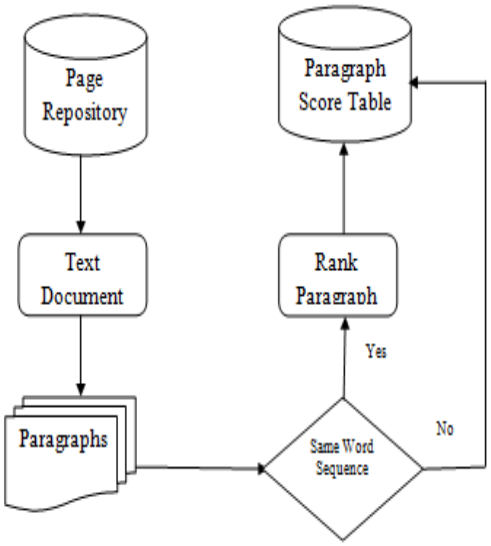

FIGURE 2: Process of Summarization

To increase the efficiency as well accuracy of the summarization process, the document is preprocessed to determine paragraph boundary, sentence boundary, stop-word removal and stemming. Tokenize the summarized document. Tokenization is the process of parsing the document and extracting token. [7] A stop word is a commonly used word (such as "the","of","a","about" etc) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. The list of words that are not to be added is called a stop list. [8] After getting a token we can check whether a particular token is a stop-word or not. Stemming [9] is a process of reducing a word into its stem like reducing cars to car, invention to invent etc. Porter stemmer algorithm is used

to perform stemming. Stemming is applied to both the query keywords and to each term of the document.

An example of the query based extractive summary is given in **[TABLE-2]**.

E. Sentence Ranker

Sentence ranker takes summarized document as input and perform sentence ranking on the sentences of the highest ranked paragraph. The sentences that contains all the query terms in the same order as they are in query are given higher score than others. Thereby discarding all the sentences that are lacking the information what is required. This is highly significant to questions that involves Who, Where, When and what that expects short answers of one word or one sentence answer length. The sentences, in which query terms are found, are analyzed to retrieve named entities like location, person, date, organization etc. These terms are then indexed. A result obtained after Sentence Ranker on highest ranked paragraph is shown in **[TABLE-3]**.

TABLE 3: Highest Rank Sentences

Sentence	Sentence Score	Sen-id & Page-id
Charles Babbage, a British professor mathematician is the man who invented the first computer in 1837.	1	1,1
It was basically an mechanical type of calculator that also had a memory.	0	2,1
The computer was powered by steam engine and used punched cards for programming	0	3,1
The Man who invented the computer is a 2010 historical biography by author Jane Smiley about American physicist John Vincent Atanasoff and the invention of the computer.	0	4,1

F. Indexing

The index for the summary obtained from summarizer is created. The indexer contains all the entities that occur in the highest rank sentences. Since whatever the answer user is expecting is in the ranked sentences, it is advantageous to make index of ranked sentences rather than of whole document. Hence it greatly reduces the memory requirement. Each sentence is searched for all the named entities it contains. The highest scored sentences are then sent to Stanford Named Entity Recognizer (NER) for the extraction of entities like Person, Organization, Location from those sentences to create an index. These entities are then indexed. Indexer contains information that involves Sen-id, Answer Type and corresponding terms in sentences. For example for the question, "Who invented first computer?" Let one of the most ranked sentence is, "Charles Babbage, a British professor mathematician is the man who invented the first computer in 1837" In this we get named entities one Person (Charles Babbage) and Date (1837) in this sentence. The indexer figure is shown in **[FIGURE-3].**

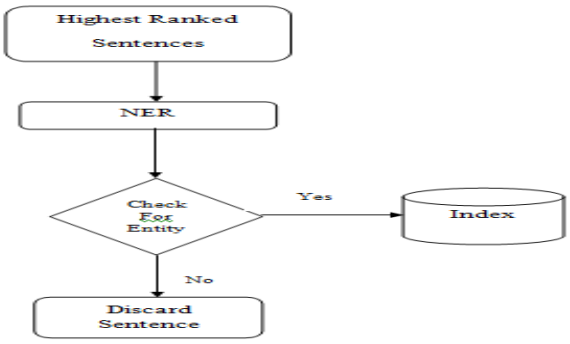

FIGURE 3: Indexing

TABLE 2: Summarization of Text

Text File	Charles Babbage is credited with the first design of a fully programmable computer in 1837. Keep in mind those early inventions that helped lead up to the computer such as the abacus, calculator, and tablet machines are not accounted for in this document. There have been a lot of major milestones throughout history that warrant mentioning. The adventure starts in the year 1936. Without getting too technical, the first "computing machine" was created by Charles Babbage in 1822. His idea was not really to create a computer as we know them today, but instead, to create a machine that would compute math problems. He was tired of human errors in completing math problems, so he sought to create an infallible math machine, but what he got instead was a machine that was the basis for what we know now as the computer. Charles Babbage invented the first computer in 1837. It was basically an mechanical type of calculator that also had a memory. The computer was powered by steam engine and used punched cards for programming. The Man Who Invented the Computer is a 2010 historical biography by author Jane Smiley about American physicist John Vincent Atanasoff and the invention of the computer. The book follows Atanasoff as he collaborates with others to develop the Atanasoff-Berry Computer (ABC), the first electronic digital computing device.He first mechanical computer created by Charles Babbage doesn't really resemble what most would consider a computer today. Therefore, this document has been created with a listing of each of the computer firsts starting with the Difference Engine and leading up to the types of computers we use today.
Summary	Charles Babbage, a British professor mathematician is the man who invented the first computer in 1837. It was a mechanical type of calculator that also had a memory. The computer was powered by steam engine and used punched cards for programming. The Man Who Invented the Computer is a 2010 historical biography by author Jane Smiley about American physicist John Vincent Atanasoff and the invention of the computer.

After applying indexing scheme system get the index in the given format shown in **[TABLE-4]**.

TABLE 4: Index

Answer Type	Term	Sen-id
PERSON	Charles Babbage	1
DATE	1837	1

G. Searcher

Searcher takes the Question Class as input and then maps the question class into appropriate Answer type(s) by using the Table. Then using index given in **[TABLE-4]**, it searches for a match to the corresponding answer type. When a match is found it extracts the Sentence id and corresponding sentence is retrieved and returned as answer to user. For example, for the question "Who invented first computer?" the Answer type will be Person or Organization. So searcher will search Indexer to find the terms for Person i.e. it selects the entry with Person or Organization as Answer Type. Therefore the first sentence, "Charles Babbage invented the first computer" is returned to user as answer and second sentence with Date is not returned. The mapping of Question class into Answer type is given in **[TABLE-5]**

TABLE 5: Question Class and its Answer Type

Question Class	Answer Type
Who	Person, Organization
Where	Location
When	Time, Year, Day, Date
Which	Location, Month, Year
Why	Reason
How	Process

4. EXPERIMENTAL EVALUATION

The evaluation criteria play an important role in the completeness of performance analysis. There are several question answering systems present like QUALIFIER, TextmA, QuASM and START. The proposed system is compared with START QA system which is a natural language base question answering system. The performance of the system is compared by using a set of 10 questions posed to the systems, our proposed system and START as well. The answers retrieved from both the system is given in [**TABLE-6**]

TABLE 6: Answers retrieved from START QAS and proposed QAS

S.NO	Questions	Proposed QAS	START
1.	Who invented the first computer?	Charles Babbage, a British professor mathematician is the man who invented the first computer in 1837	(first design of analytical engine) Charles Babbage, 1830s; (ENIAC, Electronic Numerical Integrator and Calculator, first all-electronic, completed) John Presper Eckert, Jr., John Mauchly, U.S., 1945; (dedicated at University of Pennsylvania) 1946; (UNIVAC, Universal Automatic Computer, handled both numeric and alphabetic data) 1951; (personal computer) Steve Wozniak, U.S., 1976.
2.	When did the titanic sink?	RMS titanic was a British passenger liner that sink in the North Atlantic Ocean on 15 April 1912 after colliding with an iceberg during her maiden voyage from Southampton, UK to New York City, US	I am sorry to say I don't know when the Titanic sank.
3.	Where the taj mahal is located	The Taj Mahal or the "Taj" is a white marble mausoleum located in Agra, Uttar Pradesh, India	I don't know the answer.
4.	When did the C language develop?	In computing, C is a general-purpose programming language initially developed by Dennis Ritchie between 1969 and 1973 at AT&T Bell Labs	In computing, **C** (/ˈsiː/, as in the letter C) is a general-purpose programming language initially developed by Dennis Ritchie between 1969 and 1973 at AT&T Bell Labs.[5][6] Like most imperative languages in the ALGOL tradition, C has facilities for structured programming and allows lexical variable scope and recursion, while a static type system prevents many unintended operations.
5.	Which is the coldest place on the earth?	The coldest place on Earth is a high ridge in Antarctica on the East Antarctic Plateau where temperatures in several hollows can dip below minus 133	Unfortunately, I don't know what the coldest place on the Earth is.
6.	Who is known as the master blaster of Indian cricket?	Sachin Tendulkar is known as the "Master Blaster" of Indian Cricket	Unfortunately, I wasn't told who is known as the master blaster of Indian Cricket.
7.	What is the location of leaning tower of pisa?	The location of Leaning Tower of Pisa is behind the Cathedral and is said to be the third oldest structure in Pisa's Cathedral Square after the Cathedral and the Baptistry.	Website: http://www.opapisa.it/en/home-page.html Location: Italy Province: Pisa **Source:** Wikipedia .
8.	When did the first email sent?	In 1971 Ray Tomlinson sent the first successful email from one computer, to another sitting right next to it	I don't know the answer.
9.	Who built the Taj Mahal?	Mughal emperor Shah Jahan built the Taj Mahal in memory of his third wife, Mumtaz Mahal	Architect: Emperor Shah Jahan
10.	Why sky is blue?	A clear cloudless day-time sky is blue because molecules in the air scatter blue light from the sun more than they scatter red light	I don't know the answer.

Thus it is very much clear from [Table-6] that proposed system has provided better answers for most of the questions. The answers from proposed system are exact and not of very much length for the factoid based questions which is not so for START QAS.

5. FUTURE SCOPE

This research is basically carried for the factoid questions inducing entities which provide the relevant and short answers. This approach can be extended for the various categories of question such as YES/NO questions, definition

type questions, procedural questions etc. Moreover different ranking techniques can also be applied based on the question category. As questions like yes/no or factoid based has short answers but definition type, differences and procedural type questions has few lines or paragraphs. Therefore some hybrid approach is required so that system can run efficiently on larger domain of questions.

6. CONCLUSION

Question classification is the main soul of this system. And then after finding the appropriate class and answer type, proper mapping of question class with the expected answer type needs to be done through which indexing takes place. Thus we can say our this technique is different from traditional classification techniques as it is based on indexing the documents on the basis of expected answer type .In the recent progress in our system we have achieved success in getting the accurate extraction of that paragraph which contains the maximum number of simultaneous terms or simultaneous query keywords as given in the query posed by the user in the user's interface. Then named entities are also extracted from that paragraph which has the maximum possibility of having the answer which is used to create an index containing the answer types and the terms. Now work has to be done in order to get the answer from that highest ranked paragraph, by collaborating all the other modules in such a way that accuracy and performance of the system will be maintained by maintaining the degree of the questions which user can pose to the interface. In the future, work can be done on the system in order to increase the domain of the variety of questions that a user can pose, addition of Multiple language support and refinement in document fetching.

7. REFRENCES

[1] ciir.cs.umass.edu irchallenges presentations summari ation3.doc

[2] Ani Nenkova, Kathleen McKeown "Automatic Summarization" , Foundations and TrendsR in Information Retrieval Vol. 5, Nos. 2–3 (2011) 103–233

[3] Terrence A. Brooks [2003] Web Search: How the Web has changed information retrieval, Information Research.

[4] Dell Zhang, Wee Sun Lee [2003] "Question Classification using Support Vector Machines",in proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval.

[5] David Pinto, Michael Branstein, Ryan Coleman, W. Bruce Croft, Matthew King, Wei Li and Xing Wei [2002] QuASM: A System for Question Answering Using Semi-Structured Data, 2nd ACM/IEEE-CS joint conference on Digital Libraries, Amherst, MA.

[6] Mani, I., MayBury, "M.T. [1999] Advances in Automatic Text Summarization", the MIT Press.

[7] Renu Mudgal, Rosy Madaan, A.K.Sharma , Ashutosh Dixit , "A Novel Architecture for Question Classification Based Indexing Scheme for Efficient Question Answering", International Journal of Computer Engineering & Applications, Vol. II, Issue II

[8] Shilpi Mahlotra,"Web Document Summarization Using Multiple Document Reference," MIT International Journal of Computer Science & Information Technology

[9] Rosy Madaan, A.K. Sharma, Ashutosh Dixit [2012] "A Novel Architecture for a Blog Crawler", 2nd IEEE International Conference on Parallel Distributed and Grid Computing (PDGC).

[10] Rafeeq Al-Hashemi, "Text Summari ation Extraction System (TSES) Using Extracted Keywords". International Arab Journal of e-Technology, Vol. 1, No. 4, June 2010.

[11]http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-3.html

[12] http://en.wikipedia.org/wiki/Precision_and_recall