International Journal of Research and Reviews in Information Sciences (IJRRIS)
Vol. 2, No. 1, March 2012, ISSN: 2046-6439
© Science Academy Publisher, United Kingdom
www.sciacademypublisher.com

A Survey on Question Answering System

Muthukrishnan Ramprasath[1] and Shanmugasundaram Hariharan[2]

[1]JJ College of Engineering and Technology, Department of Information Technology, Trichy, Tamil Nadu, India
[2]TRP Engineering College, Department of Computer Science and Engg, Trichy, Tamil Nadu, India

Email: mramprasath@gmail.com , mailtos.hariharan@gmail.com

Abstract – Question Answering (QA) is a specialized form of information retrieval. Given a collection of documents, a Question Answering system attempts to retrieve the right answers to questions posed in natural language. Generally question answering system (QAS) has three components such as question classification, information retrieval, and answer extraction. These components play a vital role in QAS. Question classification play primary role in QA system to classify the question based on the type of its entity. Information retrieval technique is take of identify success by extracting out relevant answer posted by their intelligent question answering system. Finally, answer extraction module is emerging topics in the QAS where these systems are often requiring ranking and validating a candidate's answer. This paper provides a brief discussion about different type of QAS and method has been used in these systems. In addition this article describes a new architecture for QAS which will be different from most QAS and its analyzer suitable QAS to find precise answer to user question.

Keywords – Question answering system, Classification, Information retrieval, Answer extraction

1. Introduction

The question answering system is very hot in natural language processing researchers can propose questions in natural language and get compact and relevant answers rather than relative web pages in the system. The goal of a question answering system is to retrieve answers to questions rather than full documents or best-matching passages, as most information retrieval systems currently do. Jibin Fu, Jinzhong Xu [35]. Eric Brill, Susan Dumais [36]

The amount of information on the web has developed exponentially over the years, with content covering almost any subject. As a result, when user looks for information, he/she is often confused by the vast quantity of results from search engines. Virtually all kinds of information are available on the World Wide Web (WWW) in one or another form. The number of web pages on the Internet increased tremendously and crossed 1 trillion landmark in 2008 which was only 200 billion in 2006 as reported in Wirken,[34], Alpert and Hajaj,[33]. Therefore, managing such a huge volume of data is not an easy task. Search engines like Google and Yahoo return links along with snippets to the documents for the user query Users browse the content carefully through a long list of outcome to look for a precise answer.

Question-answering (QA) research emerged as an attempt to tackle this information-overload problem. As mentioned above shortly QA systems are classified in two main parts: namely open domain QA system and closed domain QA system. Question which deals about nearly are everything and can only relies on universal ontology and information, such type systems are called as open domain question answering system. On the other hand, closed-domain question answering deals with questions under a specific domain (music, weather forecasting etc.) The domain specific QA system involves heavy use of natural language processing systems formalized by building a domain specific ontology.

QA research attempts to deal with a wide range of question types including: fact, list, definition, paragraph and cross-lingual questions. Search collections vary from small local document collections, internal organization documents to be complied with newswire reports on the World Wide Web. A QAS returns answer of a user question in concise form. In order to provide the specific answer, the system must know what precisely a user wants. The prior knowledge of the estimated answer type helps the QAS to extract correct and precise answers from the document collection.

A typical pipeline Question Answering System consists of three distinct phases: Question classification, information retrieval or document processing and answer extraction. Question classification is the first phase which classifies user questions, derives expected answer types, extracts keywords, and reformulates a question into semantically equivalent multiple questions. Reformulation of a query into similar meaning queries is also known as query expansion and it boosts up the recall of the information retrieval system. Information retrieval (IR) system recall is very important for question answering. If no correct answers are present in a document, no further processing could be carried out to find an answer. Precision and ranking of candidate passages can also affect question answering performance in the IR phase.

Answer extraction is a final component in question answering system, which is the tag of discrimination between question answering system and the usual sense of text retrieval system. Answer extraction technology becomes an influential and decisive factor on question answering system for the final results. Therefore, the answer extraction technology is deemed to be a module in the question answering system.

The rest of the paper organized as follows In Section II we discussed the proposed architecture of the question answering system, Section III discussed the related work of this article. Section IV discusses on question answering system based on information retrieval. Section V includes the discussion about proposed work on question answering system and final section VI presents the conclusion

2. Architecture of QA system

The architecture of QA system consists of several components as shown (Figure 1). The role and responsibility of each such as

1. Query interface is used to retrieve the question posted by the user.
2. Next the query analyzer phrase the question into subject, verb, object etc. it also used to improve the performance of the QA system.
3. Question classification used to identify type of the question after that the type of the answer will be specified.
4. Query reformulation play vital role in QA system because this component used to find the correct answer to user question
5. Search engine module which is used to retrieve the document based upon important keyword present in the question.
6. This component also focuses the pattern of the question to retrieve the relevant document.

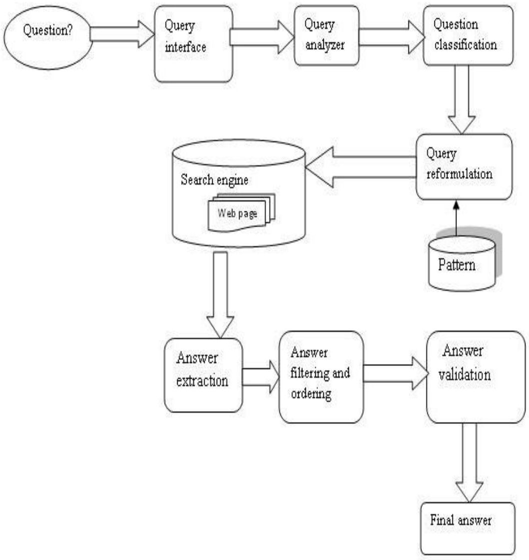

Figure 1. Question answering system architecture.

7. Search engine send candidate answers collection to next answer extraction module which extract candidate answers from retrieved documents.
8. Then these candidate answers pass to filtering and ordering unit.
9. Based on co-occurrence words and semantic relations existing in database ontology, answer type and keywords which extracting in question processing module, system filter candidate answers collection.
10. As a result some answers which are not related with the asked question will be eliminated. The answers with high priority show to user for validation. Then the answers get a validation answer grade and save it in usage knowledge. If the user accepts the suggested answer which system presented as an exact answer.

3. Related Work

General question answering system composed of three main modules: question classification, information retrieval and answer extraction. Yong-le sun [37] presented chinese question classification based on mining association rule which extract word and bi-gram from question as classic features. Kepei Zhang and Jieyu Zhao [38] presented a Chinese Question-Answering System with Question Classification which uses word, named entity, part of speech (POS) and semantics as a classic feature to classify the question. Santosh Kumar Ray [2] discussed some of the existing approaches for question classification and proposed a new method based on the usage of the Word Net. Wikipedia also tested their approach on TREC data set and achieve better classification accuracy which is comparable to earlier reported research works.

Svetlana Stoyanchev [1] presented a document retrieval experiment on a question answering system, and evaluates the use of named entities and of noun, verb, and prepositional phrases as exact match phrases in a document retrieval query. While [3] presented simplest approach to improve the accuracy of a question answering system might be restricting the domain it covered. Discussed Information retrieval mechanism considers the inverted index technique, which is widely used in the existing Information Retrieval (IR) field and showed in their performance study that RDBMS implementation almost always outperforms the IR implementations [4].

Paloma Moreda, Hector Llorens et al [5] presented two proposals for using semantic information in QAS, specifically in the answer extraction step. Its aim is to determine the improvement in performance of current QA systems, especially when dealing with common noun questions. Liang Yunjuan , Ma Lijuan, [6] discussed the design of dynamic knowledge-based full-text retrieval system, inverted index technology research and analysis, given some of indexing code, in order to improve the retrieval accuracy and to achieve a reasonable. The following table presents comparison about different types of question answering system and methods used in this system. Li Peng, Teng Wen-Da, Zheng Wei [7] presented formalized answer extraction method based on pattern learning and achieves a formalized template for automatic learning and obtaining. It avoids the defect of time consuming of manual development pattern, poor adaptability, low coverage etc. For correct answer extraction, some patterns should be defined for system to find exact type of answer and then sends to document processing. [8][9].

In answer Extraction approach the decision about the correctness of an answer is based on its entailment with a given support text. This way of answer selection not only allows assuring the rightness of answers but also their consistency with the snippets that will be showed to the users. This approach was suggested by Penas et al (2007), and has been implemented by Glöckner et al (2007) [10][11].

Anne R.Diekema,OzgurYilmazel, and ElizabethD. Liddy[35] examined the evaluation requirements of restricted domain systems and incorporates evaluation criteria identified by users of an operational QA system in the aerospace engineering domain. It demonstrates that user-centered task-based evaluations are required for restricted applicable to open domain systems.domain systems; these evaluations are found to be equal.

Table 1

S.NO	Types of question answering system	Method used in question answering system
1	A prototype Question Answering system using syntactic and semantic information for answer retrieval.	Tagging and chunking, Named Entity Recognition, Measuring Semantic distance
2	A Question Answering System Supported by Information Extraction	NE-Supported question answering, Question Processing, Text Processing, text matching
3	Analysis of the Asks Question-Answering System	Query Reformulation, N-Gram Mining, N-Gram Filtering, N-Gram Tiling
4	Multilingual Question/Answering	Tokenization and pos tagging., Word sense disambiguation, Answer type identification, Keywords expansion, Semantic Disambiguation
5	A specifiable-domain multilingual Question	
Answering architecture	Multilinguality,Spatial–temporal context awareness, Textual entailment	
6	A Question Answering System based on Information Retrieval and Validation	Expected Answer Type, Named Entities Presence, Acronym Checking
7	A Hybrid Question Answering System based on Information Retrieval and Answer Validation	Pattern Generation Module,Hypothesis Generation Module,Document Processing and Indexing

Jochen L. Leidner and Chris Callison-Burch[36] proposed a new evaluation schema that uses the insertion of answers from Frequently Asked Questions collections (FAQs) to measure the ability of a system to retrieve it from the corresponding question. Hyo-Jung Oh and Sung Hyon Myaeng[37] examine the roles and effects of the answer verification and weight boosting method, which is the main core of the automatically generated strategy-driven QA framework.

4. 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, Askme, Alta vista etc...).

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. Sang-Won Leeb, Hyoung-Joo Kim [15] stated that inverted index techniques has widely used in the information retrieval field. In order to support the containment queries for structured document such as XML, its need to be extended.

The number of document returned by the IR system huge means paragraph filtering concept has used to reduce the no of candidate document and to reduce the amount of candidate text from each document. The steps involved for QA system based on information retrieval is given below:

4.1.1. Filtering candidate document

The idea of paragraph filtering is based on the principle that the most relevant documents should contain the question keywords in a few neighboring paragraphs, rather than dispersed over the entire document. To exploit this idea, the location of the set of question keywords in each document is examined. If the keywords are all found in some set of N consecutive paragraphs, then that set of paragraphs will be returned, otherwise, the document is discarded from further processing. 'N' is again a configurable number that could be tuned based on an evaluation of system performance under different tolerances of keyword distance in documents.

4.1.2. Identifying quality of the document

To estimate the quality of the selected paragraph quality component has used. If the quality of paragraphs is deemed to be inadequate, then the system returns to the question keyword extraction module, and alters the heuristics for extracting keywords from the question. Then the IR can performed by using new set of key word retrieved from scratch. The cause of re-determining question keywords stems from having either too many or too few candidate paragraphs after paragraph filtering. In either case, new queries for the information retrieval system are created by revisiting the question keywords component, and either adding or dropping keywords. This feedback loop provides some form of retrieval context that ensures that only a 'reasonable' number of paragraphs are passed onto the Answer Processing module. Like many other parameters, exactly how many paragraphs constitute a 'reasonable' number should be configured, based on performance testing. Next paragraph ordering is to rank the paragraphs according to a plausibility degree of containing the correct answer.

4.1.3. Standard radix sort algorithm for paragraph ordering

This algorithm uses different scores to order the paragraph. The number of words from the question that are recognized in the same sequence within the current paragraph window, the number of words that separate the most distant keywords in the current paragraph window and the number of unmatched keywords in the current paragraph window. Paragraph window is defined as the smallest span of text required to capture each maximally inclusive set of question keywords within each paragraph. Radix sorting is performed for each paragraph window across all paragraphs.

4.1.4. *Term detection must satisfy the proximity (TP) restriction*

To improve precision results in IR systems by means of the incorporation of new terms to the query. In our work incorporate the dependency information between index terms by using the concept of Term Proximity (TP) information. Information retrieval system calculates the Term proximity (TP) between pairs of query terms, specifically between all possible combinations of query pairs. For example, in the query "**Letter gun for Kiesbauer"**, TP is calculated for the pairs ''gun – Bomb", "gun –Kiesbauer","gun – Kiesbauer''. This is a problem in long queries, in which there is not a clear dependency relation between some query terms. Therefore many researchers evaluate the proposal on short queries in order to reduce the number of possible query pair. This problem is also overcome in [19] Mitra, Buckley, Singhal, and Cardie by means of selecting only those query pairs that co-occur in the corpus at least 25 times. Rasolofo and Savoy [20] presented that considers the instance of each pair of terms in the document if they are within a maximal distance of five terms. Nevertheless, as the authors conclude, the weakest point of this work is the restriction of the maximal distance of the pair of terms, because it does not work for every pair of terms: its success depends on the terms, the query and the documents.

4.1.5. *Lexical and Syntactic Knowledge for IR*

In our suggestion we adopt the format of parsing the query to acquire the set of query terms to calculate the TP information, instead of calculating TP among all possible combinations of query pairs, but we vary from previous approaches in the following three points: first we do not carry out a full parsing of the query but chunking the queries into sets of simple phrases such as noun, prepositional phrases and sequences of verbs. For example, The query "A letter bomb from right-wing radicals sent to the black TV personality Arabella Kiesbauer in 1995" is segmented into the following five phrases: "[letter bomb]$_1$ from [right-wing radicals]$_2$ [sent]$_3$[black TV personality Arabella Kiesbauer]$_4$ in [1995]$_{5.}$In order to reach a more consistent behavior for different queries, we apply different TP measures depending on the lexical type of each query term. We apply TP measures to phrases as well as terms because phrases represent the concepts expressed in a text more accurately than single words.

Discourse representation theory has used in our system to deal with discourse anaphora. Grosz and sidner stated that any discourse is composed of three separate but interrelated parts [21]. Sentences are composed of phrases and phrases are composed of individual words in the linguistic structure. For that reason first we segment the documents into sentences that contain entities represented as phrases. Thus, we measure TP distances between entities and not between single terms or pairs of terms. There are several ways of interaction between these phrases, mainly through anaphora references that we consider in order to measure TP between entities. Example for our application: Assassination of Abraham Lincoln who killed Abraham Lincoln and why?

Pos-tagging (lemma; word; tag):

(assassination; Assassination; NN) ;(of; of; of); (abraham; Abraham; NP); (Lincoln; Lincoln; NP); (killed; Killed; VBD); (ybraham; Abraham; NP); (Lincoln; Lincoln; NP)

Chunking of phrases and sentences:

#1# Assassination, #2# Abraham Lincoln, #3# shot, #4# Abraham Lincoln,

Similar situations occur with the entities in the intentional and attentional structure of the discourse, such as the temporal location of the document, in which the distance between it and the remaining terms in the document are not so important. Finally, our algorithm concludes that no penalization is applied on the document, because although the document and the query do not share the same syntactic structure.

4.2. *Question Classification*

Question answering is an alternate of information retrieval, which retrieves detailed information rather than documents. A QA system takes a natural language question as input, transforms the question into a query and forwards it to an IR module. When a set of appropriate documents is retrieved, the QA system extracts an answer for this question. There are different ways of identifying answers. One of them makes use of a predefined set of entity classes. Given a particular question, the QA system classifies it into those classes based on the type of entity it is looking for, identifies entity instances in the documents, and selects the most likely one from all the entities with the same class as the question. There are different types of methods available for classify the question. In the following section we are going to discuss important technique for question classification. Such as identification of question pattern, semantic approach for question classification, sub tree kernel using support vector machine to improve the performance of the question classification

Patterns are used to identify the nature of the question posted by the user. Here we explain some of the pattern found in the experimental question database collection from UIUC Hovy et al [32]. Subsequently, we propose an question classification algorithm to classify questions using WordNet and Wikipedia. The question database consists of 5500 training and 500 test questions collected from English questions. The test questions have been collected from TREC 10 question dataset. All questions of the dataset have been physically labeled according to the coarse and fine grained categories. In the question database having eight main patterns: seven for standard 'Wh' questions and eight for other questions. Each of these patterns further consists of several sub-patterns.

Table 2

Coarse class	Fine classes
Abbreviation	Abbreviation, expression abbreviated
Entity	animal, body, color, creative, currency, diseases and medical, event, food, instrument, lang, letter, other, plant
Description	definition, description, manner, reason
Human	group, ind, title, description
Location	city, country, mountain, other, state
Numeric	code, count, date, distance, money, order, other, period, percent, speed, temp, size, weight

Functional Word Questions: All Non-Wh questions (except how) fall under the category of Functional Word Questions. These questions generally start with non-significant verb phrases.

Example: Name the Ranger who was always after Yogi Bear.

When Questions: When Questions start with ‘‘When” keyword and are temporal in nature. The general pattern for When Questions is When (do|does|did|AUX) NP VP X”, where AUX, NP, and VP auxiliary verbs, noun phrases, and Verb phrases. ‘|’ indicates Boolean OR operation and ‘X’ can be any combination of words playing insignificant role in answer type determination.

Example: When did Israel begin turning the Gaza Strip and Jericho over to the PLO?

Where Questions: ‘‘Where Questions” start with Where keyword and are related to the location. These may represent natural entities such as mountains, geographical boundaries, manmade locations such as temple, or some virtual location such as Internet or fictional place. The general pattern for Where Questions is Where (do|does|did| AUX) NP VP X?”

Example: Where is US?

Which Questions: The general pattern for Which Questions is Which NP X”? The expected answer type of such questions is decided by the entity type of the NP.

Example: Which company manufactures car parts?

Who/Whose/Whom Questions: Questions falling under this category have general pattern(Who|Whose|Whom) [do|does|did|AUX] [VP] [NP] X? Here [word] indicates the optionalpresence of the term word in the pattern. These questions generally ask about an individual or an organization.

Example: Who wrote ‘Hamlet’?

Why Questions: Why Questions always ask for certain reasons or explanations. The general pattern for Why Questions”\ is ‘‘Why [do|does|did|AUX] NP [VP] [NP]” X”.

Example: Why do heavier objects travel downhill faster?

How Question: ‘‘How Questions” have two types of patterns: ‘‘How [do/does/did/AUX] NP VP X?” or ‘‘How [big|fast|long|many|much|far] X?” For the first pattern, expected answer type is description of some process while second pattern returns some number as answer.

Example: How did the jack-o ’-lantern gets its name?

What Questions: What Questions have several types of patterns? The most general regular expression for What Questions can be written as ‘‘What [NP] [do/does/did/AUX] [functional-words] [NP] [VP] X? What Questions can ask for virtually anything.

Example: What is considered the costliest disaster for insurance industry? Many What Questions are disguised in the form of ‘‘Functional Word Questions”.

4.2.1. Semantic approach for question classification

In the field of question answering especially in query expansion and question classification processes, WordNet have been used as the online semantic sources. We propose an algorithm that combines the semantic features of the WordNet with the exhaustive and precise description of terms provided by Wikipedia. There are more than a few reasons to include Wikipedia in the question classification. First, it contains descriptions of a large number of term combinations. Secondly, in dissimilarity to WordNet, Wikipedia describes various actions that consist of compound Noun phrase. As a final point, articles on the Wikipedia are written by general users. This helps us to learn about what does a general user expects while using a particular word or any combination of words. Currently, we explain two procedures which will be used in this algorithm. These procedures use an entity type tree, a tree consisting of entity types as nodes. Initially entity tree has constructed with basic entity types and new entities are added as and when they are identified. This takes a word or a phrase as an input and uses online resources, Wikipedia and WordNet, to determine the type of expected entity. Procedure online passes the phrase or word to both WordNet and Wikipedia. Entity type from WordNet is selected from synonyms and direct hypernyms for the given word or phrase. The first paragraph returned by Wikipedia is analyzed for the presence of entities related to the word or phrase in the question. We observe that a typical article in Wikipedia starts like ‘‘X is a Y, Z,”. In general Y, Z are synonyms, hypernyms, hyponyms or some semantically related term to X and these are considered to be possible entity types. That is, if a sentence written in Wikipedia is ‘‘X is Y, Z, the procedure online takes Y, Z, as possible entity type of X.

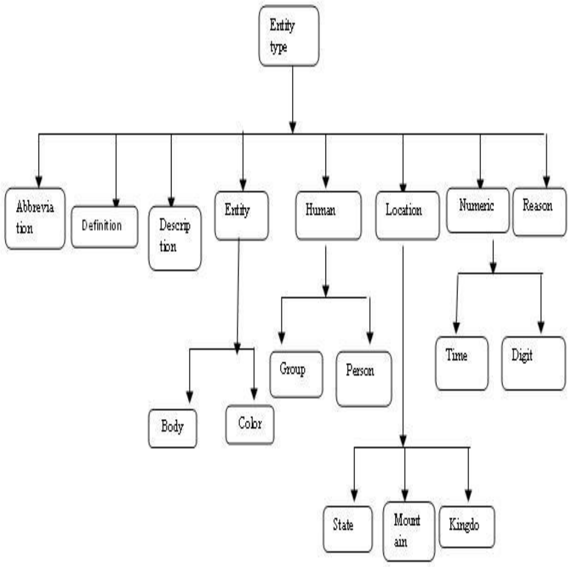

Figure 2. Demonstration of initial entity type tree.

The procedure online stores the entities returned by Wikipedia and WordNet in two sets TE1 and TE2 respectively. Elements in the set TE1 are stored in same order as mentioned in Wikipedia. Elements in the set TE2 (synonyms and hypernyms) are stored in the order of their frequency of occurrence as computed by WordNet.

The elements common in both sets are more likely to be candidate entity types. If the set (TE1\TE2) is not empty, the Procedure online searches the entity type tree for the presence of the elements in the set (TE1\TE2). If only one element of the set is present in entity type tree, it is returned as expected entity type. If more than one element is present in the entity type tree, the element at the lower level of the entity type tree is returned as expected entity type.

When we traverse the entity type tree from the root towards the leaves, the elements at lower level of the tree are representing more specific subclasses of the entity types than those at higher levels or the root and hence, more suitable for returning as expected entity type. If none of the elements in the common set is present in the entity type tree, the most frequently occurring hypernym of the first element in the common set is searched in the tree. If the hypernym is present in the entity type tree, the first element in the common set is added as child node of the hypernym node in the entity type tree. Otherwise, the first element in the common set is added as new entity type as a child entity of the root node. Proper nouns are not added to the tree.

The articles on the Wikipedia are written by general users in specific contexts, the entity type returned by Wikipedia is in general more relevant to those by WordNet. Therefore, in cases where entity types returned by Wikipedia and WordNet differ (i.e. the set (TE1\TE2) is empty), higher priority is assigned to the entity type returned by the Wikipedia. If the entities returned by Wikipedia (elements in set TE1) are present in the entity type tree, it is returned as expected answer type. In case of failure of Wikipedia, expected entity type is searched into the set returned by WordNet (elements in set TE2). If any match is found, it is returned as expected answer type otherwise a failure is reported to the calling procedure. Only Procedure online has direct access to the current entity type tree.

4.2.2. Sub set tree kernel using Support vector machine and language modeling for classification of question.

Vapnik [12] discussed Support vector machine based on the Structural Risk Minimization principle from statistical learning theory. The idea of structural risk minimization is to find a hypothesis h for which we can guarantee the lowest true error. The bounds are connected on the true error with the margin of separating hyper planes. In their basic form support vector machines find the hyper plane that separates the training data with maximum margin. We li [11] presented language modeling for question classification is to determine the question type based on the sentence structure and key words, which represent syntactic and semantic information respectively. A set of patterns are defined and hard-coded, often with regular expressions. When a new question comes, it is matched against those patterns to find the class it belongs to. As the pattern set gets more complete and accurate, the performance of this approach will become better. Language modeling techniques used to make the classification process more dynamic ans automatic. J. M. Ponte and W. B. Croft, [13] Presented a statistical approach that has gained much attention recently in the information retrieval area.

One of the most difficult tasks on applying machine learning for question classification is the feature design. Feature should represent data in a way that allows learning algorithm to separate positive from negative examples. In SVMs, features are used to build the vector representation of data examples and the scalar product between example pairs quantifiers how much they are similar. Instead of encoding data in the features vectors, kernel functions can be designed that provide such similarity between example pairs without using an explicit feature representation [14]. The kernel we considered in this paper represents trees in terms of their substructure. Such fragments define the feature space which, in turn, is mapped into a vector space.

The kernel function measures the similarity between trees by counting the number of common fragments. These functions have to recognize if a common tree subpart belongs to the feature space that we intended to generate. Here we considered Subset Trees (SSTs). A Sub Tree is defined as any node of a tree along with all its descendants. A Subset Tree is more general structure which not necessarily includes all the descendants. The only restriction is that an SST must be generated by applying the same generated the original tree [14].Tree kernel mainly used to compute the number of common sub structure between two trees $T1$and $T2$ without explicitly considering the whole fragment space.

Subset tree kernel using Support vector machine question classification method based on the use of linguistic knowledge and machine learning approaches and it exploit different classification features and combination method, also. Though among all the experiments one or two data set didn’t provide distinguishable hyperplane in every cases thereafter the outcome of experiments done using the tool SVM light on Li and Roth question classification data sets demonstrate some optimal set of values. The draw backs of this method is it cannot applicable if we are using different parameter in this subset tree kernel method.

4.3. Answer extraction

Answer extraction is a sub-area of question answering system, which is the tag of bias between question answering system and the usual sense of text retrieval system. Answer extraction technology becomes an important and decisive factor on question answering system for the final results.The feature based methods of sorting become main stream of answer extraction technology in recent years, for instance neural network [22], maximum entropy [23], SVM [24], logistic regression [25] etc. Nevertheless, there is no semantic feature adding together into the quality system because of the slow development of semantic processing of natural language processing technology. This makes the outcome of answer extraction which based on feature reach a bottleneck. How to improve the correctness of answer extraction under existing technology is the on of the issue of question answering system.

In order to reduce this problem new alternative approaches known as meta QA systems and multi-stream QA systems has been proposed .Meta-stream QA systems internally combine several components or techniques at each QA process. For instance Pizzato and Molla-Aliod[26] describe a QA architecture that uses several document retrieval methods, and Chu-carroll, Czuba, Prager, and Ittycheriah [27] present a QA system that applies two different components at each process. On the other hand, multi-stream QA approaches go a step forward by achieving a superficial combination of several QA systems. Still, in this case, they are primarily focused on selecting the correct answer for a given question rather than on ranking all candidate answers. Subsequently we introduce the traditional approaches for multi-stream QA.

There are two approaches namely system centered approach and answer centered approach. Answer-centered approaches, which select the final answer exclusively based on its regularity of occurrence across streams. Answer centered approach can be divided in to three types such as answer chorus approach, hybrid approach and wed chorus approach. Answer centered approach which relies on the answer redundancies it selects as the final respond the answer with the highest frequency across streams.

Some systems based on this approach are described in de Chalendar et al. [29] Hybrid approach which considers the combination of criteria from the system and answer-centered approaches. The method described in Jijkoun and de Rijke [30] is an example of this kind of approach. It uses the system's confidences to differentiate between answers having the same frequency of occurrence. Its evaluation results indicated that this combination could outperform the results obtained by other multi-stream QA systems based on one single strategy. Web Chorus approach uses information from the Web to evaluate the relevance of candidate answers. It selects the answer with the greatest number of Web pages containing the answer terms along with the question terms. It was proposed by Magnini, Negri, Prevete, [31] and Tanev, and subsequently it was evaluated in Jijkoun and de Rijke[29].

4.3.1. Architecture for Multi-stream question answering system

In addition to the traditional approaches adopted from IR, more recently emerged a new type of multi-stream QA approach based on the application of textual entailment recognition (RTE) techniques. The idea behind this approach is to decide about the correctness of answers based on their textual entailment with a given support text therefore, using these decisions to identify the more appropriate answer for the question at hand. The following diagram shows general scheme of the proposed multi-stream QA approach. It consists of two main stages. In the first stage, called QA stage, several QA systems extract in parallel a candidate answer and its corresponding support text for a given question. Then, in the second stage, called selection stage, a classifier evaluates all candidate answers and assigns to each of them a category (correct or incorrect) as well as a confidence value (ranging from 0 to 1). At the end, the correct answer having the greatest confidence value is selected as the final response. In the case that all answers were classified as incorrect, the system returns a nil response.

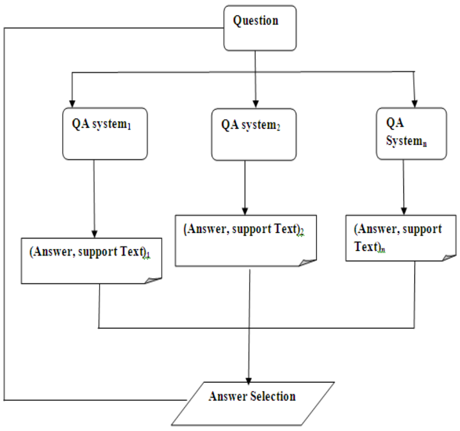

Figure 3. General diagram of multi-stream question answering system.

4.3.2. Answer selection

This phase focuses on the choice for the final answer by combining evidence from the replies of different input QA systems. In particular, the classification of the candidate answers (in correct and incorrect categories) as well as the estimation of their confidence values is carried out by means of a supervised learning approach that considers three main processes: preprocessing, feature extraction and answer classification. Preprocessing process used to extract the main content element from question answer and support text which will be subsequently used for deciding about the correctness of the answer.

Feature extraction stage gathers a set of processes that allow extracting several features from the question, answer and support text. These features can be categorized in two groups: those that indicate characteristics of the question and the answer as well as their relation, and those that measure the entailment relation between the question–answer pair and the support text.

Answer classification process determines if each candidate answer is correct or incorrect, and also estimates a classification confidence for each of them. The confidence values help to select the final response in situations where several candidate answers are classified as correct; in such cases the answer with the greatest confidence is selected as the final response. Finally, if we do any improvement in answer extraction module will directly impact the performance of the multi-stream method.

5. Proposed work on question answering system

In question answering (QA), users could use sentences in everyday life to rise questions and the system will return answers to users directly after analyzing and comprehending these questions. Therefore, the question answering system better satisfies the users' requirements. It can be said that question answering is a new generation of intellects search engine. General question answering system consists of different component such as question classification, information retrieval and answer extraction.

Question classification is the first phase of the QA system. This section discussed about many approaches used to classify the question posted by the user. Specifically language modeling for question classification is a probabilistic approach imported from IR systems. The models are constructed in a more flexible way and built two types of model a linear combination of unigram and bigram models with an absolute-discount smoothing technique. Next sub set tree kernel using Support vector machine for classification of question will used to maximize the performance level of system.

In next phase discussed about information retrieval method by the incorporation of the lexical and syntactic knowledge generated by a POS-tagger and a syntactic Chunker. It is based on theories of discourse structure, in which documents and sentences are segmented into sentences and entities. We adopt the scheme of parsing the query to obtain the set of query terms to calculate the TP information, instead of calculating TP between all possible combinations of query pairs. The tested result shows that it can applicable only shorter document and shorter collection.

Answer extraction phase discussed about the multi-stream question answering method supported by supervised learing approach. Finally, it is obvious that any development in the answer extraction module will directly impact the performance of the proposed multi-stream method. The proposed architecture overcomes this problem. In particular we plan to consider some new features in the entailment recognition process. We plan to include some additional discriminative features that allow describing with more detail the overlap between the question answer pair with the support text.

The entire component used in the question answering system having problem with the method used in those level such as classification, information retrieval and answer extraction. As a result the proposed architecture presented in this article will overcome the imperfection by using query interface, query analyzer, question classification, query reformulation, search engine, answer extraction, answer validation modules used and return the precise answer to the user.

6. Conclusion

The goal of a question answering system is to retrieve answers to questions rather than full documents or best-matching passages, as most information retrieval systems. In this paper we discussed some of the approaches used in the existing QA system and proposed a new architecture for QA system to retrieve the exact answer. In the classification module we have discussed identification of question pattern, classification of question base on semantic approach, performance maximization using sub tree kernel method using support vector matching and language modeling for question classification.

Information retrieval module discussed the scheme of parsing the query to obtain the set of query terms to calculate the TP information, instead of calculating TP between all possible combinations of query pairs. In answer extraction module discussed multi-stream QA method supported on supervised learning approach. It is used to combine the output of different question answering system to produce the better answer to user. Although our method takes advantage of the redundancy of answer across stream and allowed significantly reduce the number of in correct answer presented to the user.

Acknowledgment

The authors would like to express their thanks to the Chairman, Principal and faculty of Information Technology Department, J.J.College of Engineering and Technology, trichy, Tamilnadu, India and TRP Engineering College for the environment provided.

References

[1] Svetlana Stoyanchev, and Young Chol Song, and William Lahti, “Exact Phrases in Information Retrieval for Question Answering”, Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering (IR4QA), pages 9–16 Manchester, UK. August 2008”

[2] Santosh Kumar Ray a,*, Shailendra Singh b, B.P. Joshi c, “A semantic approach for question classification using WordNet and Wikipedia” 0167-8655/$ - see front matter _ 2010 Elsevier B.V. All rights reserved. doi:10.1016/j.patrec.2010.06.012

[3] Mohammad Reza Kangavari, Samira Ghandchi, Manak Golpour,” Information Retrieval : Improving Question Answering Systems by Query Reformulation and Answer Validation” World Academy of Science, Engineering and Technology 48 2008

[4] Chiyoung Seoa,*, Sang-Won Leeb, Hyoung-Joo Kima” An efficient inverted index technique for XML documents using RDBMS” Information and Software Technology 45 (2003) 11–22

[5] Paloma Moreda *, Hector Llorens, Estela Saquete, Manuel Palomar” Combining semantic information in question answering systems” Information Processing and Management 47 (2011) 870–885

[6] Liang Yunjuan , Ma Lijuan, Zhang Lijun, Miao Qinglin” Research and Application of Information Retrieval Techniques in Intelligent Question Answering System” 978-1-61284-840-2/11/$26.00 ©2011 IEEE.

[7] Li Peng, Teng Wen-Da, Zheng Wei, Zhang Kai-Hui “Formalized Answer Extraction Technology Based on Pattern Learning”, IFOST 2010 Proceedings.

[8]] Figueira, H. Martins, A. Mendes, A. Mendes, P. Pinto, C. Vidal, D,Priberam's “Question Answering System in a Cross- Language Environment",LECTURE NOTES IN COMPUTER SCIENCE, Volume 4730, 2007,PP. 300-309.

[9] Dan Moldovan, Sanda Harabagiu, Marius Pasca, Roxana Girgu, " The Structure and Performance of an Open-domain Question Answering System", Proceedings of the 38th Annual Meeting on Association for Computational Linguistics Hon Kong, 2000, PP. 563-570,

[10] Peñas A, Rodrigo Á, Sama V, Verdejo F (2007) Testing the reasoning for question answering validation. Journal of Logic and Computation (3), DOI 10.1093/logcom/ exm072 [

[11] Wei Li “Question lassification Using Language Modeling”Center for Intelligent Information Retrieval

[12] V. Vapnik, “Statistical Learning Theory”, Wiley, New York, USA 1998.

[13] J. M. Ponte and W. B. Croft, “A Language Modeling Approach to Information Retrieval”.

[14] Roberto Basili, Alessandro Moschitti, “Automatic Text Categorization From information Retrival to Support Vector Learning”, ARACNE editrice,November 2005. [15] M. Collins, N. Duffy, “New Ranking algorithm forparsing and Tagging: Kernels over Distance structure, and the Voted Perception”, Association for Computational Linguistics (ACL), Philadelphia, USA, 2002.

[15] Chiyoung Seoa,*, Sang-Won Leeb, Hyoung-Joo Kima ,” An efficient inverted index technique for XML documents using RDBMS” Information and Software Technology 45 (2003) 11–22

[16] Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval (pp. 21–29).

[17] Roberston, S. E., Walker, S., & Beaulieu, M. (2000). Okapi at TREC. Information Processing and Management, 36(1), 95–108.Santana, O., Carreras, F. J., Hernández, Z., & Gonzalez, A. (2007). Integration of an XML electronic dictionary with linguistic tools for natural languageprocessing. Information Processing and Management, 43, 946–957.

[18] Amati, G., Carpineto, C., & Romano, G. (2004). Comparing weighting models for monolingual information retrieval. In Comparative evaluation of multilingual information access systems. Lecture notes in computer science (Vol. 3237, pp. 310–318).

[19] Mitra, M., Buckley, C., Singhal, A., & Cardie, C. (1997). An analysis of statistical and syntactic phrases. In Proceedings of RIAO-97 (pp. 200–214).Mittendorfer, M., & Winiwarter, W. (2002). Exploiting syntactic analysis of queries for information retrieval. Data & Knowledge Engineering, 42, 315–325.

[20] Rasolofo, Y., & Savoy, J. (2003). Term proximity scoring for keyword-based retrieval systems. In Proceedings of the 25th European conference on IR research (ECIR 2003). Lecture notes in computer science (Vol. 2633, pp. 207–218).

[21] Grosz, B., & Sidner, C. (1986). Attentions, intentions and the structure of discourse. Computational Linguistics, 12(3), 175–204. Hirst, G. (1981). Anaphora in natural language understanding. Berlin: Springer-Verlag

[22] Marius A Pasca. High performance, open-domain question answering from large text collections[D]. USA; University of Southern Methodist, 2001.

[23] Abraham Ittycheriah. Trainable question answering systems[D]. USA: The State Univerity of New Jersey, 2001

[24] Jun Suzuki, Yutaka Sasaki, Eisaku Maeda. SVM answer selection for open-domain question answering[A].l9th International Conference on Computational Linguistics (Coling-2002) [C] Taipei: Howard International House, 2002. 974- 980.

[25] Peng Li, Yi Guan, Xiao-Iong Wang. Answer extraction based on system similarity model and stratified sampling logistic regression in rare data [J].International Journal of Computer Science and Network Security, 2006,6(3):189-196

[26] Pizzato, L. A. S., & Molla-Aliod, D. (2005). Extracting exact answers using a meta question answering system. In Proceedings of the Australasian language technology workshop. Sydney, Australia (pp. 105–112).

[27] Chu-carroll, J., Czuba, K., Prager, J., & Ittycheriah, A. (2003). In question answering, two heads are better than one. In In Human language technology conference of the North American chapter of the Association for Computational Linguistics (HLT-NAACL) (pp. 24–31).

[28] Clarke, C. L. A., Cormack, G. V., Kemkes, G., Laszlo, M., Lynam, T. R., Terra, E. L., et al. (2002). Statistical selection of exact answers (multitext experiments for TREC 2002). In Text retrieval conference (TREC) TREC 2002 proceedings. Department of Commerce, National Institute of Standards and Technology.

[29] De Chalendar, G., Dalmas, T., Elkateb-Gara, F., Ferret, O., Grau, B., Hurault-Plantet, M., et al. (2002). The question answering system QALC at LIMSI, experiments in using web and wordnet. In Text retrieval conference (TREC) TREC 2002 proceedings. Department of Commerce, National Institute of Standards and Technology.

[30] Jijkoun, V., & de Rijke, M. (2004). Answer selection in a multi-stream open domain question answering system. In S. McDonald & J. Tait (Eds.), ECIR. Lecture notes in computer science (Vol. 2997, pp. 99–111). Springer

[31] Magnini, B., Negri, M., Prevete, R., & Tanev, H. (2001). Is it the right answer? Exploiting web redundancy for answer validation. In ACL '02: Proceedings of the 40th annual meeting on Association for Computational Linguistics (pp. 425–432).

[32] Hovy, E., Gerber, L., Hermjakob, U., Lin, C.Y., Ravichandran, D., 2001. Towards semantics-based answer pinpointing. In: Proc. First Internat. Conf. Human Language Technology Research. Association for Computational Linguistics, Morristown, USA, pp. 1–7.

[33] Alpert, J., Hajaj, N., 2008. We knew the web was big. 7/25/2008. <http:// googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>. CIA the world Factbook, <http://www.cia.gov/library/publications/the-worldfactbook/>.

[34] Wirken, D., 2006. The Google Goal of Indexing100 Bil lion Web Pages.<www.sitepronews.com/archives/2006/sep/20.html>. WordNet. <http://wordnet.princeton.edu>.

[35] Jibin Fu, Jinzhong Xu(3), "Domain Ontology Based Automatic Question Answering"School of Computer Science & Technology Beijing Institute of Technology Beijing, China fujibin@gmail.com, Keliang Jia(2) School of Information Management ShanDong Economic University JiNan

[36] Eric Brill, Susan Dumais and Michele Banko" Ask MSR question answering system". Microsoft Research One Microsoft Way Redmond, Wa. 98052

[37] Hyo-Jung Oh a,⇑, Sung Hyon Myaeng b, Myung-Gil Jang" Effects of answer weight boosting in strategy-driven question answering" Information Processing and Management xxx (2011)