A New Architecture of Automatic Question Answering System using Ontology

Pawan Kumar
Research Scholar

Raj Kumar Goel
Assistant Professor
Noida Institute of Engineering & Technology, Gr. Noida (UP)

Prem Sagar Sharma
Assistant Professor
J.P. Institute of Engineering & Technology, Meerut (UP)

ABSTRACT
Question answering system (QAS) consists of three modules question processing, document processing and answer processing. Question processing is the main module of QAS. If it doesn't work properly, it causes problem for the whole system. A further answer processing module is a spring up the topic in Question Answering. These systems are often required to rank and validate candidate answers. These modules are finding the exact answer of given query by the user. In this paper, we have discussed a new model for question answering system which is find exact answer and improved their module.

Keywords
Question processing, Document processing, Answer processing, Sentence extraction algorithm, NER algorithm, and Question generation algorithm.

1. INTRODUCTION
In recent years, the growing popularity of web services user wants to get the relevant information about their questions. An existing information retrieval system or search engine can search few keywords, i.e. document retrieval; which gives only relevant documents that contain the keyword. Generally user wants a rigorous answer to his question. For example, "Who is the first prime minister of India? ", users only wants their answer in simple words as "Jawahar Lal Nehru", but not to read through a lot of documents that contain the words, **i.e.** "first", "India" and "prime minister". A question answering system provides an exact answer to the user questions by consulting its knowledge base. Therefore, an automated system converts the documents into textual answers. Earlier, QAS provides only simple answers of the user's questions. Therefore, the research has been focused on complex questions to get better output. Question processing has two main components, question classification and question reformulation. Question classification specifies the questions and answers. Question reformulation converts user's proper or improper questions into natural questions and specify in its related domain. Answer processing module has two main component first is answer filtering and the second is the answer validation.

2. SYSTEM
QAS is an upgraded form of FAQ (frequently asked question). FAQ is used for predefined database where as QAS is used to web document for finding the exact answer of a given query. In QAS questions are classified into two parts Factual Questions and Definition Questions. Factual questions are those which provide accurate information about any event. Questions starting from such as who, when comes under this category. [2] Definition questions or feedback questions that need an understanding of the subject question. Questions contains why and how comes this category. QAS are classified in two main domain Open domain QAS and Close domain QAS. Open domain QAS deals with almost everything. It generally relies on the ontology and the knowledge of the world. Closed domain QAS deals with questions under a specific domain (for example medicine or weather forecasting etc).which seems to be an easier task. QAS produces small, accurate and helpful answer of natural language questions. The goal of QAS is to find the exact and correct answer for given query. Analysis of questions, searching and choosing answers are three important parts in a QAS.

2.1 QUESTION PROCESSING
It consists of question representation, derivation of expected answers and keyword extraction. It has two main components, Question classification and Question reformulation.

2.1.1 Question Classification:-
One part of the question processing stage in question classification. Question classification will be prior to reformulation. This is for retrieve types of questions and answers. QA system first should know type of question. It also helps system to eliminate the question in final type of answer.[10] Table 1 shows question words, type of questions and answers. Totally questions can be divided into two parts first is question with 'WH' question words such as what**,** where, who, whom, which, how, why and etc. and second is questions with 'modal' or 'auxiliary' verbs that their answers are Yes/No. For correct answer extraction, some patterns should be defined for system to find exact type of answer and then sends to document processing. [4], [5]

QUESTION CLASSIFICATION	NAMED ENTITY	EXAMPLE
When	DATE/TIME	When was Jawaharlal Nehru born?
Which	LOCATION	Which city has minimum temperature?
	PERSON	Which person did invent the instrument of aero logy?
Why	REASON	Why don't we have enough rain this year?
Who	PERSON	Who is the president of India?
What	NUMBER	What is temperature of Delhi?
	DATE	What year do we have max rain?
	LOCATION	What is the capital of India?

Table 1 Question classification and answer [2]

2.1.2 Question Reformulation:

It tries to identify various ways of expressing an answer given a natural language question. It is used in Question Answering system to retrieve answer in a large document collection. The question reformulation converts the question into set of questions that will be sent to the search engine for parallel evaluation. In question reformulation we use of syntax relation among words of asked question sentence. We use of semantic relations among words of asked question sentence. We use of existing information of previous asked and answers in which a part of totally in same to user's asked question. In this case, system can use type of previous answer for new asked question. It causes that the process of finding proper pattern and type of answer become shorter and reduces the necessary time for submitting correct answer. [8] It would be possible if the system has the ability of saving information in 'knowledge base' database. When a user asks a question, first sentence parses to its syntax components and then its keywords are selected to use in reformulation. [9]

3. ANSWER PROCESSING

After the question type has been identified, the system extracts all such type information from the web documents as plausible answers, using named entity recognizer. Answer processing has two main components are Answer extraction and Answer validation. [7] Answer extract from web documents which are retrieving by search engine in answer extraction module. After answer extraction we validate answer with filtering and ranking user answers and final system's suggested answers with user voting. We approach to automatic answer validation relies on discovering relations between asked question and answer by mining the web documents or specific domain ontology. Once the candidate answers are selected, all those answers will be listed to the user. Now user can select the most appropriate answer which is best suiting to the question. This will be recorded in the question answer database in the rank column.

4. SYSTEM ARCHITECTURE

In user interface user write his question by an interface. In domain specific ontology it is used as dictionary and contains all words that are in related domains. In this database ontology questions and answers are surveyed semantically. Semantic relation among keywords saved in this database. Domain information is saved in this part and will submit the user's answer when a web services connects to internet. Question classification is the one of the most important function of most QA systems. In this part all questions are classified regarding WH question words or other question words with Yes/No answer. In question reformulation we change main question with using rules changes to question with new format. Usage Knowledge is the one of the most useful ways for finding answers of question is to be used library of the previous question and answer. In the question filtering the candidate answers will be filtered based on question type and answer type which was created in system. Answer Validation is the human assessment role which checks the correctness of answer and fills the validation grade in usage knowledge for the next validating which affect on timeline.

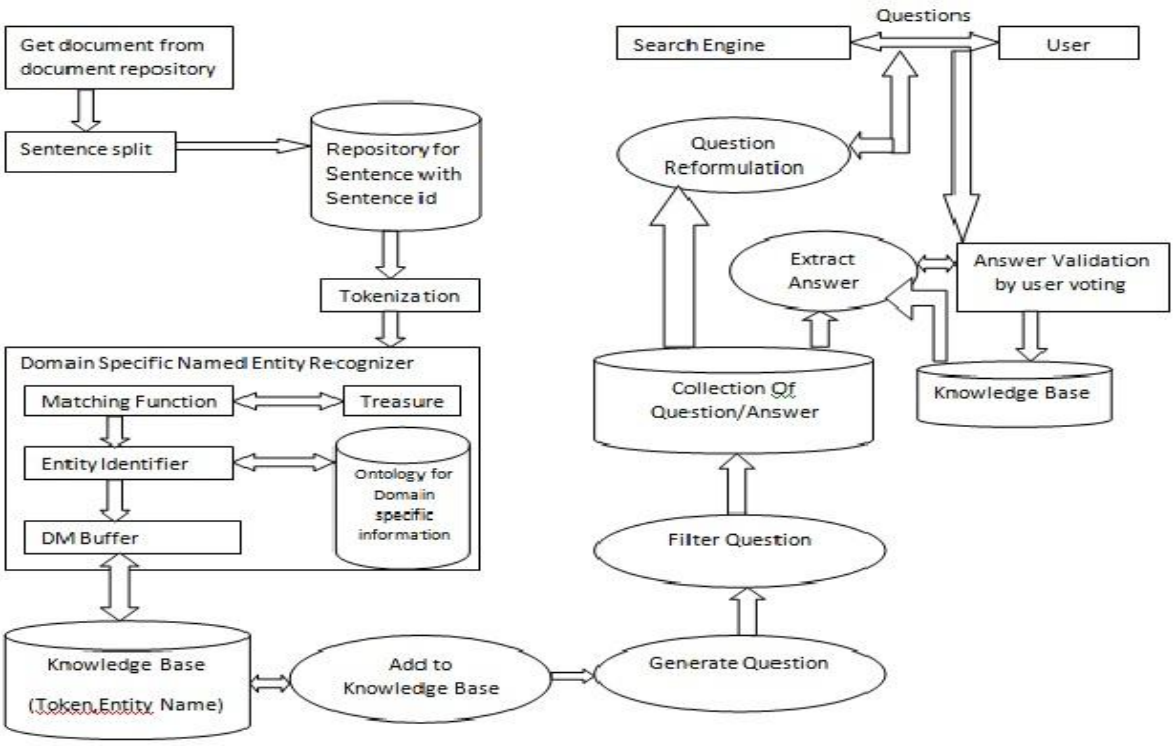

Fig 1.The System Architecture

Sentence split get the documents from the document repository. Generate the sentence id from the page of web documents and store the index with sentence id of the entire sentence in a page. Repository for the sentence with sentence id uses a repository for storing the sentence id of the pages and store the reference of those pages**.**

4.1 Sentence Split Algorithm

S_S ( )

Do

1. [Extract web document(WP) one by one from web repository(WR) while it is not empty]
   While (WR[i]!=NULL)
   WP=WR[i];
2. [Convert the all web pages in the text documents]
   Textdoc=WP;
3. [Extract the sentence from text documents and store in sentence repository(SR) with sentence id]
   SR [j+1] = Textdoc;
4. [Token create for sentence and it store token repository(TR)]
   TR [k+1] = Token (SR);
5. [Remove all end keywords from token repository]
   While (TR[i] !=NULL)
   Token=TR[i];
   If (endkeyword (token) ==true)
   Remove (token);

End

Tokenization creates the token from the sentence from the sentence id. That token will be use for the name entity recognization. Domain Specific Named Entity Recognizer multiple module Matching function, Treasure, Entity identifier, Ontology for domain specified, Temp buffer.

4.2 Named Entity Recognizer Algorithm:-

NER ( )
{

1. [Extract token from token repository]
   While (TR != NULL)
   Token = TR[i];
2. [Each token match from matching function in treasure and return base token]
   T = MF (token, treasure);
3. [Identify the entity by NER using ontology in entity identifier(EI)]
   If (T!=NULL)
   R =EI (T, ontology);
   Else
   R = EI (token, ontology);
4. [Return the Named Entity with the token id and stored in knowledge base (KB)]
   KB (token, entity);

}

Matching function is use for the match user's word from the stored documents and it will change the word in base word and stored in treasure. Treasure is use for the store the base word with its similar word. Entity identifier is use for identify the entity, which is belongs from which entity like location; person etc. and its store the entity in knowledge base. Ontology is use for the specific domain which identify from the entity. If entity is identify from person entity the ontology use the ontology of person entity. The last module is the temp buffer and it is use for the store the entity with token and if user sends the previous queries then the temp buffer send question generation for saving retrieving time. Temp buffer send the entity with token in knowledge base. Knowledge base is use the database for store the document id with the

token. Question generation generate the multiple question of given question from entity identifier. Question filtering filter the question generation which is use off keywords like is, am, are, was etc. Collection of question answers store the questions and answers from previous phase. Answer extraction extract the answer from the knowledge base and give to user. Answer validation by user voting use take the feedback of the user for their question and give the correct answer and it is useful or not. Validated answer stored in the knowledge base. Question reformulation creates the question in the new format.

5. CONCLUSION

We improve the accuracy of a question answering system might be restricting the domain it covered. By restricting the question domain, the size of user answer collection becomes smaller. Question reformulation is a main part for understanding the interplay of information retrieval. There are three steps in patterns for reformulation by user: format, contents and source. The main aim of rewriting question is asking question in another new format by user in which less time sources are used for search. In addition the question processing module, the improvement of answer processing module will be complete the question processing task in efficiency of the QA system, because the system must return exact answer. We find the exact answer of the user's query. We create the multiple questions of given word or given question.

6. REFERENCES

[1] Demner-Fushman, Dina, "Complex Qestion Answering Based on semantic Domain Model of Clinical Medicine", OCLC's Experimental Thesis Catalog, College Park, Md.: University of Maryland (United States), 2006.

[2] Doan-Nguyen Hai, Leila Kosseim, "The Problem Of Precision in Restricted Domain Question Answering. Some Proposed Methods of Improvement", In Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains, Barcelona, Spain, Publisher of Association for Computational Linguistics, July 2004, PP.8-15.

[3] Green, W.Chomky, C., Laugherty, K.BASEBALL: "An automatic question answer". Proceeding of the western Joint Computer Confrence, 1961, PP. 219-224.

[4] Figueira, h. Martins, 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.

[5] Dan Moldovan, Sanda Harabagui, Marius Pasca, Roxana Girgu, "The Structure and Perfomance of an Open Domain Question Answering System", Proceedings of the 38th Annual Meeting on Association for Computational Linguistics Hon kong, 2000, PP.563-570.

[6] Cody kwok, Oren Etzioni, Daniel S. Weld, "Scaling Question answering to the web", Proceedings of the 10th international conference on World Wide Web, Hon Kong, 2001, PP. 150-161.

[7] ADL Technical Team, "Content Object Repository Discovery and Registration/Resolution Architecture", ADL 1'st International Plugfest, June 2004.

[8] Rohini Srihari, Wei Li, "A Question Answering System Supported by Information Extraction". Proceedings of 6th conference on applied natural language processing, 2000, PP. 166-172

[9] Erik F. Tjong Kim Sang, Fien De Meulder. "Inroduction to CoNLL-2003 Shared Task: language –independent named entity recognition", Proceedings of th 7th conference on Natural language learning at HLT-NAACL 2003, 2003, PP. 142-147.

[10] R. B.-Y. a. B. Ribeiro-Neto, Modern Information Retrieval: Addison Wesley, 1999.

[11] P. Kumar, Domain Specific Named Entity Recognizer (DSNER) from Web Document International Journal of Computer Applications (0975 – 8887) Volume 86 – No 18, January 2014