A FRAMEWORK FOR RESTRICTED DOMAIN QUESTION ANSWERING SYSTEM

Payal Biswas, Aditi Sharan, Nidhi Malik
Jawaharlal Nehru University
New Delhi-110067
Email id: payal.biswas786@gmail.com

Abstract— **This paper proposes a framework for developing Question Answering System for restricted domain using advanced NLP tools. The proposed model basically works over the concept of Information Extraction rather than the old technique of information Retrieval used by the search engines. The main objective of the model is to extract the exact and precise answer for the given question from a large dataset. This framework is simple and easy to implement against the previously developed complex architectures. The Framework is divided into four modules namely: Question Processing Module, Document Processing Module, Paragraph extraction module and Answer extraction module. The paper also proposes various algorithms separately for Definition Type, Descriptive Type and Factoid Type of questions for extracting most potential answer from the large dataset.**

Keywords—**Question Answering System, Question Processing Module, Document Processing Module, Paragraph Extraction Module, Answer Extraction Module, Definition Type Question, Descriptive Type Question, Factoid Type Question.**

I. INTRODUCTION

Now a days there are lots of search engines available having many remarkable capabilities but the problem with these search engines is that they do not have deduction capabilities. That is instead of giving a straight forward, accurate and precise answer to the user's query or question they usually provide the links or URL's of those websites which might contain the answer of that question.
Although the set of documents which are retrieved by the search engine contain a lot of information about the search topic but it may or may not contain exactly that information which the user is looking for [1]. The basic idea behind the QUESTION ANSWERING SYSTEM is that the users just have to enter the question and the system will retrieve the most appropriate and precise answer for that question and return it to the user. Hence in those cases where the user is looking for a short and precise answer, QUESTION ANSWERING System plays a great role rather than Search Engines which usually provide a large set of links of those web pages which might contain the answer of that question.
For example, for the following set of questions such as:

" Which vitamin is present in Citrus Fruits ".
" What is the birth place of Barak Obama ? "
" Who is the first President of India ? "
" What is the height of Mount Everest ? "

Users are more interested in the answers such as ***Vitamin C, Honolulu Hawaii, Rajendra Prasad,*** and ***8848 m*** rather than a large document where the user has to go through to find out the exact answer for which he or she is actually looking for.

Therefore to resolve this problem at present the QUESTION ANSWERING SYSTEM has gained the interest of great researchers working in various fields such as Web Mining, Natural Language Processing, Information retrieval and information extraction etc. Question Answering System can be define as: "A system, whose main objective is to determine WHO did WHAT to WHOM, WHERE, WHEN, HOW and WHY?" [2]. That is a system which is capable of answering all the WHO, WHAT, WHOM, WHERE, WHEN, HOW and WHY type questions of the user?

The basic difference between the generally used search engines and the new concept of QA system is that the former works over the concept of Information Retrieval, while the QA system works over the concept of Information Extraction whose aim is not just to retrieve the relevant documents from large dataset alike search engines, but to retrieve the useful part of information from those relevant documents.

Question Answering Systems can be classified on the basis of the domains over which it has been constructed. For example:

1) Closed Domain QA System
2) Open Domain QA System.
3) Restricted Domain QA System

Close domain question answering systems deal with questions in a specific domain [3] having limited amount of focused and structured information. In this case of deep reasoning is possible but due to the very small size of data set they are not more than a "Toy Systems"[4] while Open domain question answering systems are domain independent having large collection of data from various fields but here deep reasoning is not possible [3].

Since both the Open Domain QA System and Close Domain QA System have their own pros and cons a new concept of Question Answering has been coined by Molla & Vicedo [4] called RESTRICTED DOMAIN QA SYSTEM which is the midway of these two domains.

Research in restricted-domain question answering (RDQA) addresses problems related to the incorporation of domain-specific information into current state-of-the-art QA technology with the hope of achieving deep reasoning capabilities and reliable accuracy performance in real world applications. In fact, as a not too-long-term vision, we are convinced that research in restricted domains will drive the convergence between structured knowledge-based and free text-based question answering.

During our study of QA System we found that although there are many question answering systems which have been developed but they are for open or closed domains. Moreover these systems are based on IR concept, hence there is need for developing QA System for restricted domain. In this paper we have proposed a frame work for Restricted domain Question Answering system which works over the concept of Information Extraction.

II. RELATED WORK

Since the idea of QA System has been coined various question answering systems have been developed apart from LUNAR [7] and BASEBALL [8] with different concepts and ideas. SAD SAM [9], Baseball and DEACON [10] have been developed using the concept of List-structured data-base. PICTURE LANGUAGE MACHINE (PLM) [11] and NAMER [12] are the two systems which depend on graphic database. PLM was designed to translate English or graphic data into a subset of the predicate calculus while NAMER takes a probabilistic learning approach for the same. Later on a new concept of *Text-Based Systems* have been evolved which attempt to find answers from ordinary English text. Protosynthez [13], Automatic Language Analyzer [14] and the General Inquirer[15] are the examples of Text-Based System.
In the last two decades dozens of question answering systems have been developed using some new concepts and techniques. Hyo-Jung et al. [16] presented Chinese question classification which extract word and bi-gram from questions as classic features. Kepei Zhang and Jieyu [17, 18] did the same but uses the concept of word, named entity, part of speech (POS) and semantics as a classic feature. Santosh Kumar Ray [19] proposed a new method based on the usage of the Word Net while Svetlana Stoyanchev [6] evaluated the use of named entities, verbs, and prepositional. Kangavari et al. [3] presented simplest approach to improve the accuracy of a question answering system. Chiyoung Seoa, Sang-Won Leeb et.al [20] showed that RDBMS implementation using inverted index technique almost always outperforms the IR implementations. Paloma Moreda et al [5] proposes the use of semantic information in QAS. Liang Yunjuan & Ma Lijuan [21] discussed the design of dynamic knowledge-based full-text retrieval system with inverted index technology. Ramprasath et.al has surveyed and compared different types of question answering system [22].

Many architecture have also been proposed for developing Question Answering Systems. Kolomiyets and Mones [23] have proposed a model based on the translation of question statement and document into a computer readable format which is a little bit sophisticated and expensive. In 2008 Mohammad Reza Kangavari et al.[3] have also proposed an architecture using dynamic patterns and semantic relations among words verb and keywords. Both these architectures may perform well but they are very complex, containing a large number of modules which is difficult to implement. Few architecture have also been proposed by Moldovan and Harabagiu [24] and Ramprasath and Hriharan [22] which seem to be simple by architecture, but the problem is that their working steps are not very clear to understand.

In our research work we have proposed a framework for developing a question answering system using advanced NLP tools which is easy to understand and implement. The next section discusses the proposed framework in detail.

III. PROPOSED ARCHITECTURE OF QUESTION ANSWERING SYSTEM

In the last two decades there have been many developments in the field of NLP and IR. Lot of tools and techniques have been proposed in order to enhance the performance of the searching and retrieval process such as: development of better similarity measure, efficient design of IR system, availability of better resources and tools such as POS tagger, NE recognizer content extractors and many more. Keeping these things in mind we have proposed a new architecture for QA system which includes these newly developed tools and techniques thus improving the overall performance of question answering system. Fig.[1] shows the proposed architecture of question answering system.

A. Question Processing Module

In this module the given Question is processed to get some important information from it. Steps through which question Processing Module passes and their descriptions are given below.

Steps in Question Processing Module:

1) Find the Type of given question using Wh word.
2) Find out the expected type of answer.
3) Get the Keywords from the Question.
4) Find out the Focus of the question.

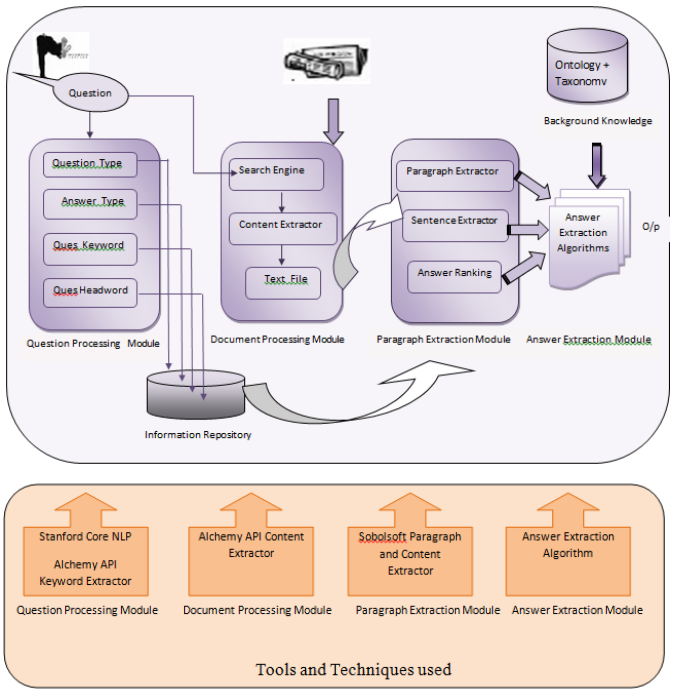

Fig. [1] Proposed Architecture Of Question Answering System.

Various information which we will get through this module are the Type of Question, Expected Answer Type, Focus or Head Word of the Question and the Question Keywords.

As specified in TREC of Question Answering track, broadly questions can be classified into three basic categories *Definition Type*, *Descriptive Type* and *Factoid type*. Definition Type of questions generally ask for exactly one or two sentences which define a particular noun in the question while Descriptive Type of questions requires few set of sentences or a paragraph which gives through information about a topic and Factoid Type of Questions are the one which look for one or two word exact answer. Table [1] categorise various Wh questions into these three classes.

Table [1] : Question and Answer Type

Question Type / Wh Word	Definition Type	Descriptive Type	Factoid Type
QUESTION	What	Why How What	Who When Where What Which

We can easily observe from the above table that, the *what* type of Questions fall under all the three categories. The definition type of *What* questions can easily be identified and answered, which we will discuss later in this paper but it is a bit difficult to distinguish between the descriptive and the factoid type of *What* questions hence it is a bit difficult to answer such *What* questions. The information about the *Wh* question word and the Head word, both can help in predicting the type of expected answer for the given questions. The concept of head word also called as the question focus, has been given by Zhiheng Huang et.al. [25]. It is the actual keyword or the expected entity about which the question is asking for.(See Table [2]).

S.no.	Question	EAT
1.	Where	Place/ Location Name
2.	When	Time/ date
3.	Who	Person Name

S.no.	Question	Head Word
1.	What is the Percentage of P2O5 in 'Pelofas' fertilizer ?	Percentage
2.	Which crop is cultivated in black soil	Crop
3.	Mastitis is a disease of which organ?	Organ

Table [2] : (a) Question and Expected Answer(b)Question and Head Word

B. Document Processing Module

This module retrieves the relevant documents to the given question from the large data set.

Steps in Document Processing Module:

1) Get the question in hand and search relevant documents using a reliable search engine.
2) Take top ten relevant documents.
3) Extract the content from these documents.
4) Save these contents in a text file.

The module first retrieves the relevant set of documents from web using any search engine. Take the top ten pages and pass the links of these pages to the Alchemy Content Extractor. It removes all the advertisements, navigation links, and other unwanted contents from these web pages and returns only the textual content. Repeat this process for all the ten selected url's and save the content of these web pages to the Text File and merge them into a single text files and pass it to the next module.

C. Paragraph Extraction Module:

This is the module where the task of Paragraph Extraction and Sentence Extraction takes place in order to find out the most probable Answer of the question in hand.

Steps in Paragraph Extraction Module

1) Run Paragraph extractor over the text file, obtained from previous module.

2) If question is Definition or Factoid Type, send the extracted paragraph for Sentence Extraction to the next sub module.

The output of the Document Processing Module that is the Text file and the output of the Question Processing Module which we have stored in the Information Repository after Processing the Question are passed to the Paragraph Extraction Module. Using the information obtained from information repository the Paragraph Extractor will extract only those paragraphs which have the same keywords as those of the Question Keyword. These paragraphs will be sent to the next sub module that is the Sentence Extractor only in case of *Definition* and *Factoid* Type of question. For descriptive Type of questions these paragraphs itself is given as final output. Ranking is basically done for Descriptive type of questions. The extracted sentences are then passed to the *Answer Extraction Module.*

D. Answer Extraction Module:

This module presents algorithms for extracting the potential answer for all the three categories of questions that is Definition, Descriptive and Factoid Type of Question.

1) Definition Type of Question

Table [2] clearly shows that only the *What* type of questions fall under the Definition type of questions. However it falls under all the three categories of questions. As we have discussed earlier in this section that it is a little bit difficult to distinguish between the descriptive and the factoid type of *What* questions hence here in this paper we have considered only the Definition Type of *What* questions.

In order to find out the grammatical structure of Definition Type of *What* questions, we have parsed all the three categories of *What* type of questions using Stanford's coreNLP [29] and analysed them. We observed that all the Definition Type of *What* questions have the following question pattern:

Question Pattern

(What + aux + [NOUN / JJ+NOUN / keyword])

Once we have found out the question pattern we can predict the potential answers from the data set. Those sentences which have the same head word as that of question and have any one of the following parsing structure are considered to be the candidate answer for that question.

Answer pattern

{ [NOUN / JJ+NOUN] + AUX + X } OR

{ [NOUN / JJ+NOUN] + can be define as + X} OR

{ X + is known as + [NOUN / JJ+NOUN] } OR

{ X + called + NOUN/ JJ+NOUN] } OR

{ [NOUN / JJ+NOUN] + means + X }

Here [NOUN / JJ+NOUN] shows the same Noun or any adjective followed by the Noun as that of the question and X shows *anything*. Say these patterns as $P_1, P_2, P_3 \ldots\ldots P_n$ then the algorithm for extracting the potential answer candidate for definition type of question can be given as:

Algorithm:1

IF
Sentence Pattern $\equiv P_1 \| P_2 \| P_3 \| P_4 \| P_5$ then
Return it as Answer
Else
Return "Document does not contain the answer".

Experimental Output
Question : What is mixed cropping ?
Output : - Mixed cropping means growing two or more crops simultaneously on the same piece of land .
- **Mixed cropping** is growing two or more crops simultaneously on the same piece of land.
- Mixed cropping is a type of agriculture that involves planting two or more plants simultaneously in the same field.

2) Descriptive Type of Questions

As we know that for descriptive type of question the output should be few set of sentences or paragraph here we just require to retrieve those paragraphs, which have the same keywords as that of given question in hand. We can use the following algorithm to extract the answer of descriptive type of question :

Algorithm: 2

Answer Extraction Algorithm for Descriptive Type Question :

1) Find out "KEYWORD" from Question

2) Retrieve relevant paragraph from data set having the "KEYWORDS"
return it as answer
Else return "Document do not contain the answer."

In our experiment we have extracted the keywords using AlchaemyAPI keyword extractor [30] and for extracting the paragraphs and sentences we have used a software called "Paragraph and sentence Extractor" [31] .

3) Factoid Type of Questions

Among all the three types of questions, answer extraction for factoid type of question is the most difficult task. The algorithm for extracting the answer for factoid type of question works similar to that of descriptive type of questions till the task of paragraph or sentence extraction, then we have to extract the exact answer keyword from that sentence. This task of extracting the exact keyword makes the factoid question most difficult to answer.

In case of factoid type of questions Head Word plays a big role in extracting the answer. As we have defined earlier "Head word are the single word specifying the object that the question seeks". Syntactic parsers can be used to find out the head word of any given sentence. In most of the time head word is the first noun word after the *Wh* word.

Head Word : First [NOUN / NOUN PHRASE] after WH word

In order to perform our experimental work we have used the output of both the Stanford and Berkeley parser to find out the head word of the question.

The following algorithm can be used to extract the potential answer for factoid type of question:

Algorithm: 3

Answer Extraction Algorithm for Factoid Type of Question:

1) Find out "KEYWORD" from Question.
2) Find out "HEAD WORD" from Question .
3) Retrieve relevant paragraph then relevant sentences from data set having the "KEYWORD"
4) Match additional word in retrieve sentences with the Expected Answer Type (EAT ,type of headword) using ONTOLOGY.
5) If matched return it as answer else
   If there is any other sentence do step (4)
   Else
   return "Document do not contain".

Experimental Output

Question : What is the national flower of Japan ?
Output : Cherry Blossom // Hypernyme using agrovok : Flower
Question : Which chemical is responsible for the color of tomato ?
Output : Lycopene // Hypernyme using agrovok : Chemical

IV. EXPERIMENT AND RESULT EVALUATION

In this work we have taken agriculture as the restricted domain. Since we did not find an appropriate benchmark dataset for agricultural domain we have created our own dataset of questions using Agricultural related chapters of NCERT syllabus, course book of Bachelors of Agriculture and various Agricultural web pages from Internet.

The output of the answer extraction algorithm for the definition type of questions is best among all the three classes of questions. The performance of all the above explained algorithms for different class of questions is shown in the following figures (See Fig.[2]):

The Answer Extraction Algorithm for Definition Type Question can return the exact answer almost 92% of time but it fails only when the answer statement is not in a proper format that is the answer does not contain the main Keyword or Head word of the question. For example consider a question, *What is mixed cropping ?* If the answer is in the form *"**Mixed cropping** is growing two or more crops simultaneously on the same piece of land."* then the algorithm can easily extract the answer but if it would be in the form "*growing multiple crops in the same field"* then the algorithm fails to extract the answer from the document. In case of descriptive type of questions since we want whole information about that particular topic hence the output is same as that of the output of the search engine while the performance of the algorithm is reflected well in Open Domain. The performance of the algorithm for extracting answers for factoid type of questions can be improved in future with the better accessibility to the ontologies and with the better performance of various tools and software used for the experiments.

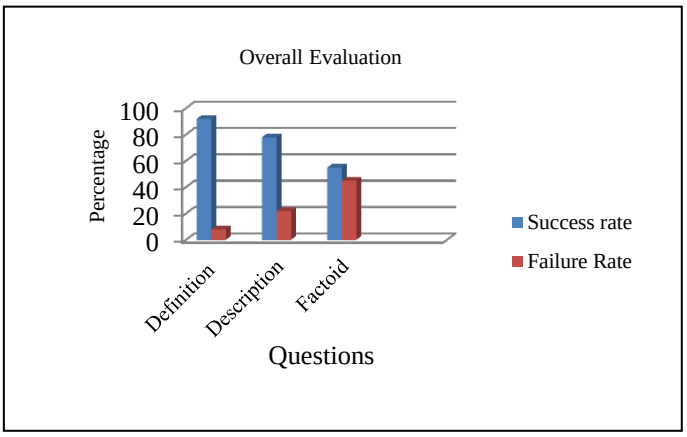

Fig.[2] Combined Performance Evaluation

V. CONCLUSION AND FUTURE WORK

In this paper we have proposed a framework for restricted domain question Answering System using advanced NLP tools and software. This framework can be used to develop a Question Answering System for extracting exact and precise answer from restricted domain textual data set. The proposed framework not only provides a simple and implementable framework for developing question Answering System but also provides a proper flow of data for answer extraction. Since the proposed model works over keywords and headword and is independent of the question or sentence structure, it has reduced the overhead of question normalization. Moreover since the framework is given for restricted domain, it also handles the issue of word sense disambiguation. The major problem which exists with the proposed framework is that it's performance is dependent on the performance of the search engine and the used NLP tools.

This problem can be resolved in future by including the concepts of gazetteer list or domain specific dictionaries.

REFERENCE

[1] Chali, Yllias: *"Question Answering Using Question Classification and Document Tagging"*, Applied Artificial Intelligence, pp.500—521,2009.

[2] Hacioglu, K., & Ward, W. "*Target word detection and semantic role chunking using support vector machines"*. In Proceedings of the human language technology conference (HLT-NAACL2003), 2003.

[3] Kangavari M.R., Ghandchi S., Golpour M.,"*Information Retrieval: Improving Question Answering Systems by Query Reformulation and Answer Validation".* Academy of Science, Engineering and Technology, pp.303–310, 2008.

[4] Molla D., and Vicedo J., *"Question answering in restricted domains: An overview",* Computer Linguist, pp.41–61, 2007

[5] Moreda P., Llorens H., Saquete E., & Palomar M., "*Combining semantic information in question answering systems"*, Information Processing & Management, pp.870 – 885, 2011.

[6] 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), pp. 9–16 Manchester,UK. August 2008".

[7] Woods W.A, Kaplan R.A, Nash-Webber.B, "The lunar sciences natural language information system" , Final report: BBN Report #2378. Technical report, Bolt Beranek and Newman Inc.,Cambridge, MA., June 1972.

[8] Green B.F, Wolf A.K., Chomsky, K. Laughery, *"BASEBALL: An automatic question answerer"*, in: Proceedings of Western Computing Conference, vol.19, pp. 219–224,1961.

[9] Kuno S., and Oettinger A.G., " *Syntactic Structure andambiguity in English".* Fall Joint Comput.Conf.241963,SpartanBooks,Baltimore,pp.397-418.

[10] THOMPSON F.B., CRAIG J., "*A DEACON breadboard summary*." RM64TMP 9, TEMPO General Electric Co., Santa Barbara, Calif., Mar.1964.

[11] Sillars W., "*An algorithm for representing English sentences in a formal language.",* Rep. 7884, Nat. Bur. Stand., Washington, D. C., Apr. 1063.

[12] Simmons R.F and Londe D , " *Namer: a pattern recognition system for generating sentences about relations between line drawings*". Doc. TM-1798, System Development Corp., Santa Monica Calif., Mar. 1964.

[13] Simmons R.F, Klein S., and, McConlogue K. L, "*Indexing and dependency logic for answering English questions"*. Amer. Documentation *15*, 3, pp.196-204, 1964.

[14] LYONS J., and THORNE J.P., Quart.rep. "*An automatic language analysis*", 1-7. ASTIS, Indiana U., Bloomington, Ind., 1960-1962.

[15] Stone P. J., Bayles R. F., *Namerwirth*, J. *Z.,* and Ogilvie D.M., "*The general inquirer : a computer system for content analysis and retrieval based on the sentence as a unit of information.*" Behav. *Sci.,* 7, 4 (1962), 1-15.

[16] Hyo-Jung , Sung Hyon Myaeng b, Myung-Gil Jang " *Effects of answer weight boosting in strategy-driven question answering"*.Information Processing and Management, 2011.

[17] Li .S, Zhang J., Huang X., and Bai. S., "*Semantic computation in chinese question-answering system"*. Computer Science and Technology, vol.9,pp.1–7, 2002.

[18] Kepei Zhang & Jieyu Zhao, "*A Chinese Question-Answering System with Question Classification and Answer Clustering*", 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 2010

[19] Santosh Kumar Ray , Shailendra Singh , B. P. Joshi, "*A semantic approach for question classification using WordNet and Wikipedia*". , Pattern Recognition Letters, p p.1935-1943, October, 2010

[20] Chiyoung Seoa,, Sang-Won Leeb, Hyoung-Joo Kima " *An efficient inverted index technique for XML documents using RDBMS*". Information and Software Technology, pp.11–22, 2003.

[21] Liang Yunjuan , Ma Lijuan, Zhang Lijun, Miao Qinglin "*Research and Application of Information Retrieval Techniques in Intelligent Question Answering System*"., IEEE, 2011.

[22] Muthukrishnan Ramprasath and Shanmugasundaram Hariharan. "*A Survey on Question Answering System*", International Journal of Research and Reviews in Information Sciences, Vol. 2, March 2012.

[23] Oleksandr Kolomiyets and Marie –Francine Moens, *" A survey on question answering technology from an information retreival perspective,"* Elsevier J. Information Sciences, 181, pp. 5412-5434, 2011.

[24] Moldovan, D., S. Harabagiu, M. Pasca, R. Mihalcea, R. Girju, R. Goodrum, and V. Rus. (2000). "*The structure and performance of an open-domain question answering system*." In 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), Hong Kong, 2000.

[25] Zhiheng Huang, Marcus Thint, and Zengchang Qin, "*Question classification using headwords and their hypernyms*". In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-08), pp. 927–936, 2008.