ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

CLOSED DOMAIN QUESTION ANSWERING SYSTEM USING NLP TECHNIQUES

Sweta P. Lende*, M. M. Raghuwanshi
Dept. of Computer Technology, YCCE, Nagpur, Maharashtra, India
Dept. of Computer Technology, YCCE, Nagpur, Maharashtra, India

DOI:

ABSTRACT

Question Answering (QA) system in information retrieval is a task of automatically answering a correct answer to the questions asked by human in natural language using either a pre-structured database or a collection of natural language documents. It presents only the requested information instead of searching full documents like search engine. This paper presents the implementation methods and experimental result with analysis for closed domain QA System which handle only documents related to education system to retrieve more precise answers using NLP techniques.

KEYWORDS: Question Answering, NLP, Information Retrieval, Education Acts

INTRODUCTION

Recently, there are many search engines available. All these search engines have great success and have remarkable capabilities, but the main problems with these search engines is that instead of giving a direct, accurate and precise answer to the user's query they usually provide list of document related to websites which might contain the answer of that question. Although the list of documents which are retrieved by the search engine has lot of information about the search topic but sometimes it may not contain relevant information which the user is looking for.

The main aim of Question Answer system is to present short answer to user query instead of searching list of document related to search topic. Users just have to ask the question and the system will retrieve the most appropriate and correct answer for that question and it will give to the user.

For implementing QA system, most of the researchers working in various domain such as Web Mining, NLP, Information retrieval and information extraction and so far focused on open-domain QA and close domain QA system. Systems can be divided into two types on the basis of the domain; Closed-domain question answering refers to specific domain related questions. Open-domain question-answering deals with the questions which are related to every domain.

PROPOSED APPROACH

According to literature survey of Question Answering Systems, we can observe that the closed domain QA system gives accurate answer than the open domain QA system. If we see the current scenario, there is no QA system for exactly answering the queries on document of education act related information, which ensures the correct answers. So, the idea for developing the Question Answering System on education act is proposed.

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

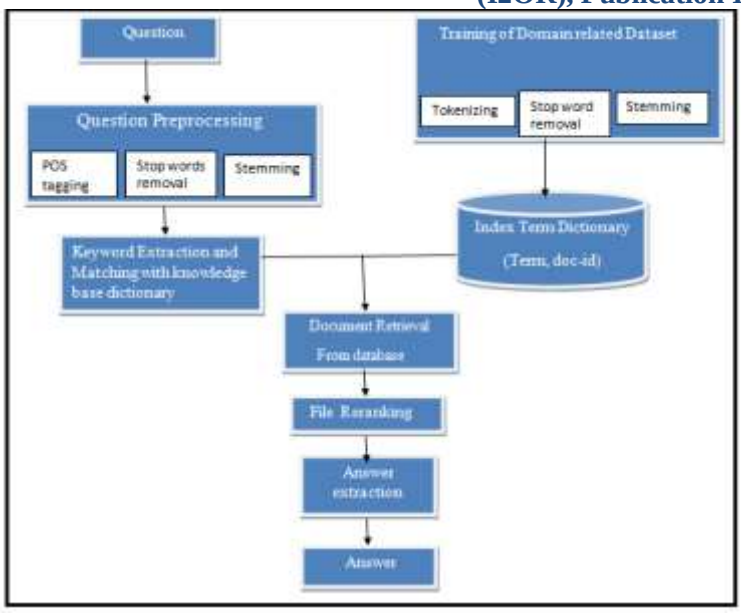


Figure 1. Proposed Architecture of QA System

IMPLEMENTATION

A. Creation of Dataset (Corpus):

The first phase of the Question Answering System is to create Corpus. As we have to design the closed domain QA system on education acts, so we have gone through the different websites and taken the text data from the websites: www.legislation.gov.uk. For each section of education acts, there is one text file; such for 583 sections, 583 text files are stored as corpus. Education acts related text file contain information about each section of education such as information about school, funding authorities of school, areas of school, duties of teacher related to student, duties of parents to secure children education and many more information related to education. Total 583 sections are available related to education acts.

B. Preprocessing:

After creating corpus some preprocessing operations are performed on each text file of corpus. Major tasks in pre-processing are stop words removal and stemming.

Stop-word Removal: The English stop words like "is, for, the, in, etc" are remove from each text files of dataset by maintaining English stopword dictionary.

Stemming: Reduce terms to their "roots" before indexing
e.g., ***automate(s), automatics, automation*** all reduced to ***automatic***.
For stemming, English stemwords dictionary (for example a set of documents that contain stem words) is maintained for extracting keywords.

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

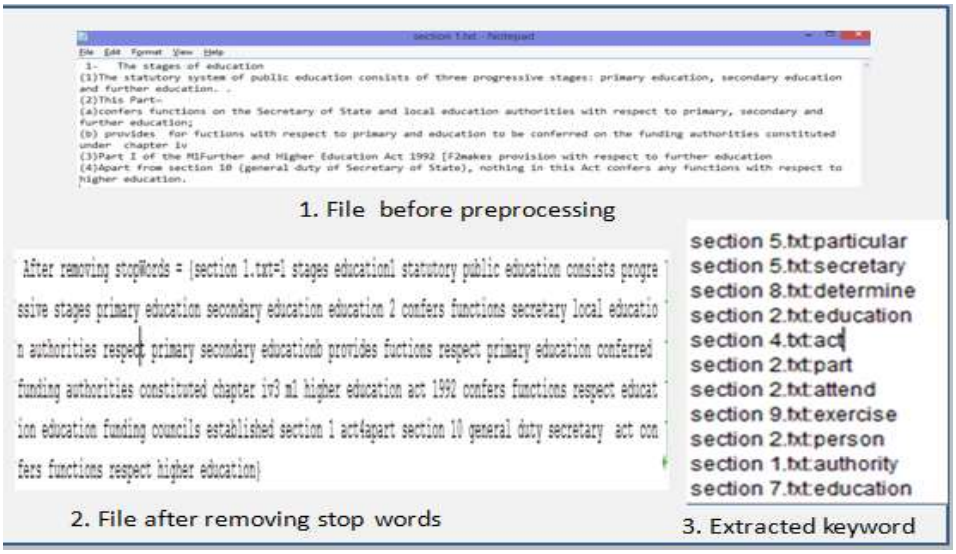


Figure 2. Output of Text file after and before Preprocessing

C. Index Term Dictionary

After pre-processing the extracted keyword are stored in index term dictionary. Extracted keywords contain only stemwords which obtain after performing stemming. Index term dictionary is created by using java and stored as table in mysql. Dictionary forms a structure containing two columns as word and file name. Word is nothing but a extracted keyword and file name is name of file which contain that keyword. The following screen shot show the structure of index term dictionary.

For each term t, we store a list of all documents that contain t.

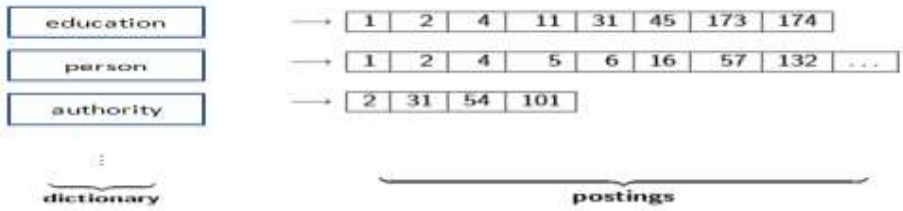


Figure 3. Structure of index term dictionary

D. Question Dataset:

As we are dealing with the closed domain Question Answering, the system need to answer the different questions related to domain. So the dataset of 300 different questions on education acts are maintained to train our QA system. These questions are taken from different users. According to these questions, the QA system is designed to give the answers.

The Examples of the Question dataset is given below.

Q1. What are the stages of education?
Q2. What is primary education?
Q3. What is secondary education?
Q4. What is higher education?
Q5. What is duty of parents to secure education of children?
Q6. What are the general duties of secretary of state?
Q7. Who determine a local educational authority?

E. Question Preprocessing:

The given input query is preprocessing by performing some preprocessing operation on it i.e. POS tagging, stop words removal and stemming.

a. User Query:

User will enter the query related to education system. For example, the user can ask the question "what is the duty of parent to secure children education?" or what primary school? Or any query related to education system.

b. POS tagging:

First we perform POS (part of speech) tagging operation on input query to tag each word of user query with its type such as verb, noun etc. For tagging each word POS standford tagger is used.

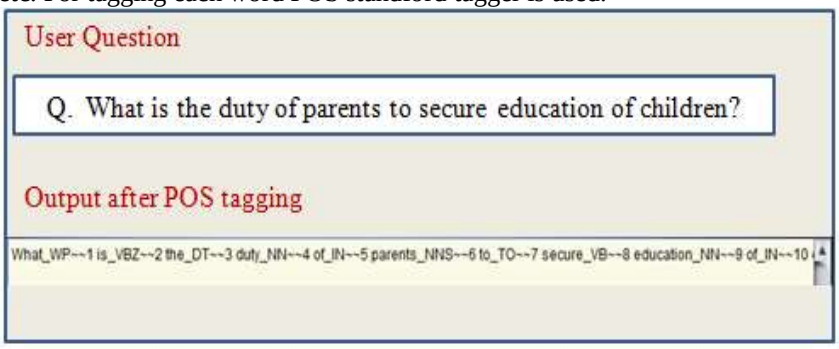


Figure.4 Output after POS tagging Operation

c. Extracted Keyword:

From the user query, the keywords are extracted. These keywords are got by removing the symbols and stopwords from user query, also stemming is applied on keywords so as match with index term dictionary term for document retrieval. English stop words and stemmed words dictionary is maintained to extract keywords

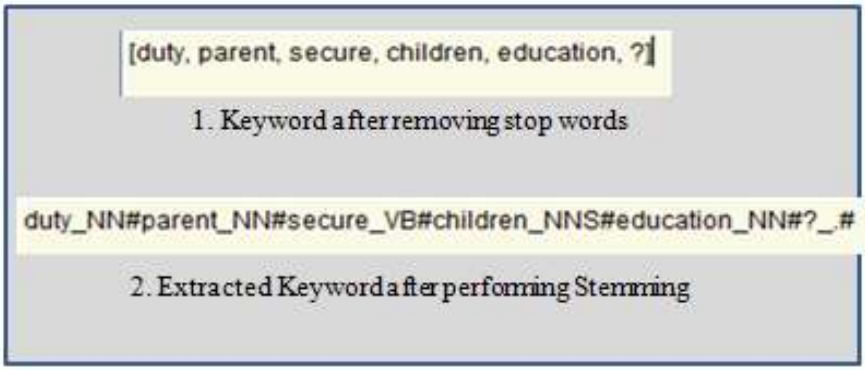


Figure. 5 Extracted Keyword from Question

F. Document Retrieval:

In document retrieval system the extracted keywords which are obtained by query are match with term of indexed dictionary. After matching only match keyword's ids are retrieved as a document. For more than one keywords, it takes the intersections of all document ids where these terms are present so that where all the keywords are finding only those document are to be retrieved for candidate answer passages.

For example

The term duty is present in file 1, 2, 3 5, 8 &

The term parent is present in 1, 2, 3, 4, 6, and 8.

It retrieves all the files as file 1,2,38 the keywords that obtained from user query are matched in documents. It can give number of document which matched keyword.

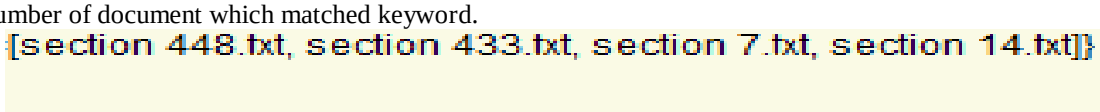


Figure. 6 Retrieved Documents

G. File Reranking:

In case of keyword ranking, first we find out score in between query keywords and all files which is obtained by document retrieval. For finding score we use jaccard similarity function. In case of jaccard similarity, we first find out the intersection in between extracted keywords of query and all files retrieved after document retrieval.

Score = (A n B) / (A u B)

Where, A = set of extracted keywords.

B= set of files keywords

Ex. Query words= {duty, parent, education, children} &

Files words = {Duty, parent, secure, education, children, compulsory, school, age}

Similarity score=4/8= 0.5

According to score all files are rank and max score files are extracting for answer extraction

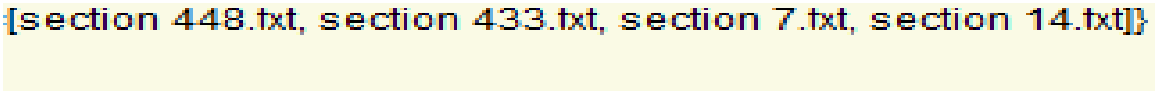


Figure. 7 Retrieved Document after reranking files

H. Answer Extraction:

In case of answer extraction, POS tagging is apply on all filter document which is obtain in keywords ranking. After applying POS tagging, we check the sense in between extracted which is obtain by query and filter document.

e.g. Suppose extracted keywords from query are [Duty, parent, education, and children]

We check where,

[Duty, parent] used as - noun in document

[Education] used as- pronoun in document

[Children] used as- NNS

After checking sense in between query and document. We extract that paragraph which contains same sense as the query sense.

I. Answer:

After answer extraction we select answer which we obtain in answer extraction and we present this answer to user for that

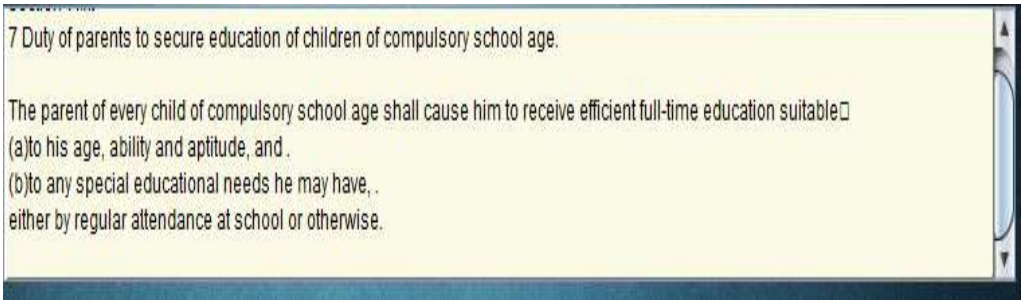


Figure. 8 Extracted Answers

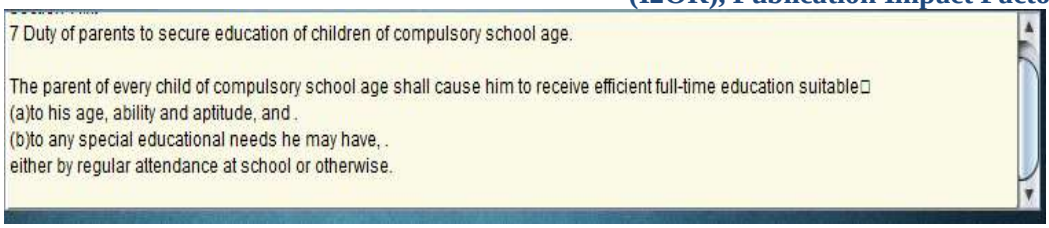
EXPERIMENT AND TEST CASES

The below all Figures illustrates the User Interface for Question Answering System. The user will enter the question in the text box. After clicking on 'answer extraction' button, the exact answer retrieved in the answer box.

The various examples of various types are as follows

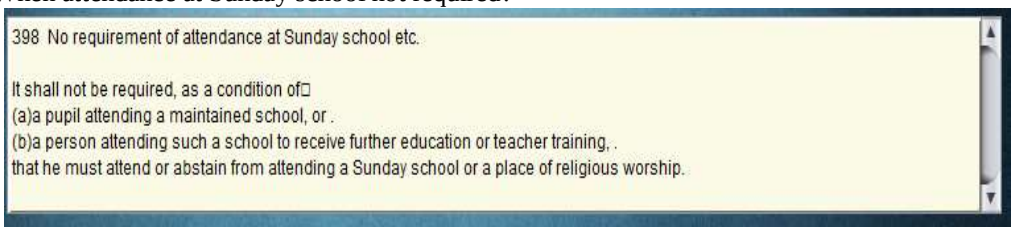
Implementation Example 1:

Question: What is the duty of parents to secure education of children?



Implementation Example 2:

Question: When attendance at Sunday school not required?

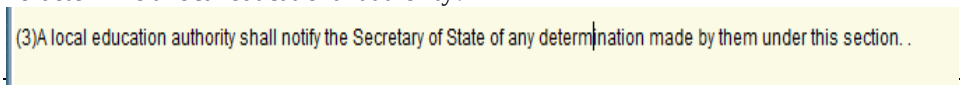


Implementation Example 3:

Question: List the stages of education?

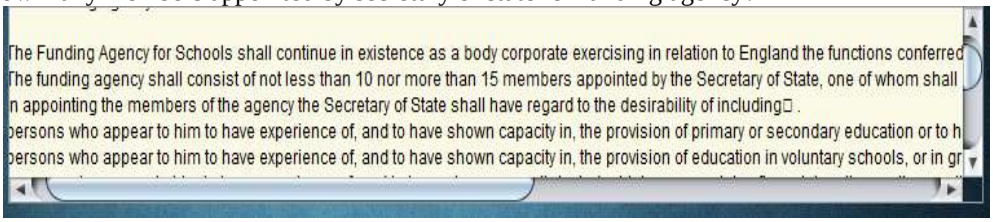
Implementation Example 4:

Question: Who determine a local educational authority?



Implementation Example 5:

Question: How many members appointed by secretary of state for funding agency?



EVALUATION AND RESULTS

There are several evaluation metrics that differ from one QA campaign to another .Moreover, some researchers develop and utilize their own customized metrics. The proposed system tries to find precise answers .The following measures are the most commonly used measures that are typically utilized for automated evaluation:

Precision, Recall and F-measure:

Precision & recall are the traditional measures that have been long used in information retrieval. While the F-measure is the harmonic mean of the precision and recall; these three metrics are given by:

Precision= number of correct answers / number of questions answered

Recall= number of correct answers / number of questions to be answered

F Measure = 2 Precision * recall / Precision + Recall

Experimental Results:

Total Questions 100

Response by the system 76
Correct answer 68
Precession 0.89
Recall 0.68

Total Question	Response by system	Correct answer	Precession= (Correct answer/ no. of questions answered) *100	Recall= (no. of correct answers / no. of questions to be answered) *100
100	76	68	89%	68%

Figure. Experimental Results showing precession and recall

The proposed system is tested with 100 different questions. These questions include various queries by the users about education sections. These questions are mostly of structured question format. Out of 100 questions, nearly 68 are of exact.

CONCLUSION

Question answering system using NLP techniques is more complex compared to other type of Information Retrieval system. The Closed domain QA Systems give more accurate answer than that of open domain QA system but this system is restricted to single domain only.

The proposed Question Answering system for closed domain gives accurate answer for the users question based on domain. The question preprocessing module will determine the target word from question. Based on the target word, it retrieved the document from corpus to get the accurate answer. The system is tested based on the various evaluation parameters. The system is tested on 100 questions showing the accurate & precise results.

The QA system for closed domain on education acts using NLP techniques is proposed to give the accurate and suitably more correct answers for user's structure queries.

REFERENCES

1. Amit Mishra, Nidhi Mishra and Anupam Agrawal, "Context- Aware Restricted Geographical Domain Question Answering System", In 2010 International Conference on Computational Intelligence and Communication Networks.
2. Rivindu Perera "IPedagogy: Question answering system based on web information clustering" 2012 IEEE Fourth International Conference on Technology for Education.
3. Pragisha K. and Dr. P. C. Reghuraj, "A Natural Language Question Answering System in Malayalam Using Domain Dependent Document Collection as Repository." International Journal of Computational Linguistics and Natural Language Processing Vol 3 Issue 3 March 2014 ISSN 2279 – 0756.
4. Payal Biwas, Aditi Sharam,Nidhi Malik" A Framework For Restricted Domain Question Answering Sysem"2014 International conference on issue and challenges in intelligent computing Technique.
5. Zeng-Jian Liu, Xiao-Long" A Chinese Question Answering System Based On Web Search"2014 the international conference on machine learning.
6. Jibin Fu, Keliang Jia and Jinzhong Xu, "Domain Ontology Based Automatic Question Answering", 2009 International Conference on Computer Engineering and Technology.
7. Abdullah M. Moussa and Rehab F. Abdel-Kader, QASYO: "A Question Answering System for YAGO Ontology", International Journal of Database Theory and Application Vol. 4, No. 2, June, 2011.
8. Moussa, Abdullah M. & Rehab, Abdel-Kader (2011) "QASYO: A Question Answering System for YAGO Ontology". International Journal of Database Theory and Application. Vol. 4, No. 2, June, 2011. 99.

9. Zeng-Jian Liu, Xiao-Long Wang and Qing-Cai Chen "A Question answering system on web search" International conference on machine learning 2014.
10. Lahiru Samarakoon, Sisil Kumarawadu "Automated Question Answering for customer Helpdesk Application" 2011 6th Internation conference on Industrial and Information System.
11. Tiansi Dong and Ulrich Furbach "A natural language QA system as a Participant in Human Q and A portal".
12. Anette Frank , Hans-Ulrich Krieger, Feiyu Xu, Hans Uszkoreit, Berthold Crysmann, Brigitte Jörg and Ulrich Schäfer, "Question answering from structured knowledge . sources", In German Research Center for Artificial Intelligence, DFKI, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany Available online 27 January 2006.
13. Pum-Mo Ryu, Myung-Gil Jang and Hyun-Ki Kim. 2014. "Open domain question answering using Wikipedia-based knowledge model." In Information Processing and Management 50 (2014) 683– 692, Elsevier.
14. Adel Tahri and Okba Tibermacine. "DBPEDIA BASED FACTOID QUESTION ANSWERING SYSTEM." In International Journal of Web & Semantic Technology (IJWesT) Vol.4, No.3, July 2013.
15. Menaka S and Radha N. "Text Classification using Keyword Extraction Tech- nique", in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013.
16. ZHANG Yu, LIU Ting, WEN Xu, "Modified Bayesian Model Based Question Clas- sificatio", , vol.19, pp. 100-105.
17. Pum-Mo Ryu, Myung-Gil Jang and Hyun-Ki Kim. 2014. "Open domain question answering using Wikipedia-based knowledge model." In Information Processing and M