Natural Language Query Based Question Answering System

Soumya P. Panda, Vikash K. Pandit, Rohit Kumar, Sudhanshu Chaturvedi, Akash Das

Abstract: ***Development of natural language query based automatic question answering system is in huge demand these days and is a rapidly growing field. It is considered to be the most powerful application for answering different user queries not only on limited domains but also in multi domain environments. In this work, a natural language query based intelligible question answering system is presented that extracts relevant answers from the documents and present the answer in a pre-defined format to the user. A comparative study of the presented model with the traditional techniques is also presented.***

Keywords: Question Answering System, Information Extraction, Natural Language Processing, Unstructured documents.

I. INTRODUCTION

Research in the field of designing Question Answering Systems (QAS) is a rapidly growing field worldwide [1][2]. The demand of system that provides short, precise and formatted answers to user queries is in huge demand due to the different available technologies based on Natural language processing[3][4]. A QAS focuses on presenting the answers of the user queries effectively in less time [5][6] [7]. It's a rapidly growing field and has several applications [8] [9] [10]. The question answering system may take the help of Information Extraction (IE) technology to extract the answers from specific document contents[11].

IE focuses on automatic extraction of specific information from offline and online repositories. The information extraction system identifies a subset of information from large repository and presents it in a pre defined format. Advantages of Information Extraction System can be seen in many text and web applications, for example, comparison of hotel prices on several websites, understanding the disease of a patient from his symptoms, searching for contacts from emails, etc. Information extraction may be applicable to different textual sources: from emails and Web pages to reports, presentations, legal documents, and scientific papers.

Medical Sector– Extracting useful information from the reports or medical summary of a patient written in an unstructured or semi-structured form becomes a very complex task. If the information in these semi-structured data is analysed and valuable information is extracted in the form of structured format, then the report of a patient can be analyzed within less time and exact information can be fetched.

Media Monitoring - Information Extraction System is very useful for automatically tracking specific events from companies, getting news updates from news sources, tracking digital fraud analysis and extraction of specific information from the community website, etc.

Automatic Form Filling– The information from the document is analyzed and extracted onto the respective fields which save time.

Information Extraction from Emails: Through information extraction, the exact information that can be extracted in a structured format may be in the form of a table or file making it easily accessible for the user to locate the exact information for the query.

Information Extraction process from textual documents and web pages is a very complicated task because they are mainly based on the automatic recognition of natural languages. Natural language is intrinsically ambiguous; the meaning of words depends on the context as well as on specific situations. Furthermore, extracting specific information from huge amount of dynamic collections of diverse materials is a critical challenge that the Information Extraction system is facing.

Human languages have a lot of associated issues while processing through machine creating problems in text extraction methods and identification of relationships between entities. The problem of appropriate extraction of answer becomes more complex due to the presence of synonymy, polysemy, antonymy, etc and the abbreviations having different meanings in different contexts. Categorizing the documents and finding the relevant text from them becomes difficult for diverse collection of documents. Few common issues are discussed below,

Synonymy: The state in which words sounding different in pronunciation but have the same or identical meanings. For example, if the system searches for the word 'politician' in but the document contains the word 'statesman' and not politician, then the text extraction system would have difficulties in finding out the answers corresponding to these words.

Polysemy: The state in which a word can have more than one meaning according to different situations and contexts. For example, the word 'fly' can be a verb or noun.

A number of applications of the QAS are available these days which is not only restricted to specific domains[12]. QASs for healthcare analytics[13][14], agriculture[15][16], virtual assistants[17][18],

Revised Manuscript Received on February 20, 2020.

Soumya P. Panda*, Department of CSE, Silicon Institute of Technology, Bhubaneswar, India. Email: sppanda.cse@gmail.com

Vikash K. Pandit, Department of CSE, Silicon Institute of Technology, Bhubaneswar, India.

Rohit Kumar, Department of CSE, Silicon Institute of Technology, Bhubaneswar, India.

Sudhanshu Chaturvedi, Department of CSE, Silicon Institute of Technology, Bhubaneswar, India.

Akash Das, Department of CSE, Silicon Institute of Technology, Bhubaneswar, India.

Natural language query based chat boats[19], virtual teachers to answer student queries [20] are a few most demanding applications of QAS. However, extracting appropriate answers and presenting it in appropriate format based on the user requirements is still a challenging task [21]. The QAS overview is shown in Fig. 1.

A number of researches are documented in the field for QAS. A discussion about a rule-based QAS is presented in [9] for clinical data. Creation of these types of systems needs domain expertise and is also a time consuming process to frame all possible rule sets. In [10] a discussion on developing a QAS for solving various plant disease problems is presented. [11] Presents a clinical information extraction application.

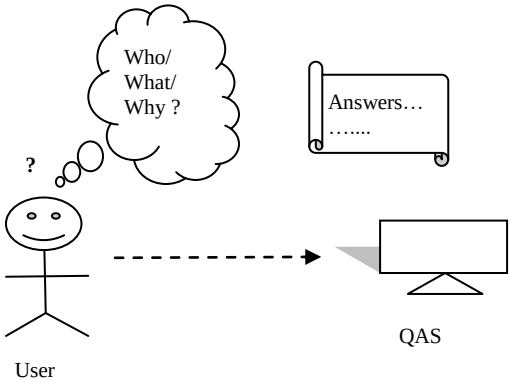


Figure.1: QAS System overview

In this work, a natural language query based QAS is presented that extracts relevant answers from the documents and presents the answer in a defined format to the user. Experiments were performed to evaluate the system performance with respect to different user queries. A comparative study of the presented model with the traditional techniques is also presented.

II. THE QUESTION ANSWERING MODEL

For the presented work, a document repository of 1500 documents is considered. The articles include different medical records, corporate reports, news items, research articles, social media data, etc. The wiki information resources are also used for extracting relevant answers. The phases involved in the QAS are presented in Fig. 2.

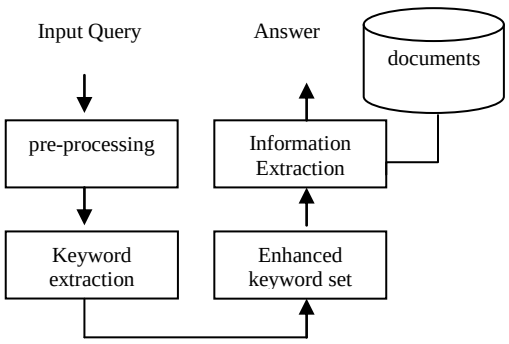


Figure.2: Phases of QAS

The presented model works by referring the pre defined rules designed for identifying the type of question and its likely answers. The common question types and its sub classes are presented in Table.1 and Table.2. However, an Index file is created initially from the document collection to match the appropreateness of the document with respect to the queries. With respect to the match score, the respective documents are analyzed using the presented algorithms. The answers are then extracted and presented as per the predefined format for each category. For the Wiki online document repository covering different domains, the web scraping method is used.

Table.1: Question classification for single format answer type

Question Type	Example	Answer Types
When	When did India get independence?	Date format
Where	Where is Tajmahal located?	Location-place format
Whom	Whom did Nancy meet at the bridge?	Person
Why	Why are leaves green?	Description

Table.2: Question classification for multiple format answer types

Question Type	Example	Answer Types
Who	Who is the prime minister of India?	Person
	Who is the largest seller of nutritional supplements in India ?	Organization
Which	Which person has maximum statue in the world?	Person
	On which river bank is Delhi located?	Location-place format
	Which date teachers day is celebrated ?	Date format
	Which organization launched Chandrayaan mission?	Organization
What	What is the value of 1 dollar in INR?	Money
	What date did India get independence?	Date
	What is the capital of India?	Location
	What is the name of the person who wrote Ramayana?	Person
	What is mitochondria?	Description
How	How old is Ramayana?	Value
	How many states in India?	Number
	How to write an application?	Description

Default	Can you tell me about Microsoft?	Description
	Define Gravity	Description

The format of the answer is created by the defined rules. However, for Auxiliary type user queries starting with- 'Does', 'Do', 'Is', 'Can', etc the expected answer type is either 'Yes' or 'No'. Therefore, the simple keyword based content matching is applied and based on the presence or absence of the required information in the document content, the answers are filled in the template as 'Yes' or 'No'. The example Aux-type words are given in Table. 3.

Table.3: Example Aux- Question type

Aux-Question Type	Answer Type
Is	Yes/No
Do	Yes/No
Does	Yes/No
Can	Yes/No

Illustration-1: single word answer

Input Query:
When did earthquake occur in Bhuj?
Wh-Question Type:
When
Wh-Question sub-Type:
NA
Answer_type:
DATE
Answer format:
Date
Output Answer:
1978

Illustration-2: Sentence based answer

Input Query:
What is the capital of India?
Wh-Question Type:
What
Wh-Question sub-Type:
Where
Answer_type:
LOCATION
Answer format:
Template
Answer Template:
<STR> is the capital of India
Extracted Answer:
New Delhi
Output Answer:
New Delhi is the capital of India

Illustration-3: Descriptive answer

Input Query:
Why are leaves green?
Wh-Question Type:
Why
Wh-Question sub-Type:
NA
Answer_type:
DESCRIPTION
Answer format:
Description template
Output Answer:
Leaves are green because they have green chloroplasts organelles that carry out photosynthesis

Illustration-4: Yes/No answer

Input Query:
Is there a flight to cuttack?
Output Answer:
No

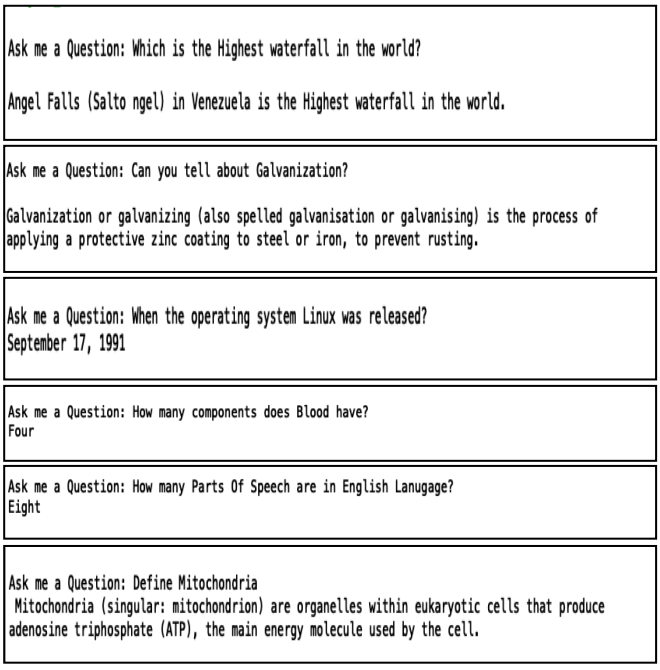
Algorithm-1:

Input: Query (Q)

Step-: Apply stop_word_Removal(Q)
Step-2: Using Regular Expression find out the question tag.
Step-3: Apply spaCy on the file content and query find out the NER tags.
Step-4: Get the relevant text from the file content using Keyword matching from the Query.
Step-5: if ques_tag = 'Who' , then tag ans_tag = 'PERSON'
Step-6: if ques_tag = 'Whom' then tag ans_tag= 'PERSON'
Step-7: if ques_tag = 'When' then tag ans_tag= 'DATE'
Step-8: if ques_tag= ''Where' then tag ans_tag= 'LOCATION'
Step-9: if ques_tag = 'Which' then tag ans_tag= 'PERSON', 'LOCATION', 'ORGANISATION', 'DATE'
Step-10: if ques_tag= 'What' then tag ans_tag= 'PERSON', 'LOCATION', 'ORGANISATION', 'DATE'
Step-11: if ques_tag = 'How' and attached with subqueries: 'much' | 'long' | 'many' | 'old' then tag ans_tag= 'DATE', 'TIME', 'MONEY', 'QUANTITY', 'PERCENT', 'CARDINAL'.
Step-12: if ques_tag= 'How' and no attached subqueries then tag ans_tag= 'DESCRIPTION'.
Step-13: if ques_tag does not match with defined tag set then tag ans_tag= 'DESCRIPTION'.

III. RESULT SCREEN SHOTS

The output screen shorts for some questions and obtained answers are shown below:



IV. IMPLEMENTATION DETAIILS AND RESULTS

For comparing the performance of the rule based question classification technique to accurately classify the wh-type, the SVM question classification model is used. For this purpose, the Scikit-learn machine learning library is used. The SVM question classification is a training procedure where the classifier learns from the features such as wh-word, its neighboring words, main verbs, word next to wh-word and auxiliary verb. The training was done taking more than 10000 questions as examples along with their respective classes.
The prediction was based on what the classifier learns at the training phase. The accuracy of the SVM question classifier is not satisfactory particula0rly for the wh-words which can have multiple sub classes like what, how, which, etc. However, the presented question classification model focuses on the varied types of wh-words and its subgroups. This model gave a satisfactory result as all classes and sub-classes are analyzed for classifying the questions to its actual type. The average accuracy percentage of SVM question classifier and presented model question classifier are shown in fig. 3 and fig. 4.

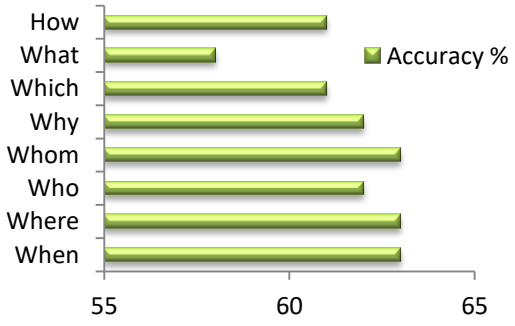


Figure.3: Average accuracy of SVM classifier on Wh-type questions

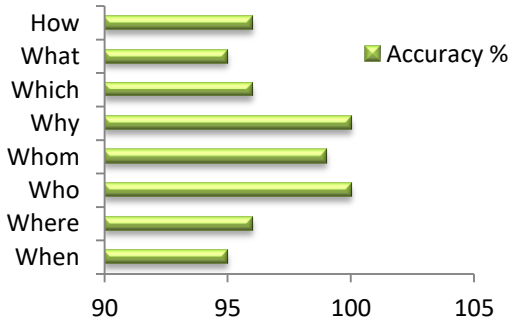


Figure.4: Average accuracy of presented technique on Wh-type questions

The precision and recall measure [10] are used for performance evaluation of the presented model. The obtained results for 50 Aux-type questions is presented in fig. 5. The results for 150 examples are shown in fig.6 and fig.7 respectively. In an average, the model achieves a precision measure of 80% and recall measure of 89%. While the corpus based question classification has an accuracy of 65 %, our rule based model achieves better results. A subjective measure is also calculated based on a 5 point scale by different uses. The average satisfaction level of the users is 87% over all the tests conducted.

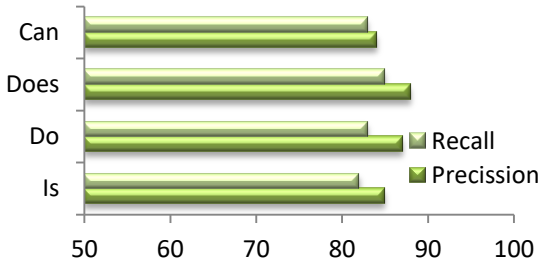


Figure.5: Average score for Aux-type questions

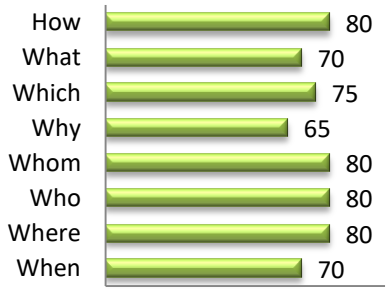


Figure.6: Average precision measure for Wh-type questions

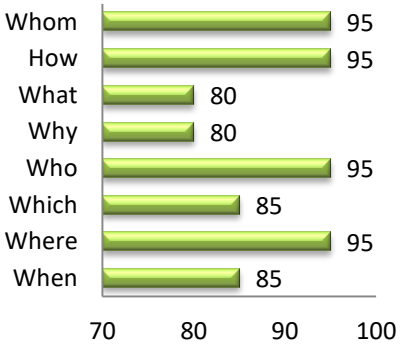


Figure.7: Average recall measure for Wh-type questions

V. CONCLUSIONS

In this paper, an intelligible Rule-based QAS is presented that takes a user queries in natural language and process it to understand the type of answer required by the user. It then analyzes the document contents and extracts the document contents related to the query and present the answer. The model can classify the wh-type question accurately and also includes the Aux- type questions. However, some advanced techniques may be included to achieve better performance. The various morphological word forms can also be addressed to increase accuracy.

REFERENCES

1. S. Stoyanchev, G. Tur, and D. Hakkani, “Name-aware speech recognition for interactive question answering”, In proc: ICASSP 2008, pp. 5113-5116, 2008.
2. A. Pradhan, V. Behera, A. Mohanty, and S. panda, “A Rule-based Information Extraction System", International Journal of Innovative Technology and Exploring Engineering, Vol. 8, No. 9, pp. 1613-1617, 2019
3. C. Mao, L. Li, Z. Yu, L. Han, J. Guo, and X. Lei, "Research on Answer Extraction Method for Domain Question Answering System(QA)", In proc: International Conference on Computational Intelligence and Security, pp. 79-83, 2009.
4. A. Pradhan, V. Behera, A. Mohanty, and S. Panda, “A Voice-based Information Extraction System” In proc: 3rd International Conference on Smart Computing & Informatics (SCI2018), Springer, pp. 593-602, 2019.
5. Z. Zhao, L. Zhang, X. He, W. Ng, "Expert Finding for Question Answering via Graph Regularized Matrix Completion", IEEE Trans. on Know. and Data Eng., Vol. 27, No. 4, pp. 993, 1004, 2015
6. H. Shen, G. Liu, H. Wang, and N. Vithlani, "SocialQ&A: An Online Social Network Based Question and Answer System", IEEE Trans. on Big Data, Vol. 3, No. 1, pp. 91- 106, 2017
7. M. R. Morris, J. Teevan, and K. Panovich, “What do people ask their social networks, and why?: A survey study of status message Q&A behavior,” in Proc. SIGCHI Conf. Human Factors Comput. Syst., 2010, pp. 1739–1748.
8. B. Li and I. King, “Routing questions to appropriate answerers in community question answering services,” in Proc. 19th ACM Int. Conf. Inf. Knowl. Manage., 2010, pp. 1585–1588.
9. Quan X., Liu W., and Qiu B., “Term weighting schemes for question categorization,” IEEE Transactions on pattern Analysis and Machine Intelligence, vol. 33, no. 5, pp. 1009–1021, May 2011.
10. Yang H., and C. Meinel, “Content based lecture video retrieval using speech and video text information”, IEEE Trans. on Learning Tech., Vol. 7, No. 2, pp. 142-154, 2014 .
11. Liu K. , Z. Yuan-Zhe, J. Guo-Liang, L. Si-Wei, and Z. Jun, `”Representation learning for question answering over knowledge base: An Overview,”, Acta Automatica Sinica, vol. 42, no. 6, pp. 807818, 2016.
12. X. Benavent, A. Serrano, R. Granados, J. Benavent, and E. Ves, “Multimedia information retrieval based on late semantic fusion approaches: Experiments on a wikipedia image collection”, IEEE trans. on multimedia, Vol. 15, No. 8, pp. 2009-2021, 2013.
13. L. Wang, Y. Zhang, and T. Liu, `”A deep learning approach for question answering over knowledge base," in Proc. 24th Int. Conf. Comput. Process. Oriental Lang., 2016, pp. 885-892.
14. L. Su, T. He, Z. Fan, Y. Zhang, and M. Guizani, "Answer Acquisition for Knowledge Base Question Answering Systems Based on Dynamic Memory Network", IEEE Access, Vol. 7, 2019.
15. Y. Lan , S. Wang, J. Jiang, "Knowledge Base Question Answering With a Matching-Aggregation Model and Question-Specific Contextual Relations", IEEE/ACM Trans. on Audio, Speech, And Lang. Proc., Vol. 27, No. 10, 2019.
16. H. Jin ,Y. Luo, C. Gao, X. Tang, P. Yuan, "ComQA: Question Answering Over Knowledge Base via Semantic Matching", IEEE Access, Vol. 7, 2019
17. A. Agarwaly, N. Sachdevay, R. K. Yadavy, V. Udandaraoy, V. Mittaly, A. Guptay, A. Mathur, "EDUQA: Educational Domain Question Answering System Using Conceptual Network Mapping", In proc: ICASSP 2019, 8137-8141
18. T. Shao, Y. Guo, H. Chen, Z. Hao, "Transformer-Based Neural Network for Answer Selection in Question Answering", IEEE Access, Vol. 7, 2019
19. S. Zhang , X. Zhang, H. Wang, L. Guo, S. Liu, "Multi-Scale Attentive Interaction Networks for Chinese Medical Question Answer Selection", IEEE Access, Vol. 8, 2018
20. Y. Lin, H. Shen, "SmartQ: A Question and Answer System for Supplying High-Quality and Trustworthy Answers", IEEE Trans On Big Data, Vol. 4, No. 4, pp. 600-613, 2018.
21. A. Shtok, G. Dror, Y. Maarek, I. Szpektor, “Learning from the past: Answering new questions with past answers,” in Proc. International Conference on World Wide Web, pp. 759–768, 2012.
22. Z. Li, H. Shen, G. Liu, and J. Li, “SOS: A distributed mobile Q&A system based on social networks,” in Proc. International Conference on Distributed Computer System, pp. 627–636, 2012.

AUTHORS PROFILE

Soumya Priyadarsini Panda, is currently working as an Assistant Professor in the Department of Computer Science and Engineering, Silicon Institute of Technology, Bhubaneswar, India. She has a M. Tech. and Ph. D. degree in Computer Science and Engineering and has published more than 25 research papers in reputed journals and conferences. Her research interest includes Speech Processing, Natural Language Processing and Machine Learning.

Vikash K. Pandit, is an undergraduate student of Department of Computer Science and Engineering, Silicon Institute of Technology, Bhubaneswar, India. His research interest includes Natural Language Processing, and Machine learning.

Rohit Kumar, is an undergraduate student of Department of Computer Science and Engineering, Silicon Institute of Technology, Bhubaneswar, India. His research interest includes Information Extraction and Machine learning.

Sudhanshu Chaturvedi, is an undergraduate student of Department of Computer Science and Engineering, Silicon Institute of Technology, Bhubaneswar, India. His research interest includes Information Retrieval and Machine learning

Akash Das, is an undergraduate student of Department of Computer Science and Engineering, Silicon Institute of Technology, Bhubaneswar, India. His research interest includes Big Data Analytics, Natural language processing and Machine learning