A Rule-based Information Extraction System

Soumya Priyadarsini Panda, Varun Behera, Alloran Pradhan, Abhisekh Mohanty

Abstract*: *Designing intelligent expert systems capable of answering different human queries is a challenging and emerging area of research. A huge amount of web data is available online and majority of which are in the form of unstructured documents covering articles, online news, corporate reports, medical records, social media communication data, etc. A user in need of certain information has to assess all the relevant documents to obtain the exact answer of their queries which is a time consuming and tedious work. Also, sometimes it becomes quite difficult to obtain the exact information from a list of documents quickly as and when required unless the whole document is read. This paper presents a rule-based information extraction system for unstructured web data that access the document contents quickly and provides the relevant answers to the user queries in a structured format. A number of tests were conducted to determine the overall performance of the proposed model and the results obtained in all the experiments performed shows the effectiveness of the model in providing required answers to different user queries quickly.

Index Terms*: *Information Extraction, Information Retrieval, Multimedia Data, Natural Language Processing, Question Answering System

I. INTRODUCTION

In the last few years, there has been a major change in the technology increasing the requirement of different sophisticated information processing systems. This leads to the development of intelligent systems based on Human Computer Interaction (HCI) technology [1]. Research in the area of language processing enables machines to understand and produce natural language text and speech [2]. The field of Artificial Intelligence added new features to the HCI technology, where the system perceives its environment and takes actions intelligently that maximize its chance of successfully achieving the goals [3]. There has been a sufficient success today in these areas having a widespread area of applications in designing different human–computer interactive systems [4] such as Expert Systems, Question Answering Systems (QAS), Information Retrieval (IR) systems, and talking computer systems [5].

A huge amount of multimedia data is available in the web resources [6]. A significant part of such information is in the form of unstructured text documents [7] and thus it becomes difficult to find the exact information from a list of documents quickly at the hand set unless all the documents are read. It is also a difficult task to develop intelligent methods that access the document content and extract relevant information quickly [8]. Information Extraction (IE) in this aspect focuses on analyzing a collection of unstructured documents and producing cleaned relevant information to the users in a structured format. The IE technology focuses on reducing the human effort and time needed to read the whole document contents to extract required information. It also focuses on extracting important information from documents and presenting it in a structured format. Information Extraction is a rapidly growing field as much of the information in the web is expressed in natural languages [7]. There are fewer researches documented in the field for Information Extraction, However, majority of the models focuses on extracting information from domain specific structured documents. [9] Describes a rule-based information extraction system for medical texts. Even though creating a rule-based system is a time consuming process and requires domain knowledge but are reliable and are useful to automate data processing. However, the model focuses on only clinical data. The authors in [10] discussed about developing a Why–How QAS on community web boards for supporting ordinary people in solving various plant disease problems. The model uses machine learning techniques for question type identification to achieve better accuracy. A review on research on clinical information extraction applications is presented in [11]. A number of literatures [12], [13] are also available that focuses on domain specific information or can handle specific types of queries on limited data set [11]. Limited availability of data increases the challenges in extracting accurate information as and when required [14], [15]. Extracting multi domain information from unstructured text on different types of user queries and presenting the answer in a structured format is still a challenging task and has achieved no significant progress till date. This work focuses on development of IE system for the unstructured web data in text form which covers a diverse range of topics and domain.

In the recent era of technology, users in need of certain information may obtain the details from different web resources easily [16], [17]. However, the web repositories are increasing day by day and most of the documents are unstructured [18]. Extracting answer from those documents requires assessing all related documents [19], [20]. Also, most of the time, users require quick answers to be obtained easily instead of surfing a large repository of documents particularly for the Wh-form of queries [21] [22]. This requires the development of expert question answering systems [23]. Just like asking the questions to a human expert and getting the answers back immediately, the expert system answers the queries of the users [24].

Revised Manuscript Received on July 05, 2019.

Soumya Priyadarsini Panda, Department of CSE, Silicon Institute of Technology, Bhubaneswar, India.

Varun Behera, Department of CSE, Silicon Institute of Technology, Bhubaneswar, India.

Alloran Pradhan, Department of CSE, Silicon Institute of Technology, Bhubaneswar, India.

Abhisekh Mohanty, Department of CSE, Silicon Institute of Technology, Bhubaneswar, India.

This work focuses on development of a rule-based information extraction system which is a hybrid model that combines the ideas of the IE technology and QAS, where the users may ask their queries on different topics and the system answers to the queries quickly like a human expert. For this purpose, a set of hand coded rules are prepared to analyze the question type and generate the template for the expected answer type. Instead of using a local offline document repository, the model uses the online available web document to extract the relevant answers on user queries. The use of the online web resources makes the model more efficient by providing answers which covers a wide range of domains eliminating the limited domain issue of the IE systems. A number of tests were conducted to determine the overall performance of the proposed model on a number of random user queries covering different domains. The results obtained in all the experiments performed shows the effectiveness of the model in providing accurate answers to different user's queries in less time.

The rest of the paper is organized into the following sections. The details of the document repository used as knowledge base for information extraction is discussed in section II. Details of the proposed model are presented in Section III with description of all phases. Various implementation details, the experiments performed and results obtained are presented in Section IV along with the discussions on performance analysis of the model with respect to different type of user queries. Section V summarizes the research findings with a discussion on future directions of the work.

II. DOCUMENT REPOSITORY CREATION

Document repository creation is an integral part of any IE system. As the paper mainly focuses on extracting information from unstructured web data, to create the knowledgebase for the system, a collection of articles covering different domains are downloaded from the web. A collection of 1000 documents repository is created covering different medical records, social media interactions and streams, online news, government documents, corporate reports, stock exchange, gold price, petrol price, etc. All the files are stored in an offline repository in .pdf format. However, to make the model more dynamic the online wiki repository is also used to extract the answers of the user queries which are not present in the offline repository.

III. PROPOSED MODEL FOR INFORMATION EXTRACTION

This section presents the description about the proposed Rule-based Information Extraction System (RIES) which extracts the answers to the user queries quickly and present the answer in a structured template. The overview of an intelligent Question Answering System is presented in Figure1. The overview of the proposed model is presented in Figure. 2 and the details of the phases involved are discussed below.

A. Query Processing

The query processing unit processes the text utterances by a text pre-processing process which involves removal of unnecessary words to obtain useful information for further use. The Stemming technique is also applied for eliminating various affixes (suffixed, prefixes, infixes, circumfixes) from the key terms present in the queries. The lemmatization process is applied on the processed text to capture canonical forms. The keywords are then extracted and classified into various named entities such as name of person, location, organizations, time, etc.

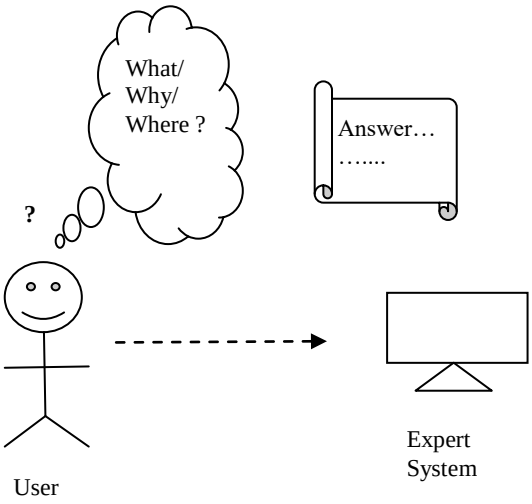

Figure.1: Overview of intelligent Question Answering System

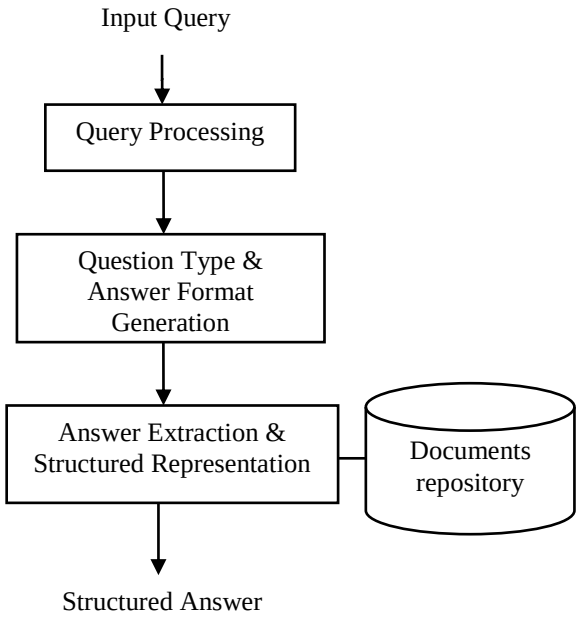

Figure.2: Overview of Proposed Information Extraction System

B. Question Type Detection and Answer Format Generation

The expected type of answer to any user query depends on the type of the question. Usually majority of the user queries focuses on extracting answers to the 'Wh-' form questions like- what, when, where, why, which, who queries.

The answers of those types of queries contain the information about the location, place, date, time event, people, etc. Considering this idea, a set of classes are formed covering all Wh-type questions with its type and sub classifications. For each of these classes, the expected answer types are determined. The classifications of the Wh-questions and their expected answer types are presented in Table.1. Based on these classifications, a set of rules are prepared to identify the type of questions entered by the users and the answer format is generated. Algorithm-1 is used to obtain the keywords for information extraction and the respective Tag sets for answer extraction. The expected answer type is then determined and the answer template is generated using Algorithm-2. The answer template is generated based on the obtained tag set for the expected answer format and the user query.

Table.1: Question type classification and expected answer format

Wh- Question Type	Expected Answer Type
When	Date
Where	Location
Who	Person
Whom	Person
Why	Reason/Description
Which-Who	Person
Which-Where	Location
Which-When	Date
What	Price/Number/ Definition/ Title
What-Who	Person
What-When	Date
What-Where	Location
How	Description
How-Many	Number
How-Much	Value
How-Long	Time/Distance

Algorithm-1:

Step-1: Apply word Tokenizer on Query Q and obtain word tokens Wt for identifying the keywords.
Step-2: Apply NER Tagger on each word token Wt and create the tag set.
Step-3: Determine the question class by processing the 'Wh' word in the query.
Step-4: Determine expected answer type using Algorithm 2

Rule set for answer type generation:

R1: IF ques_class is 'when' THEN answer_type is DATE
R2: IF ques_class is 'where' THEN answer_type is LOCATION
R3: IF ques_class is 'who' THEN answer_type is PERSON
R4: IF ques_class is 'whom' THEN answer_type is PERSON
R5: IF ques_class is 'why' THEN answer_type is DESCRIPTION
R6: IF ques_class is 'which' and sub class is 'who' THEN answer_type is PERSON
R7: IF ques_class is 'which' and sub_class is 'where' THEN answer_type is LOCATION
R8: IF ques_class is 'which' and sub_class is 'when THEN answer_type is DATE
R9: IF ques_class is 'what' and sub_class is 'who' THEN answer_type is PERSON
R10: IF ques_class is 'what' and sub_class is 'when' THEN answer_type is DATE
R11: IF ques_class is 'what' and sub_class is 'where' THEN answer_type is LOCATION
R12: IF ques_class is 'what' and sub_class is 'number' or 'price' THEN answer_type is VALUE
R13: IF ques_class is 'how' and sub_class is 'many THEN answer_type is NUMBER
R14: IF ques_class is 'how' and sub_class is 'much' THEN answer_type is VALUE
R15: IF ques_class is 'how' and sub_class is 'long' THEN answer_type is TIME | DISTANCE
R16: IF ques_class is 'how' and sub_class is 'description' | 'reason' THEN answer_type is DESCRIPTION

Algorithm-2:

Step-1: Determine the answer type for Query Q using Rules R1 to R16
Step-2: Generate the answer template from the answer type and query

C. Answer Extraction & Structured Representation

In order to extract the relevant answers to the user's queries, the Wiki online document repository is used which covers a huge collection of topics covering different domains. Based on the obtained keywords and tag sets by Algorithm-1 and Algorithm 2, the document contents are accessed using the web scraping method and the respective answers are extracted and filled in the generated answer template using Algorithm 3.

Algorithm-3:

Step-1: Find the online wiki document for the obtained keywords.
Step-2: Apply tokenizer and NER Tagger to document contents with respect to the generated NER tag set.
Step-3: Extract the relevant data from the wiki document
Step-4: Fill the answer template with the extracted information in respective tag position to provide the structured output or answer to the user query.

IV. EXPERIMENTATION AND RESULT ANALYSIS

The proposed model is implemented in Python and for processing the input text queries, the NLTK (Natural Language Toolkit) tool is used. To obtain the features from text about different entities spaCy is used. In order to extract the information from online document repository, the wiki web resources is considered. For this purpose, the web scraping method which focuses on transforming unstructured data on the web into structured data is used.

A set of questions covering different domains are prepared to evaluate the performance of the presented model. The proposed model (RIES) is also tested on a number of user's queries asked by different users and the results obtained are evaluated. The user queries are of the form- what, when, where, who, whom, which, why, how. The performance is evaluated based on the precision and recall measure [10], which are presented in equation (1) and equation (2) respectively. The average precision and recall measure for 100 example questions covering the wh-questions of the form: when, where, who, whom, why, which, what and how is presented in figure 3 and figure 4 respectively. In each of the tests conducted, the model successfully obtained the answer to the query. The average precision measure is 75% while average recall measure is 89% as shown in figure 5.The proposed model achieves relatively better results in all the experiments performed.

$$\text{Precision} = \frac{\text{Number of relevant answers}}{\text{Total number of answers extracted}} \ldots\ldots (1)$$

$$\text{Recall} = \frac{\text{Number of relevant answers obtained}}{\text{Total number of available answers}} \ldots\ldots (2)$$

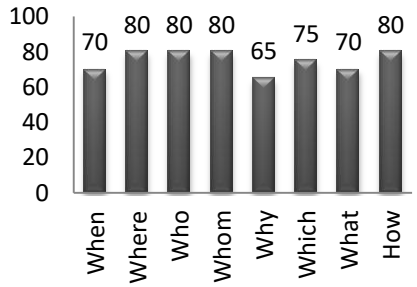

Figure.3: Average precision measure of RIES

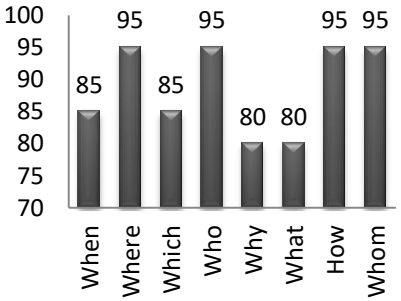

Figure.4: Average recall measure of RIES

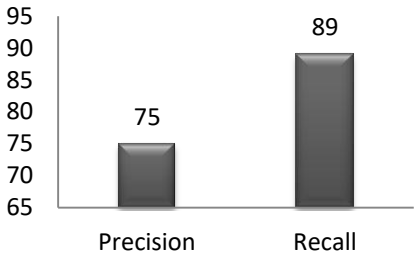

Figure.5: Average precision and recall measure of RIES

A subjective evaluation test is also conducted to judge the level of user satisfaction. A five points Mean Opinion Score (MOS) scale (1: poor, 2: average, 3: average, 4: satisfied, 5: highly satisfied) is used for this purpose [25], where, the users are asked to rate the IE system based on their experience on a set of random questions of the Wh-category. The average MOS results for all considered Wh-type of questions is presented in figure 6 which shows the level of satisfaction of the users of the IE system on their random queries.

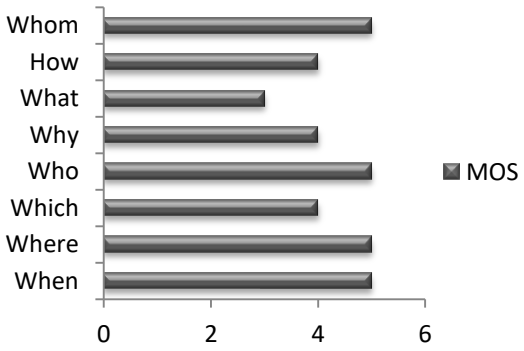

Figure.6: Average MOS score on RIES performance

Even though, the model achieves high quality results in terms of the recall measure in all the experiments performed, still the percentage of accuracy in terms of precision is less. As the model only considers the offline repository and online wiki documents, the system is unable to extract the answers to some specific queries whose information are not available in the document repository. Therefore, the model may further be enhanced to access other online information by referring to other web links instead of only the wiki link so that the domain restriction can be overcome and the model can achieve more better results.

V. CONCLUSIONS

This work presents an intelligent Rule-based Information Extraction System (RIES) that analyzes the document contents and presented the answer in a structured format. The model is tested on a number of user's queries and in each of the tests conducted; the model successfully obtained the answer to the query. The model can classify all types of wh-type question accurately. However, the model may further be enhanced by using machine learning techniques to increase accuracy.

Also, the model may further be extended to address the auxiliary verb type questions as well as the questions with conjuncts. The variability of inflectional forms may also be included to address various morphological word forms.

REFERENCES

1. R. Glauber, and D. B. Claro, "A Systematic Mapping Study on Open Information Extraction", Expert Systems with Applications, Vol. 112, pp. 372-387, 2018.
2. S. P. Panda, and A. K. Nayak, "An efficient model for text-to-speech synthesis in Indian languages", International Journal of Speech Technology, Vol. 18, No. 3, pp. 305-315, 2015.
3. K. Rajbabu, H. Srinivas, and S. Sudha, "Industrial Information Extraction Through Multi-Phase Classification using Ontology for Unstructured Documents", Computers in Industry, Elsevier, Vol. 100, pp. 137-147, 2018.
4. S. P. Panda, and A. K. Nayak, "A waveform concatenation technique for text-to-speech synthesis", International Journal of Speech Technology, *Vol. 20, No.* 4, pp. 959-976, 2017.
5. G. Chen, C. Wang, M. Zhang, Q. Wei, and B. Ma, "How Small Reflects Large? -Representative Information Measurement and Extraction", Information Sciences, Elsevier, Vol. 460, pp. 519-540, 2017.
6. D. Dupplaw, M. Matthews, R. Johansson,… and A. Moschitti, "Information extraction from multimedia web documents: an open-source platform and testbed, International Journal of Multimedia Information Retrieval, Springer, Vol. 3, pp. 97-111, 2014.
7. M. R. Bouadjenek, H. Hacid, M. Bouzeghoub,"Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms", Information Systems, Vol. 56, pp. 1–18, 2016
8. Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical Information Extraction Applications: A Literature Review", Journal of biomedical informatics, Elsevier, Vol. 77, pp. 34-49, 2017.
9. A. Mykowiecka , M. Marciniak, A. Kups´c, "Rule-based information extraction from patients' clinical data", Journal of Biomedical Informatics, Vol. 42, pp. 923–936, 2009.
10. C. Pechsiri, R. Piriyakul, "Developing a Why–How Question Answering system on community web boards with a causality graph including procedural knowledge", Information Processing in Agriculture, Vol. 3, pp. 36–53, 2016.
11. F. S. Alotaibi, V. Gupta, "A cognitive inspired unsupervised language-independent text stemmer for Information retrieval", Cognitive Systems Research, Vol. 52, pp. 291–300, 2018.
12. Z. A. Merrouni, B. Frikh, and B. Ouhbi, "Toward Contextual Information Retrieval: A Review And Trends", Procedia computer science, Vol. 148, pp. 191-200, 2019.
13. J. Lai, Y. Mu, F. Guo, P. Jiang, and W. Susilo, "Privacy-enhanced attribute-based private information retrieval", Information Sciences, Vol. 454, pp. 275-291, 2018.
14. Y. Djenouri, A. Belhadi, P. Fournier-Viger, and J. C. W. Lin, Fast and effective cluster-based information retrieval using frequent closed itemsets. *Information Sciences*, Vol. *453*, pp. 154-167, 2018.
15. R. C. Santos, and M. A. S Machado, "FIRDoR-Fuzzy information retrieval for document recommendation", *Procedia computer science*, Vol. *139*, pp. 56-63, 2018.
16. A. Rorissa, and X. Yuan, "Visualizing and mapping the intellectual structure of information retrieval", Information processing & management, Vol. 48, No. 1, pp. 120-135, 2012.
17. S. Karimi, J. Zobel, and F. Scholer, "Quantifying the impact of concept recognition on biomedical information retrieval", Information Processing & Management, Vol. 48, No. 1, pp. 94-106, 2012.
18. X. Liu, and C. Hu, "Research and design on e-government information retrieval model", Procedia Engineering, Vol. 29, pp. 3170-3174, 2012
19. D. Martinez, G. Pitson, A. MacKinlay, and L. Cavedon, "Cross-hospital portability of information extraction of cancer staging information", Artificial intelligence in medicine, Vol. 62, No. 1, pp. 11-21, 2014.
20. A. Aljamel, T. Osman, G. Acampora, A. Vitiello, and Z. Zhang, "Smart Information Retrieval: Domain Knowledge Centric Optimization Approach", IEEE Access, Vol. 7, pp. 4167-4183, 2018.
21. H. Yang, and C. Meinel, "Content based lecture video retrieval using speech and video text information", IEEE Transactions on Learning Technologies, Vol. 7, No. 2, pp. 142-154, 2014
22. X. Benavent, A. Garcia-Serrano, R. Granados, J. Benavent, and E. Ves, "Multimedia information retrieval based on late semantic fusion approaches: Experiments on a wikipedia image collection", IEEE transactions on multimedia, Vol. 15, No. 8, pp. 2009-2021, 2013.
23. H. M. Kim, and A. Sengupta, "Extracting knowledge from XML document repository: a semantic Web-based approach", Information Technology and Management, Vol. 8, No. 3, pp. 205-221, 2007.
24. X. Liu, C. Wan, and D. Liu, "Keyword query with structure: towards semantic scoring of XML search results", Information Technology and Management, Vol. 17, No. 2, pp. 151-163, 2016.
25. M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale", Computer Speech and Language, Vol. 19, pp. 55–83, 2005.

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 20 research papers in reputed journals and conferences. Her research interest includes Speech Processing, Natural Language Processing and Machine Learning.

Varun Behera is an undergraduate student of Department of Computer Science and Engineering, Silicon Institute of Technology, Bhubaneswar, India. His research interest includes Natural Language Processing, Big data Analytics and Machine learning.

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

Abhisekh Mohanty 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