Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 14 (2020) Pp-282-286

BBRC
Bioscience Biotechnology Research Communications

Comparative Analysis of Keyword and Semantically Enhanced Question Answering System on Law Domain

Shubhangi C. Tirpude*[1] and Devishree Naidu[2]

[1, 2]*Shri Ramdeobaba College of Engineering & Management, Nagpur, India*

ABSTRACT

As compared with the traditional approach to a search based on keyword, semantic analysis and semantic based search are advanced techniques which understands the linguistic of the search query & makes it accurate intelligent search engine for any domain. We have developed a Question Answering system that considers the semantic information of question inputted and answer retrieved focussing on retrieval of information which is context based. The question answering system is based on syntactical & semantical analysis by creating the semantic graph and defining the semantic relationship between semantic entities. We have designed closed domain question answering system on law documents dataset which answers the queries related to law domain. We tested the queries using keywords matching approach and compares it with by considering the semantic involved in the query, & we found that the semantic based approach produces the result with high accuracy than keyword-based approach, because it considers the Conceptualization and user intents involves in the user's query.

KEY WORDS: QUESTION ANSWERING SYSTEM (QAS), SEARCH ENGINE, COMPUTATIONAL LINGUISTIC, SEMANTIC WEB, N GRAM, FEATURE VECTOR.

INTRODUCTION

The Question Answering (QA) system interprets the question specified in natural language and returns the correct information(answer) using collection of documents. A lot of research has been done in the keyword-based information retrieval which retrieves correct answer of the query based on keyword matching. But the problem with this approach is that if two or more queries with same keywords but different meanings will give the same result because it doesn't focus on understanding the meaning of the query posed in natural language. Even most of the search engines like Bing, Yahoo etc. are continually identifying & enhancing the new features to increase the user experience [1]. Even retrieving the data form the large repository of documents & finding the accurate & correct answer is a complex task in terna of complexity and time. To solve this problem, we have developed the semantically rich Question Answering Systems.

The semantic based information retrieval understands the meaning of the query which improves the accuracy of the information retrieval. To understand the semantic of the query we need to define the complex structure. The lot of research has done in keyword-based information retrieval where only the keywords are matched. But the problem with this approach is that if the two queries having the same keywords but have different meaning, then for both the queries will give the same result. If we use the semantic based approach, it understands the meaning & user indentation involved in the query which retrieves the accurate answer

ARTICLE INFORMATION

**Corresponding Author:* tirpudes@rknec.edu,naidud@rknec.edu

Received 19th Oct 2020 Accepted after revision 25th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31)
A Society of Science and Nature Publication,

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.14/66

Literature Review: Miriam Fernandez, Iván Contador, Vanesa López, David Valet, Pablo Castells, Enrico Motta (1) suggested the fully-fledged ontologies in the semantic-based perspective. The paper describes a semantic search model which integrates the benefits of keyword and semantic-based search and addresses the challenges of the huge and diverse web environment. The target search space is defined as a collection of unstructured content. The results describe that as compared to the best TREC automatic system; the semantic search model attains better performance results.

Athira P. M., Sreeja M., P. C. Reghu Raj (2) describes suitable methods which process the complex questions by enhancing the capabilities of current QA system. They have used the ontology and domain knowledge for redeveloping queries and detecting the relations. The system will generate short and precise result to the question asked in the natural language in a specific domain. The system will be implemented and result shows accuracy of 94 % in natural language question answering.

Maksym Ketsmur,, Mário Rodrigues, and António Teixeira[3] design a QA system to knowledge bases such as DBpedia, using factual questions in Portuguese, English, French and German. The system was tested with 30 random questions from QALD 7 (Question Answering over Linked Data) training set. Considering that the answer existed in the knowledge base, a correct answer was produced for 67% of the questions for the Portuguese version and up to 55% (for English) of times for multi-language version. Results proved that this approach is promising and further investigation should be carried out to improve it. The robustness observed, and capability to handle several languages, fosters future work to expand the system to answer

Table 1. Comparisons of different types of Semantic based QA System.

Type of Question and Answering System	Methods used	Dataset or Corpus	Result
Semantic question answering model for question answering system [1]	User modeling & relevance feedback methods are used for semantic question analysis	GeoBase Ontology dataset consisting of 880 annotated user questions of US Geographical information.	Compare with Aqualog and FREya& achieved 0.947 f- measure
Ontology based question answering system on software test document domain [2]	Model view controller pattern is used	software test document domain	Retrieves answers to factoid type questions
Ontology based question answering using semantic similarity matching [3]	Blooms Taxonomy questions of various levels of are generated (i. e. low to high, slight to large, modest to composite)	Open domain questions using Google	Find different patterns for the same questions
Open Domain Real-Time question answering based on semantic & syntactic question similarity [6]	Latent based question similarity, WorldNet based semantic & syntactic similarity	Yahoo answer corpus site includes topics as Art & Humanities, health, home, sports & travel etc.	Extract answer in less than 1000 characters in less than 60 sec. 899 questions answered out of 1088 questions
Towards a Question Answering system over the semantic Web [9]	Multilingual KB-agnostic approach	Knowledge base Wikipedia, DBpedia MusicBrainz, DBLP & Freebase	Uses five different languages as English, German, French, Italian & Spanish for the evaluation of system

Vivek Datla, Sadid A. Hasan, Joey Liu, Yassine Benajiba Kathy Lee, Ashequl Qadir, Aaditya Prakash, Oladimeji Farri[5] implements a real-time question answering system using the syntactic and semantic similarity on open domain. The system defines real time user questions. These questions are extracted from the stream of most recent questions and it is given to the participants via a socket connection. The systems in turn provides an answer with length of 1000 characters in less than 60 seconds. These answers are evaluated by the human expert in terms of accuracy, legibility, and precision and so that the correctness of the generated answers is checked. For generating the answers, question disintegration, question relatedness and answer generation strategies are used. The following table shows details about the question answering system and their working.

MATERIAL AND METHODS

The proposed system retrieves the correct answer by increasing at the semantic level, the linguistic knowledge and also the better understanding of the domain used.

The System is divided into 3 components:
i) Query understanding
ii) Document Retrieval
iii) Ranking and exact answer selection

To develop a question answering system, we used the data set of law domain. We develop the corpus of the 500 files related to various sections & articles of law domain. To create the knowledge base, the basic concept is to identify the syntactic information which gives us the lexical construct like noun, verb and other terms. The noun, verb and the adjective keywords are analysed with the semantic meaning using WordNet with hyponym & synonym of the semantic entities if any exists.

Figure 1: Proposed Approach

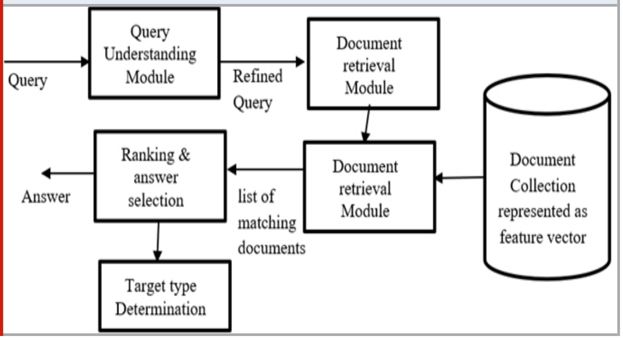
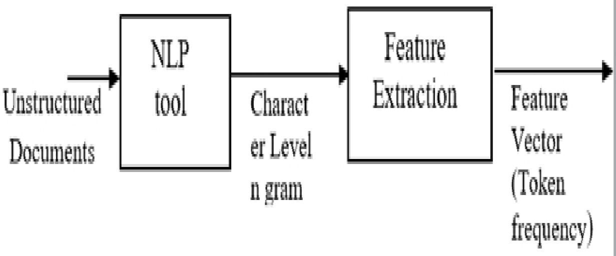


Figure 2: Creation of Corpus

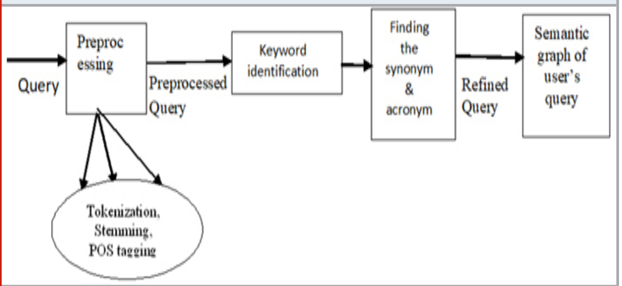


Corpus: The corpus of the text files related to the IPC sections are created. To create the knowledges base of the files for the fast retrieval, the pre-processing like tokenization, stop word removal, stemming etc. be carried out. After the pre-processing, we have to identify the keyword present in the files & creating the dictionary of keyword. It also determines, for each keyword the number of documents & the documents where that keyword is present. To do this, we divide each document into n-grams (sequences of n tokens). For Natural Language Processing (NLP) tasks, n-grams are often computed at the word level, but we found that computing it at the character level gives better performance in our source code task. Another advantage of character level model is that we don't have to tokenize the text or don't have to break it into words. For each document the frequency of each possible token is computed. We convert the character level bi-gram combined with other n-grams to an array of numbers by counting the occurrences of the token sequences in the document. We call this numeric representation of a document the feature vector. Next, we will create the semantic graph of the all the documents with important keyword & also maintains the semantic relationship between them.

Query Understanding: The query(structured/unstructured) is inputted to the system. Understanding the semantic of query is important to get the accurate result. Initially the query gets analysed and identifying the keywords. The complex query is divided into small parts for better finding the semantic. After that the query get expanded to find its synonym so that the query with the different meaning of the same keyword is not to be missed out. Also, the acronym expansion carried out expanding the terms like "IPC" to "Indian penal code". Now the query gets classified so that the target of query will be determined. The target type indicates what type of answer is expected to retrieve. The target type which we considered here is "what", "when", "yes/no" short descriptive" & "factual". After the classification, the refined query is the passed to the Document retrieval module.

To understand the semantic, we created the semantic graph which captures the internal structure such as syntactic & semantic which determines the semantic relationship between the semantic entities. It then searches the documents using the semantic graph created. Some pre-processing task that we apply are tokenization, lemmatization, stop words removal and stemming for better matching the documents.

Figure 3: Query Refinement Process



Question Dataset: We need the question dataset to train our system to deal with the closed domain, The question data set of 150 questions was designed related to total 511 IPC section.
Examples:
1. What is punishment for attempt to murder
2. Which IPC is applied for attempt to murder?
3. Under which section the offences affecting the human body and punishment for the crime are defined.

Document Retrieval: The keywords in query is gets matched with the documents represented by feature vector. All the relevant documents get identified & retrieved based on keyword present in the query. The union of all the keywords present in the query & thus it finds the relevant documents. They will be extracted from the documents stored in our data corpus. Below is the screenshot of search module representation: Both the keyword based & semantic based search is taken into account. For semantic based search, the semantic graph of the query is considered, which gets matched with the documents represented by semantic graph.

Figure 4: Search Module

Ranking & Answer Selection: For the factual questions, what type of question, explanative answers, fact-based questions, the system determines precise answers to the query. The ranking of the documents is provided based on the semantic matching. For the keyword-based search, the result is less effective compared to that of semantic based search.

For example:
If the query 1 is "Which punishment is given for attempting to murder?" & query 2 is "Are there any punishment for murder?" Since the keyword defined in both the queries are same, it returns the same answer for both the queries. But as the first query is descriptive type & second query is Yes/ No type. So the answer also expected to be of descriptive type answer & answer in the form of yes/no. So, by understanding the user intent & conceptualization behind the query, it retrieves the accurate result. We have specified classification of the questions based on the target of each question depends on the tokens present in the query as per the following table:

As the documents are represented as the feature vector, this feature space is transformed into the latent semantic space. In this new space, we used the similarity to find the relationship between the queries and documents. We have used the popular ranking model, BM25 model for ranking the documents. After retrieving the collection of passages from the various documents based on the target type, the next task is the exact answer section for the given query. The ranking will return the documents with highest relevance, then next highest up to the lowest relevance. We use the graded relevance to determine the measure of usefulness & accordingly selection of the final answer carried out.

Table 2. Target type and tokens

Target type	Tokens
Description type	What, Define, give reason, Suggest, tell us, what happens
List type	List various section, List the punishments, List the IPC
Yes/No type Factual type (When type) Location based type (Where type)	Whether, Can, Is, Would, Will When, how long, how much, Where, at

RESULTS AND DISCUSSION

The system is evaluated on the basis of Precision, recall & F-Measure, which is most commonly used metrics for evaluating the performance of the information retrieval. The following table indicates the comparison of the results of keyword-based approach & semantic based approach. The proposed system is tested with 150 different questions which are in structured form.

Table 3. Experimental Results showing precession & recall of both keyword & semantic approach.

	Keywords based search	Semantic based search
Question dataset used for testing the system	150	150
Answers generated by the system	95	125
The correct answer generated	90	120
Precision	0.94	0.96
Recall	0.6	0.8

FUTURE SCOPE AND CONCLUSION

The system is tested for 150 question (structured and unstructured) on the various evaluation parameters, showing the accurate & precise results for the semantic based search. In future, we can implement the summarization which will summarize the results of the top 3 retrieved results so as to get detail description of the query. We can use the concepts for developing the question answering system for COVID-19 data set to answer COVID-19 related queries.

REFERENCES

Miriam Fernández, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, Enrico Motta "Semantically enhanced Information Retrieval: an ontology-based approach", Journal of web Semantics, volume 9, issue

4 December 2011, pp 434-452.

Athira P. M., Sreeja M., P. C. Reghu Raj "Architecture of an Ontology-Based Domain-Specific Natural Language Question Answering System" , International journal of Web & Semantic Technology(IJWesT), volume 4,Number 4, Oct.. 2013, pp 31-39.

Maksym Ketsmur, Mário Rodrigues, and António Teixeira "DBpedia Based factual question Answering System" IADIS International Journal on WWW/ Internet, Vol 15, No. 1, ISSN 1645-7641, December 2017, pp 80-95

Iryani Saany , Ali Mamat, Aida Mustapha, Lilly S. Affendeyand M. Nordin A. Rahman "Semantics Question Analysis Model for Question Answering System" Applied Mathematical Sciences, Vol. 9, no. 130, 2015, pp 6491 – 6505 .

Vivek Datla, Sadid A. Hasan, Joey Liu, Yassine Benajiba, athy Lee, Ashequl Qadir, Aaditya Prakash, Oladimeji Farri "Open Domain Real-Time Question Answering Based on Semantic and Syntactic Question Similarity" Artificial Intelligence Laboratory, Philips Research North America, Cambridge, MA, USA

Harb, Hany, Fouad, Khaled, M. Nagdy, Nagdy, "Semantic Retrieval Approach for Web Documents" International Journal of Advanced Computer Science and Applications volume 2, number 9, September 2011, pp 67-76

Eckhard Bick. "The Parsing System PALAVRAS Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework". Aarhus University Press, ISBN 87 7288 910 1, 2014.

Miltiadis D. Lytras, Naif R Aljohani, Ernesto Damiani, Kwok Tai Chui "Semantic Web Search Through Natural Language Dialogues" In book: Innovation, Developments, and Applications of Semantic Web and Information Systems", Chapter: 12, Publisher: IGI Global, 2018, pp.329-349

Dan Su1, Yan Xu1, Tiezheng Yu1, Farhad Bin Siddique, Elham J. Barezi, Pascale Fung1 "CAiRE-COVID: A Question Answering and Multi-Document Summarization System for COVID-19 Scholarly Information Management," Centre for Artificial Intelligence Research (CAiRE), Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP, Association for Computational Linguistics The Hong Kong University of Science and Technology, 2020,