# Domain Specific Question Answering System

[1]Rohini P. Kamdi, [2]A. J. Agrawal

M.Tech Scholar, Associate Professor
Computer science & engineering SRCOEM, Nagpur, Maharashtra, India
Email : [1]rohinikamdi29@gmail.com, [2]avinashjagrawal@gmail.com

**Abstract - Question Answering (QA), in information retrieval, is the task of automatically answering a question posed in natural language (NL) using either a pre-structured database or a collection of natural language documents. As with the excessive information growth in the web, retrieving the exact fragment of information even for a simple query, it requires large and expensive resources. Additionally the need to develop exact systems gains more importance due to available structured knowledge-bases and the continuous demand to access information rapidly and efficiently. The domain specific question Answering System gives suitable solution for this. This paper proposes the closed domain QA System for handling the legal documents of IPC sections and Indian Laws to retrieve more precise answers.**

## I. INTRODUCTION

In early 90s, the first Question Answering task in TREC 8 (Text Retrieval Conference), revealed an increasing need for more sophisticated search engines able to retrieve the specific piece of information that could be considered as the best possible answer for the user question. Such systems must go beyond document selection, by extracting relevant part. They should either provide the answer if the question is factual or yield a summary if the question is theoretic.

Question Answering (QA) is an area of natural language processing research aimed at providing the users with a convenient and natural interface for accessing information. The typology of question relied on 13 categories. To each category was associated a search strategy of the answer in the knowledge base. The textual database replaces the knowledge base in the previous works. An information retrieval based system exploiting only statistic knowledge of the corpus leads to the elaboration of a system able to answer less than half of the questions.

The problem intersects two domains: Information Retrieval (IR) and Natural Language Processing (NLP). IR is improved by integrating NLP functionalities at a huge scale, i.e. independently of the domain, and necessarily having a large linguistic coverage. This integration allows the selection of the relevant passages by linguistic features at the syntactic or even semantic level. NL document collections used for QA systems include: a local collection of reference texts, a set of Wikipedia's pages and a subset of World Wide Web pages. QA deals with a wide range of question types which includes: fact, list, definition, How, Why, hypothetical, semantically constrained, and cross-lingual questions.

This paper is classified as follows: Section 2 comprises basic elements of QA System; section 3 comprises related work in Question Answering, section 4 fallows proposed approach and section 5 gives the conclusion.

### A. Basic Elements Of QA System

Every QA System has the basic elements for implementation as:

1) Question Processing

The input to the Question Processing is question asked by the user. The Question Processing captures the semantic of question that is for what the question is asked by the user. The Question Processing has three tasks as:

a) Determining the question type
b) Determining the answer type
c) Extracting keywords from the question and formulate a query.

a) Question Types
There are five classes of questions according to the answers as:

| | |
|---|---|
| Class 1 | Answer: single datum / list of item<br>C: who, when, where, how (old, much, large) |
| Class 2 | A: multi-sentence<br>C: extract from multiple sentence |
| Class 3 | A: across several text<br>C: comparative/contrastive |
| Class 4 | A: an analysis of retrieved information<br>C: synthesized coherently from various retrieved fragment |
| Class 5 | A: result of reasoning<br>C: word or domain knowledge and common sense reasoning |

Figure.1. Classes of Questions

b) Types of QA

There are two types of Question Answering systems according to the domain of answers as:

Closed-Domain QA System: Closed-domain question answering deals with the questions under a specific domain, and can be seen as an easier task because NLP systems can exploit domain-specific knowledge frequently formalized in ontologies. It has very high accuracy but requires extensive language processing and limited to one domain. The example of such a system is medicines or automotive maintenance.

Open-Domain QA System: Open-domain question answering deals with the questions about nearly everything, and can rely on general ontologies only and world knowledge. And these systems usually have much more data available from which to extract the answer. It can potentially answer any question but has very low accuracy as the domain is not specific.

c) Keyword Selection

The keywords are helpful for finding the relevant text in question to give a specific answer. For better matching, these keywords can be expanded with lexical or semantic alternations like, the word “producer” can be taken as “produce”, the phrase “has been sold” can be taken as “sell” and the specific category as “dog” can be referred as “animal” for keyword selection. Also some words based on importance are focused for keywords like non-stopwords in quotations, all complex nominal (plus adjectives), all other nouns, all verbs (not focus on tense), and potential answer type.

2) Document Retrieval

From the keywords that selected, a query is formulated and is given in the Passage retrieval component. In this, all the passages are extracted that contains the selected keywords. The quality of passage depends upon the loops. It follows some simple heuristic algorithms to decide whether the certain keyword is added or dropped for candidate answering text.

For example, if in the first iteration, it uses the initial 6 keywords selection heuristics it fallows the algorithm like: if the number of passages is less than a threshold then the query is too strict, therefore drop a keyword otherwise if the number of passages is greater than a threshold then query is too relaxed, and therefore add a keyword.

The ranking of passages is done by constructing the keyword windows in which; it searches how many times certain keywords are found in the passages. The passage scoring is depends upon

- The number of question keywords obtained in the same sequence in the window.
- The number of keywords separating the most distant keywords in the window.
- The number of unmatched keywords.

According to passage score, more relevant passage is selected for an answer. The passage retrieval component deals with document retrieval from the database for extracting the passage that contains the candidate answer text.

3) Answer Extraction

In the answer extraction, the representation of the question and the representation of candidate answer bearing texts are matched against each other to give a specific and correct answer. From this set of such candidate answers are produced and then ranked according to the likelihood of correctness. The answer ranking features are:

- Question term numbers matched in the answer passage.
- Question terms numbers matched in the same phrase or sentence as the candidate answer.
- Number of question terms matched, separated from the candidate.
- Number of terms occurring in the same order in the answer passage as in the question.
- Average distance from the candidate answer to the question term matches.

## II. RELATED WORK

The open domain QA System [1] described the use of Wikipedia as a rich knowledge source in a question answering system with multiple answer matching modules based on different types of semi-structured knowledge sources of Wikipedia, including article content, infoboxes, article structure, category structure, and definitions. These semi-structured knowledge sources each have their unique strengths in finding answers for specific question types, like as infoboxes for factoid questions, category structure for list questions, and definitions for descriptive questions.

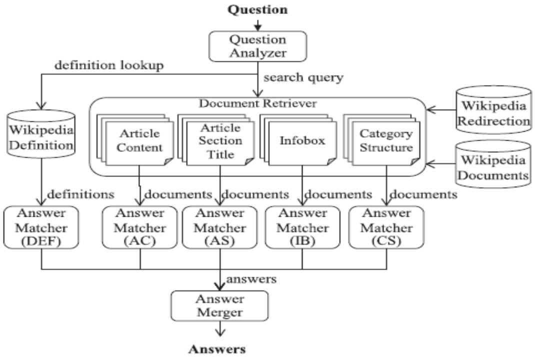

Figure.2. System overview [1]

In this for Question Analysis, questions in natural language form are analyzed using multiple linguistic analysis techniques, including POS tagging, chunking, and named entity tagging[12] and then analyzed result into the form of answer format (AF), answer theme (AT) and question target (QT). The AF has three possible values as factoid, list, and descriptive, an AT is the class of the object or description sought by the question and A QT consists of two parts as object that the question is about and property of interest that a question attempts to get at regarding the object.

For retrieving an answer, it selects the best answer a for given question q that maximizes the multiplication of question analysis score SQ(r|q), document retrieval score SD(d|r) and the answer matching score SA(M)(a|q,r,d) where r, a, d are question analysis result, answer candidate and retrieved document, respectively and scores are normalized between 0 and 1. The answers extracted from multiple modules are merged using an answer merging strategy that reflects the specialized nature of the answer matching modules. The main motivation behind this work was to devise a way to utilize the existing semi-structured, large-size Wikipedia database as a knowledge source for a QA system without building high-cost knowledge base.[1]

For semi-structured knowledge-bases QA System [2], a new architecture to develop a factoid question answering system based on the DBpedia ontology and the DBpedia extraction framework. Dbpedia is a project that aims at extracting information based on the semi-structured data presented within the Wikipedia articles.

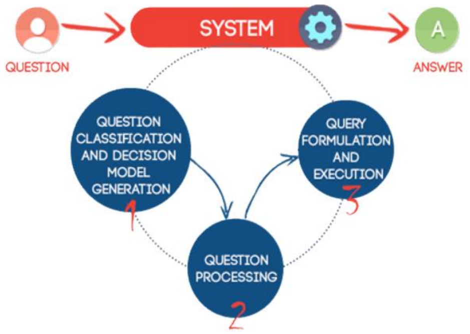
QUESTION
SYSTEM
A
ANSWER
QUESTION CLASSIFICATION AND DECISION MODEL GENERATION
1
QUESTION PROCESSING
2
QUERY FORMULATION AND EXECUTION
3

Figure.3. Global Question Answering System Architecture [2]

This paper [2] is divided into 3 parts as Question Classification and Decision Model Generation**,** Question Processing and Query Formulation and Execution.

(i) Question Classification and Decision Model Generation: It is given by two components; first component is the question classifier which pre-processes the question dataset and trains it by the SVM algorithm [2], [11] which is a binary classifier giving two classes as the coarse-grained classes and the fine-grained classes for their proposed QA System. And the second component is the decision model generator which fallows tokenization, stop words removal and features extraction using the bag of words. Then it creates trained and tested file for estimating accuracy.

(ii) Question Processing: It allows identifying the question type of given question by extracting resources using DBpedia spotlight tool and extracting keywords using processing.

(iii) Query Formulation: It involves Ontology Class Determination, which determines ontology classes and their properties used to construct the final SPARQL query and the result of the query is an RDF file holding ontology classes and properties, Query formulation used to retrieve the answer from DBpedia composed from resources and ontology classes determined by the keyword set and Execution in which the system interrogates the DBpedia Server to get a response as an RDF file and then parsed to get the answer of the given question.

Authors in [3] given the system which finds answers of Malayalam factual questions by analyzing a repository of Malayalam documents for handling the four classes of factual questions in Malayalam for closed domain. The QA system is divided into three modules as Question Analysis, Text Retrieval and answer snippet extraction and Answer identification.

(i) Question Analysis: It takes single sentence level questions as an input. The aim of this module is to identify the question word, the query and expected set of answer templates and fallows the NLP algorithm for preprocessing.

(ii) Text Retrieval and answer snippet extraction: Based on the query words the answer candidates are retrieved from the document collection for answer identification.

The document collection is indexed and, which has total keyword match with the question are selected for answer snippet extraction. For this, it checks the count of match of the query with each sentence. The sentences which have a fuzzy match to the query are selected as the answer candidates are represented using a triplet containing the sentence, index and count of the match. The index is used to extract the actual sentence. The index value is assigned at the time of text splitting and count of the match gives the value of match with the question. These answer candidates are passed to the next module selection of answer candidates,

(iii) Answer identification: It has two sub-modules as,

Scoring and Ranking of the answer candidates which performs the scoring and ranking, and selects the winner candidate using matching window sizes. This answer candidate which has the highest score is selected as the winner candidate and this snippet is further processed for answer extraction. And the second is Answer Extraction using Named Entity Recognition in which the expected named entity of the question is identified by analyzing the question word and then nearest surrounding words of the question word are analyzed to identify the expected answer entity.

Jibin Fu in the paper [4] proposed a music knowledge question answering system on the ontology knowledge base through which the users can ask a question about music knowledge in natural language, and the system automatically extracts relative knowledge to give answer based on FAQ and ontology knowledge base. It has three processes as:

(i)Question Classification: It uses the ontology and improved Bayesian-based method [15]. First, the concepts in user's questions are extracted in support of ontology knowledge base and then the frequency of terms calculated using "word-bag" model for finding class of question.

(ii) FAQ and Question Analyzer: Frequently asked questions are stored in FAQ module which can quicken the processing. The similarity of user's question and question in FAQ candidate question set is computed. If user's question can't match in the FAQ, the question is transferred to question analyzer module in which, question template method is used to extract semantic representation for a simple question and for complex question and abnormity question; keyword association method is used for probability of semantic representation. For each template, its semantic representation is extracted, once a question can match a question template, the semantic representation of the question can be located.

(iii)Answer Extraction: It fallows two strategies: In first, one directly match question with the question in FAQ, for frequently asked question, and in second strategy, for a question not included in FAQ, analyze the question and extracts the answer in support of ontology and logic reasoning. First has higher priority than strategy two.

The relative concepts are extracted from ontology and the relations between concepts are reason and knowledge point is extracted to form answer.

## III. PROPOSED APPROACH

Using the literature survey of Question Answering Systems, we can say can that the closed domain QA System is more accurate than the open domain QA System. If we see scenario of queries related to legal documents of IPC sections and different Indian laws, there is no such QA system, which ensures the correct answers. The user generally asks the query in the unstructured form, for example "If Ram killed Shyam, then punishment to Ram". Also for same query there can be different answers as like for previous and the question "Charges for murder". So, the idea to develop the Question Answering System for IPC Sections and Indian Laws is proposed as shown in figure 4.

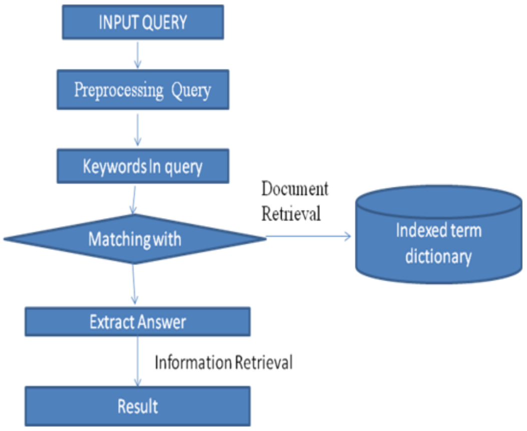

Figure.4. Block Diagram

An input for the system will be a query related to IPC sections or different Indian laws like constitution amendment, parent amendment, company amendment laws etc. The Question Information Extractor gives the specific target and constraints of the question which is used to extract the answer. The corpus of IPC sections and Indian laws document is generated to form indexed term dictionary as metadata knowledge base storing the related keywords of each document. Using these keywords, the original passage or sentences are tagged to give candidate answers from answer extractor. According to given question, the constraint and candidate answer matched against each other and highest score probable answer is retrieved as a final answer. The system will produce the accurate answer for trained questions and then will test to measure the accuracy of untrained questions

## IV. CONCLUSION

Question Answering requires more complex NLP techniques compared to other forms of Information Retrieval. QA Systems can be developed for resources like web, semi-structured and structured knowledge-base. The Closed Domain QA Systems give more accuracy in finding answers but restricted to single domain only.

The QA system for closed domain of legal documents of IPC sections and Indian Laws using machine learning approach and information retrieval is proposed to give the accurate and suitably more correct answers for user's structured or unstructured queries in efficiently.

## ACKNOLEDGEMENT

The authors are grateful to the Head, Computer Science and Engineering Department and the Principal, Shri Ramdeobaba College of Engineering and Management for providing adequate facilities to conduct a research. Authors are thankful to the faculty members of

Computer Science and Engineering Department for their support and cooperation during this work.

## REFERENCE

[1]. 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.

[2]. 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.

[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]. Jibin Fu, Keliang Jia and Jinzhong Xu, “Domain Ontology Based Automatic Question Answering”, 2009 International Conference on Computer Engineering and Technology

[5]. 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

[6]. Perera, Rivindu (2012) “IPedagogy: Question Answering System Based on Web Information Clustering”, In Proceedings of the 2012 IEEE Fourth International Conference on Technology for Education (T4E '12). IEEE Computer Society, Washington, DC, USA

[7]. Menaka S and Radha N. “Text Classification using Keyword Extraction Technique”, in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013

[8]. MatthewW. Bilotti and Eric Nyberg,” Improving Text Retrieval Precision and Answer Accuracy in Question Answering Systems”, the 2nd workshop on Information Retrieval for Question Answering (IR4QA), pages 1–8 Manchester, UK. August 2008

[9]. 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

[10]. Eric Brill, Susan Dumais and Michele Banko, “An Analysis of the AskMSR Question-Answering System”, Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 257-264. Association for Computational Linguistics.

[11]. Yaoyong Li, Kalina Bontcheva, and Hamish Cunningham, “SVM Based Learning System For Information Extraction”, Department of Computer Science, The University of She_eld, She_eld, S1 4DP, UK

[12]. 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.

[13]. W.A.Woods, R.M. Kaplan, B.L. Nash-Webber, “The Lunar sciences natural language information system: Final report”, Technical Report BBN Report 2378, Bolt Beranek and Newman Inc., Cambridge, MA, 1972.

[14]. Lee, C., Hwang, Y.-G., Oh, H.J., Lim, S., Heo, J., Lee, C.-H., et al (2006). “Fine-grained named entity recognition using conditional random fields for question answering”. In Proceedings of Asia information retrieval symposium (pp. 581–587).

[15]. ZHANG Yu, LIU Ting, WEN Xu, "Modified Bayesian Model Based Question Classificatio", 2005, vol.19, pp. 100-105.

[16]. 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

◈◈◈