Abstract

Question Answering (QA) is a specialized area in the field of Information Retrieval (IR). The QA systems are concerned with providing relevant answers in response to questions proposed in natural language. QA is therefore composed of three distinct modules, these three core components are: question classification, information retrieval, and answer extraction. This survey paper provides an overview of Question-Answering and its system architecture, as well as the previous related work comparing each research against the others with respect to the components that were covered and the approaches that were followed. At the end, the survey provides an analytical discussion of the proposed QA models. The question answering system deals with both structured & unstructured query which retrieves answer paragraph as most suitable answer.

1. Introduction

The Text Retrieval Conference (TREC), a conference series co-sponsored by NIST, initiated the Question-Answering Track in 1999 which tested systems’ ability to retrieve short text snippets in response to factoid questions revealed an increasing need for it revealed 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.

Question Answering (QA) is a fast growing research area that combines research from different, but related, fields which are Information Retrieval (IR), Information Extraction (IE) and Natural Language Processing (NLP). 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. It delivers only the requested information unlike search engines which refers to full documents. As with the excessive information growth in the web, retrieving the exact fragment of information even for a simple query 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. QA research attempts to deal with a wide range of question types including: fact, list, definition, how, why, hypothetical, semantically constrained, and cross lingual questions

There are two types of Question answering Systems

- Closed-domain question answering: closed-domain question answering deals with questions under a specific domain (music, weather forecasting, Tourism, Medical etc.) The domain specific QA system involves heavy use of natural language processing systems formalized by building a domain specific ontology.
- Open-domain question answering :Open-domain question answering deals with questions about nearly everything and can only rely on universal ontology and information such as the World Wide Web.(2). The domain specific Question Answering System gives more specific and correct answers than web based QA system as it is restricted for only one domain resource to answer.

2. Question Answering System Overview

As shown in (Figure 2), a typical QA system consists of three distinct modules, Query Processing Module” whose heart is the question classification, the “Document Processing Module” whose heart is the information retrieval, and the “Answer Processing Module” whose heart is the answer extraction.

A. Question Processing Module
Given a natural language question as input, the overall function of the question processing module is to analyze and process the question by creating some representation of the information requested. The Question Processing has 3 tasks as:
- Determining the question type
- Determining the answer type
- Extract keywords from the question and formulate a query
B. Keyword Selection
The keywords are helpful for finding the relevant text in question to give specific answer. For better matching, these keywords can be expanded with lexical or semantic. Also some words based on importance are focused for keywords like non-stop words in quotations, all complex nominal’s (plus adjectives), all other nouns, all verbs (not focus on tense), and potential answer type.

C. Document Processing Module
From the keywords that selected, a query is formulated and is given to the component that is document retrieval. In this, all the passages are extracted that contains the selected keywords.

D. Answer Processing Module
In the answer extraction, the representation of the question and the representation of candidate answer bearing texts are matched against each other to give specific and correct answer. From this set of such candidate answers are produced and then ranked according to likelihood of correctness.

3. Literature Review
Open domain question answering using Wikipedia-based knowledge model [2] It describes the use of Wikipedia as a rich knowledge source for a question answering (QA) system with multiple answer matching modules based on different types of semi-structured knowledge sources of Wikipedia, including article content, info boxes, article structure, category structure, and definitions. These semi-structured knowledge sources each have their unique strengths in finding answers for specific question types, such as info boxes for factoid questions, category structure for list questions, and definitions for descriptive questions.
For semi-structured knowledge-bases QA System [3], 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.
Authors in [4] 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.
Jibin Fu in the paper [5] 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.
Bhoir V & M. A. Potey[6] proposed solution of Question Answering system works for a specific domain of tourism. The crawler developed in the system gathers web page information which is processed using Natural Language Processing and procedure programming for a specific keyword. The system returns precise short string answers or list to natural language questions related to tourism domain like distance, person, date, list of hotels, list of forts, etc.
Dell Zhang and Wee Sun Lee[7] describes a Web-based question answering system LAMP, which is publicly accessible. A particular characteristic of this system is that it only takes advantage of the snippets in the search results returned by a search engine like Google. Because it is time consuming.
to download and analyze the original web documents. The performance of LAMP is comparable to the best state-of-the-art question answering systems. Following table shows details about the literature survey of Question Answering System:

<table>
<thead>
<tr>
<th>QA system</th>
<th>Working</th>
</tr>
</thead>
</table>

4. Proposed Approach
In current scenario, there is No QA system for exactly answering the queries on legal documents of Indian laws. Also the answers differ from Wikipedia and other sites for legal documents for same question and the accuracy for unstructured query by users is very less. The proposed system for current scenario is to create such a closed domain QA system that retrieve the answer related to question efficiently. For this the Corpus or Knowledge base is constructed from legal documents of Indian Laws. Then the matching of terms in queries with possible candidate terms in each document to retrieve the best answers stored in corpus. The system will extract and gives the specific and correct answer to query (structured/unstructured) by user.
Closed Domain Question Answering System - A Survey

The Figure illustrates Proposed Approach

Input Query (Structured/Unstructured) → Preprocessing Query (Stop word Removal, Stemming, Tokenization) → Extract Keywords → Keywords Matched With Metadata (Indexed Term Dictionary) → Processed Documents → Preprocessing Documents (Stop word Removal, Stemming) → Document Corpus

The approach is divided into following steps:

**Input**: Structured/Unstructured Query.

Step 1: Pre-processing of Query.
Step 2: Extract the keywords by TF-IDF calculation
Step 3: Match the keywords with Knowledge base for Document Retrieval
Step 4: Extract the Answer using Searching Technique
Step 5: Correctness for derived candidate answers checked with Classifier algorithm.

**Output**: Specific and correct answer with high accuracy.

**4.1 Corpus**: The first phase of the closed domain Question Answering System is to create Corpus or Knowledge base. A corpus of Indian Laws is created in Text file format.

**4.2 Pre-processing**: Major tasks in pre-processing are tokenization, stop words removal and stemming.
Tokenization: Perform linguistic tokenization which includes each word and also the numbers as single token.

Stop words Removal: Removing stop words reduce the dimensionality of term space. The most common words are in text documents are prepositions, articles, and pro-nouns etc that does not provide the meaning of the documents. These words are treated as stop words. Example for stop words: the, in, a, an, with, etc. Stop words are eliminated from documents because those words are not considered as keywords in text mining applications.

Stemming: Stemming is most important process by which the different forms of word is replaced with basic root word.

4.3 Indexed Term Dictionary
The processing of text involves two main problems, first problem is the extraction of feature terms that become effective keywords in the training phase and then the second is actual classification of the document using these feature terms in the test phase. After pre-processing the TF-IDF is calculated for keyword term in a document [8]. The TF-IDF is stored as a table as the Term (keyword) ,count & document id. It represents that the term is appeared in the number of files & the count of the term in each file.

4.4 Information Retrieval
It has 3 components as:

- **User Query**: user will enter the query related to domain in structured or in unstructured form.
- **Extract Keywords**: From the user query, the keywords are extracted. These keywords are obtained by removing the symbols and stop words from user query. Also the stemming is applied on keywords so as to match with term indexed dictionary terms for document retrieval.
- **Document and Passage Retrieval**: The keywords so obtained from query are matched in Term Indexed Dictionary to find the document ids where these keywords are present. 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 present only those document are to be retrieved for candidate answer passages. For Ex The term law is present in file 1, 5, 8 & the term murder is present in 1, 4, 6, 8. It retrieves all the files as file 1, 4, 5, 6, 8 the keywords that obtained from user query are matched in documents and where it meets with threshold value, it gives the certain passage from text file corpus. It can give one or more paragraphs which satisfies the keywords matching
- **Answer Representation**: After Retrieving all the passages containing the terms, only top results are shown based on some threshold value. For Ex. If the QA System retrieves 15 passages that likely to contain the answer or sometimes duplicate passages. So the filtering is applied to present the answer. If threshold value is 6, then only the first 6 results are presented.

5. 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 classifier is proposed for retrieving the accurate answer efficiently for user’s structured or unstructured queries.

For future work, it retrieves the exact answer so that accuracy of QA system will be increased

References

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