Applied GIS

ISSN: 1832-5505
Vol-6 Issue-03 July 2018

Data mining-based intelligent question answering

Kalyani Shinde*1, Anjika Singh*2, Pallavi Singh*3, Reshma Shitole*4, and Prof. Sumeet Harale*5
Department of Computer Technology, Indira College of Engineering and Management, Pune, Maharashtra, India

***Abstract*—** In recent years, there has been a lot of interest in the difficult problem of question answering (QA). Using a structured database or a corpus of natural language texts, QA systems attempt to provide the most relevant response to questions posed by humans in natural language. One of the most important steps in QA is selecting the best possible response from a pool of candidates. The ultimate answer(s) may be chosen in many ways, but one frequent method is to see it as a ranking issue. There have been several suggested techniques so far, most of which aim to create a single, optimal ranking model that can be applied uniformly to all question formats. In contrast, this research suggests a data mining strategy for rating candidate solutions and picking the best one. In QA systems, the TF-IDF method is used to determine which top-ranked response should be chosen. The Knowledgebase (KB) is searched for the most relevant results by utilizing similar queries. To increase the quality of the best response delivered to a human, an experimental session was conducted over a series of questions in the Cultural Heritage domain using a manually annotated gold-standard dataset.

***Keywords*—** *Question Solving, Natural Language Processing, Data Mining, and Databases.*

INTRODUCTION

Because it is seen as the foundation of the next generation of digital assistants like Apple's Siri, Google Assistant, Amazon Alexa, and Cortana, and of artificially intelligent systems like IBM's Watson [8], question answering system (QA) is a quite difficult task that has garnered considerable attention in recent years. Quality assurance system, often known as an HMC. QA is able to deliver a decisive, correct, and clear response, often in the form of a brief text type answer [6], to questions posed by different users in their natural language. In comparison to other methods, a QA system is able to provide a single answer from collection of related documents given by some keywords [8], and this has attracted the attention of researchers looking for automatic ways to get answers to natural language questions submitted by users in different languages.

The idea behind a QA system is that users may type in a question, and the system will provide the best possible response based on the question's category and other factors. Take the question, "Which vitamin is present in milk?" as an example. [2].

Many other, related fields may benefit from the concepts and procedures evolved from question answering, including document fetching, time, and named-entity expression identification. While researchers first concentrated on "factoid" inquiries, they have since shifted their emphasis to "complex" questions such as "definitional," "entity," "date type," "location," and "person" [3].

RELATED WORK

Many people in the field of AI are interested in creating a Question Answering System. Many programmers have already set up quality assurance infrastructure for both open-domain and closed-domain software. In contrast, a closed-domain QA system [4] only fields inquiries about the specified domain. When we have a query on a certain topic, but don't know where to go for the answer, we may turn to a closed-domain system for help. Therefore, the knowledge base compiles the responses to such queries and stores them in a database. The most relevant result from the database is returned throughout the response retrieval process. Template matching is used to infer closed domain systems. Answers to questions from any topic may be found in an open-domain question-answering system. For this purpose, these systems use the use of search engines [5].

In the study [2], the authors provide a closed-domain QA system that pulls knowledge bases (KB) from various sources on the internet. Closed domain QA systems are more precise than open domain QA systems, but they can only be used with data from a single domain. Different steps such as corpus collection and research, index term dictionary construction, question pre-processing, document retrieval, keyword ranking, answer extraction, and answer are all used in this NLP-based article.

The kinds of QA systems are described in [3], and both the implementation of a Web-based QA system and its assessment are shown. The development of a QA system is one of the most recent discussions in the field of NLU. Web-based QA has significant potential since it enables users to pose queries in more natural language than is possible with a keyword-based search engine. Author presents IPedagogy QA system in article [7] that uses natural language questions to get responses from specified information clusters, thereby narrowing the IR search area. There are a number of natural language processing algorithms included into IPedagogy that enable the system to be guided to the precise response to any question. The assessment of the system using mean reciprocal rank yields an average accuracy of 0.73 on 10 sets of questions, each of which consists of 35 questions.

The author of article [1] developed a novel approach to developing a Chinese question-and-answer system that searches for solutions to queries posed in Chinese by analyzing and processing text written in that language. It discusses the four uses of Chinese closed-domain questions. Each of these subsystems—Question Processing, Document Retrieval, and Answer Selection—has its own dedicated function inside our QA system.

THE PROPOSED APPROACH

The architecture of a system to answer closed-domain queries in an unstructured knowledge base using the TF-IDF algorithm is shown here. There are four distinct components to the quality assurance system: question processing, knowledgebase processing, document retrieval, and answer selection.

1) Handling the Question

In this step, the inquiry content is parsed in order to extract a

ISSN: 1832-5505
Vol-6 Issue-03 July 2018

collection of keywords that will be used to build a list query for IR. It consists of many parts.

a) Tokenization, also known as keyword extraction, separates the query into its constituent parts that have significance.

Stop word removal is a filter that determines the inquiry type and then removes any words from the query that have a poor information content [8].

For the sake of optimal document retrieval, the list is subsequently transformed to lowercase.

d) Query Formulation: This step involves compiling a list of search terms to be used in a query against the KB.

2) Preparedness for KB

Unstructured text files (.txt) are the focus of the KnowledgeBase under consideration for this quality assurance system. Pre-processing the papers is necessary for the system to function well. When using the QA system, the pre-processing is only done once every cycle.

For the first part of the lowercase conversion procedure, the document is transformed into a list. All the words in the list will soon be lowercase after another round of lowercase conversion.

b) Tokenization, which disassembles sentences into meaningful words and phrases. The list is where the keywords are saved [4].

c) Remove Stop Words: Doing so reduces the number of dimensions in the term space. Prepositions, articles, pro-nouns, etc., are the most prevalent terms found in text documents, although they do not contribute to the texts' content. In order to improve the performance of information retrieval systems, "stop words" are removed from texts. By keeping an English stop word dictionary [8], we can ensure that common phrases like "is, for, the, in, etc." are removed from all of the dataset's text files.

Thirdly, Data Mining

The papers that are pertinent to the subject at hand are obtained and processed in the module devoted to document processing.

Module of Document Processing Steps

a) Exact Phrase Matches: Tokens produced by the question processing module are compared to the pre-processed KB to determine which document most closely resembles the query. The answers are filtered by this function first before being returned from the pre-processed KB. This function's output is transformed into a list for convenient further processing.

All instances of the terms in the list generated by the previous module will be returned by the function you provide in b). In this function, we do a second round of filtering, during which we get rid of any duplicate responses and try all possible permutations and combinations until we get the best one.

a) All nouns and Non-Nouns: According to different categories and position of the verbs, nouns, adjectives, etc the system is trying to understand the fetched answers to eliminate the wrong answer. Understanding the nature of question different parts of speech POS were added. This is again the third filtration to find the best answer through the system. This phase helps next module to rank and decide the best answer.

2) Answer Selection:

This module is made of a collection of components, each one assigning a score to the candidate answers. These scores are employed as possible features to be processed by the proposed TF-IDF approach for ranking answers. The selection of the best answer is chosen from the TF-IDF algorithm. The use of this algorithm is for ranking the retrieved answers. The best possible answer is given by this which is the final product of QA system.

a) Term Frequency - Inverse Document Frequency algorithm:

TF-IDF is one of the simplest ranking functions which is computed by summing the Term Frequency and Inverse Document Frequency for each answer retrieved. Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Inverse Document Frequency, measures how important a term is like "is", "of", and "that", may appear a lot of times but have little importance. Thus, we need to weigh down the frequent terms. [7]

In traditional TF-IDF the idea is: if a word appears in a document with high frequency and rarely appears in other documents, then thinks that the word has the very good category to distinguish ability and suitable for classification. Commonlyused TF - IDF formula is as follows:

$$w(t,d) = tf(t,d) + \log[N/N_t]$$

$tf(t,d)$, represents the frequency of the keyword t appear in text d, N_t represents the number of full texts, represents the numberof the texts which have the word. [1]

EVALUATION OF QA SYSTEM

A. Collection and Study of corpus:

The closed domain considered for this work is about Cultural Heritage of the city of Naples. This corpus of unstructuredknowledge is of 16 documents taken from reference paper [8]. The set of questions for system testing consists of 200 questions, along with associated possible correct answers. In particular, 40 questions are of type Description, 30 questions are of type Date, 50 questions of type Entity, 40 questions of type Location and 40 questions of type Person.

ISSN: 1832-5505
Vol-6 Issue-03 July 2018

B. Table for evaluation

Factors	Questions Type				
	Date	Entity	Location	Person	Description
Recall	0.88	0.8750	0.8857	0.8857	0.8824
Precision	0.91	0.8974	0.9118	0.9118	0.9091
F-measure	0.8603	0.8806	0.8985	0.8985	0.8955
Accuracy	0.833	0.82	0.8250	0.8250	0.8205

Fig. 1. Accuracy testing for

Recall, Precision are used to calculate accuracy of each type of questions. 200 questions are asked to system to check QA system then the overall accuracy of our QA system is going to be 82.47%.

CONCLUSIONS

The objective of this paper is to review some of the methods, technologies and implementation techniques which are used for implementing QA system. QA Systems can be developed for resources like web, semi-structured and structured knowledge- based domain. The Closed domain QA Systems give more accurate answer than that of open domain QA system but this system is restricted to single domain only. The QA system for closed domain of documents of related to education acts using NLP techniques and information retrieval are proposed to give the accurate and suitably more correct answers for user' queries.

An experimental evaluation for a collection of questions on the Cultural Heritage domain, using a manually annotated gold-standard dataset, has shown that for each question type gives the possibility of improving the accuracy of the best answer returned back to the user. This system results in the betterment of answers along with an increase in accuracy of the answer. The

further work is to improve the system on the basis of security for online availability of system and implement in e-learning environment as well as tries to improve the accuracy and speed of correct answer delivery.

V ACKNOWLEDGMENT

We would like to sincerely thank the faculty of Computer Department of Indira College of Engineering and Management Pune for the helping and guiding us. We express our sincere gratitude to Prof. Sumeet Harale, faculty, Computer Department, Indira College of Engineering and Management Pune for his support. This research was financially supported by Cultural Heritage of the city of Naples, Institute for High Performance Computing and Networking (ICAR) National Research Council of Italy (CNR) Naples, Italy.

VI REFERENCES

[1] Wei Wang and Yongxin Tang: Improvement and Application of TF-IDF Algorithm in Text Orientation Analysis International Conference on Advanced Material Science and Environmental Engineering (AMSEE 2016)

[2] Sweta P. Lende, Dr.M.M. Raghuwanshi: Question Answering System on Education Acts Using NLP Techniques, IEEE Sponsored World Conference on Futuristic Trends in Research and Innovation for Social Welfare (WCFTR'16).

[3] Deepa Yogish, Prof. Manjunath T. N., Prof. Ravindra S Hegadi: A Survey of Intelligent Question Answering System Using NLP and Information Retrieval Techniques, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 5, May 2016.

[4] Veena G., Deepa Gupta, Athulya S., Salma Shaji: A Graph-Based Relation Extraction Method for Question Answering System 2017 IEEE.

[5] Prof. Dhanshri Patil, Abhijeet Chopade, Pankaj Bhambure, Sanket Deshmukh, Aniket Tetame," A Proposed Automatic Answering System for Natural Language Questions", International Journal of Engineering and Computer Science ISSN:2319-7242, Volume 4 Issue 4 April 2015, Page No. 11310-11312.

[6] Wenpeng Lu, Jinyong Cheng, Qingbo Yang: Question Answering System based on Web, 2012 Fifth International Conference on Intelligent Computation Technology and Automation.

[7] Rivindu Perera "IPedagogy: Question answering system based on web information clustering" 2012 IEEE Fourth International Conference on Technology for Education.

[8] Pota, M., Esposito, M., De Pietro, G.: Learning to Rank Answers to Closed-Domain Questions by using Fuzzy Logic In: IEEE Institute for High Performance Computing and Networking (ICAR) National Research Council of Italy (CNR) Naples, Italy (2017)

[9] http://www.tfidf.com/