A Restricted Domain Medical Question Answering System

Vijeta[1], Basant Verma[2]

[1]Student, Computer Science & Engineering, Institute Of Engineering and Technology, Alwar, India
[2]Associate Professor, Computer Science & Engineering, Institute Of Engineering and Technology, Alwar, India

Abstract: ***Physicians have many questions when diagnosis patients, and frequently need to seek answers for their questions. Information retrieval search systems typically return a list of documents in response to a user's query. In this aspect web search engine designed for search of information on internet, give sets of relevant documents related to user query .Thus this makes flooding of documents where user spends healthy time to scan and read all documents to get desired outcome. So there is great need to automate the overall process, so that user gets answer in the form of compact text, rather than list of documents containing answer of user query. This aims at developing medical question answering system which is quite similar to search engine but rather composite in its methodology.***

Keywords: Information Retrieval System, Question Answering System, Knowledge Base, Web Pages.

1. Introduction

Information Retrieval refers to process of analyzing, structuring, storing, accessing information from large data set. In this modern age information happens to be invaluable asset. Nowadays, Medical Information Retrieval (MIR) is getting much attention [7]. Systems that access resources on the basis of medical context are having many applications, both in the academic, personal or commercial domains, these applications refer to the context they belong. With the increasing usage of the internet into daily lives, it is an important need to get medical-specific information from the local database to satisfy the user's information requirements. There is large collection of the text based information which exists in structural form [database] and unstructured form [report, book, and article]. In this regard, Web Search Engine which is aimed for search of information on internet returns sets of documents queried by the user. Hence, user spends his time to scan all returned list of documents and has to go through each line to get desired answer. Thus this causes swamping of documents where more time is needed to get desired outcome. Hence, there is a great requirement to automate the whole process, so that user receives final answer in compact and precise text, rather than list of documents containing answer of user query. This aims at developing question answering system [QAS]. Question Answering System (QAS) performs the task of retrieving its correct answer in a compact form from large text based documents [1], [3], [11]. Researchers and engineers are trying to develop QAS as another tool for fetching information from the set of documents or web.

The QAS question answering system is divided into two systems as follows:

- Open Domain QAS
- Restricted Domain QAS.

Open domain QAS deals with user question in almost every aspect. It is the QA system which is able to generate the final answer to the user's question which is directly derived from unrestricted-domain knowledge base. "Restricted-domain QAS deals with question which falls under a specific domain such as medical, tourism etc. Restricted-domain QAS is able to generate final answer from only restricted domain knowledge source that has a specific domain. Here in this thesis we are confined to restricted domain i.e. medical domain.

2. Related work

A Question answering system generates answer to the user question automatically. Since the 1970's, several QAS have been developed. One such QAS is BASEBALL, developed by Green, Chomsky, and Laughery [7]. This system provides information related to baseball league played in America over one particular season. Another system LUNAR designed by Woods [16], gives answer to the question regarding analysis of soil samples taken from Apollo lunar exploration. Both the system produced good results but they were very specific in their domain. The user query was transformed into database query hence its respective answer was generated. The knowledge source for both the system was database containing significant information about their respective domain.With the rapid use of the internet into user's daily lives, it is imperative to capture medical-specific information from local database to satisfy the user's local information needs. Hence the system would be beneficial for users searching answers to medical queries from documents. Text mining methods are being successfully used in MIR to catch and ascertain medical references in text documents [8], [20]. The data present is noisier and more versatile in form, and there is great deal of misspellings, multilingualism and acronyms. So to automatically deduce what the user aspires to, given a search query, without putting the loads on the user himself, remains an open text mining problem [6].

The Answer Extraction module does extraction of the possible candidate answers, and further does ranking of the candidate answers before finally presenting the best answer to user.

There is lots of Question answering system developed in 21st century named as:

- Webclopedia [17]
- AnswerBus [18]
- WEBCOOP [5]

3. System description

Steps involved in medical domain question answering system are:

- Data Description of offline corpus [Knowledge Base]
- Preprocessing of text based documents.
- Tagging of Document.
- Classification of Question
- Extraction of Answer
- Ranking of Candidate Answers based on similarity check

A. Data Description (Knowledge Base)

The knowledge base is the main source of data in our Question Answering System. This knowledge base could be in structured or unstructured form. Generally in DBMS, type of a structured data can be easily accessed. But the large amount of information storage occurs in form of the text files which are unstructured [15]. In past, Question Answering Systems were interfaced against the DBMS. The resources present as web files are being already indexed so as to retrieve the documents containing the answer is not a cakewalk. The main handy task is to find out the answer from a text document. For building knowledge base we collected information from Wikipedia [14] and other sites which are divided into various headings which are referred as context information of the disease. The updating can be easily performed manually by taking information from net, and should be placed in the document falling under appropriate headings as mentioned above.

B. Pre-processing Of Text Based Documents

There are various steps employed before doing tagging as a part of preprocessing of document, these are as follows

- Removal of Noise
- Tokenization of Document Sentences
- Splitting of Sentences

Noise removal from documents- Noise removal transforms the raw document into the document which comprises of only content related to subjects of the document (not contain images, header, footer and other irreverent text). The Text is the sequence of characters. Tokenization is fragmenting a string of characters into its lexical elements such as words or punctuations [11]. Sentence splitting is the process in which we do splitting of the words and punctuation into their separate sentence [11].

C. Tagging of Document

This module does tagging of useful information from knowledge source to bring out the cosine information. For the specific document we used name of disease as document name. We thereby used different tools which are outside systems for doing document tagging.

- Parser (Stanford Parser) [13]
- Word Net [12].

The parser is component of compiler or interpreter that considers the grammatical construction of sentence, it can find out which words are used as subject, verb or object, phrases [3], [19]. We used Stanford Parser, a tool which generates the Parts of Speech (POS) of each word of the inputted user query and the candidate answers selected from documents. The Word Net characteristics are utilized to point out the relationship existing in between words of user query and data source. Documents but we use its features to help tagging of documents and with other modules in our system. Words are grouped in one of these four categories: nouns, verbs, adjectives and adverbs. WordNet is utilized in most question answering systems as it is a useful tool when dealing with words.

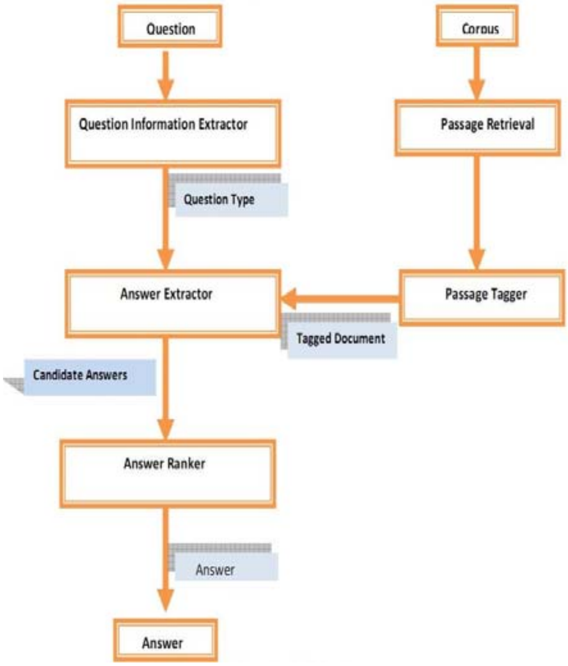

Figure 1: Proposed Architecture of QAS

D. Classification of Questions

Our system categories efficiently the user question into following categories [4], [9]:

- When
- What
- Who
- Which
- Where
- How

Table 1: Question Classifications

Question Type	*Answer Type*
When- Day	Date NUMBER
How-Treatment	TREATMENT
Who/Whom-Person	PERSON
What-Definition	DISEASE
What-Symptoms	SYMPTOMS
What-Cure	CURE
What-Treatment	PRECAUTIONS
Where medical	LOCATION

E. Extraction Of Answer

The Answer Extraction module does extraction of the candidates answer, and further does ranking of the candidate before finally presenting it to user[2],[9].

F. Ranking of Answer

Answer Ranking module perform ranking for each candidate answer and show it to user based on the ranking [10]. This answer ranking module should only produce exact answer. We calculated the similarity check between each word of inputted user question with each word of the candidate answer if the Part Of Speech of both words is same. Thus, answer to the user query is generated.

4. Evaluation and Experiment

For purpose of evaluation we developed an offline corpus. The documents in this corpus contain information about various diseases such as Tuberculosis, Stress and Gastritis. This database is collected from internet. The document is searched on the basis of disease name which is basis for the name of the particular document. As our system deals with question (When, Who, Where, Which, How, What), so system generate answers to these questions. We do subjective evaluation of our system. The purpose of is to find out the user's satisfaction, related to the system. Our colleagues' were asked to mark whether the system gives relevant answer or not.

5. Results and Discussion

The overall system performance of the system is up to 70% from questions staring with WH words. The answer extractor uses the answer type of the answer, and extracts all the sentence of that type. All the candidate answer passed to the answer ranker. Evaluation of this module will be based on how many times an answer is unsuccessfully extracted when the answer is known to be in the retrieved documents.

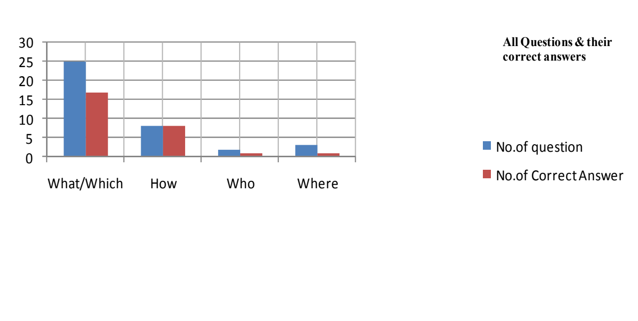

Figure 2: Comparison of Question Types and Correct answer

6. Conclusion

Our question answering system does extract most relevant answer to the question from the large document which contains information about different diseases. Our approach of question answering system firstly classifies the query, then extracts the candidate answers and finally ranks these candidate answers. The most significant part of our approach is Question Classifier. Our system categorizes questions based on question type and extracts answer also based on question type. When the user posed the question, the system extracts the answer information from the tagged document passages in terms of named entities. From these named entities some will represent the candidate answers of the question. The other part of the system assigns the weights to the candidate answers and ranks the candidate answers according to the weighs which gives the best accuracy with the use of WordNet [12]. The achievements of the present work are, We have developed a Semantic Based Question answering system which extracts the answer of a user query from the Our system is flexible, we can add information related to a disease under the appropriate sub-headings (context) to keep it updated. The system is also scalable and allows adding information of any disease in the document.

7. Future Work

To further enhance the capabilities of the proposed QA system in medical domain we can do classification of the questions with machine learning techniques. Machine learning techniques could find the focus of question and do classification of the questions. The performance could also be enhanced by the use of web crawler which automatically extract information from WebPages and add it to document to keep it updated. Cosine similarity could be performed to search the particular document from thousands of documents available on the net. The performance could be also enriched with the help of query expansion, query reformulation and answer validation. We can give input in form of speech by using speech to text converter and vice versa for generating results.

References

[1] Min-kyoung Kim, and Han-joon Kim, "Design of Question Answering System with Automated Question Generation", Fourth International Conference on Networked Computing and Advanced Information Management, Washington USA, 2008 , pp. 365-368.

[2] Grant Ingersoll, Ozgur Yilmazel and Elizabeth D.Liddy, “Finding Questions to Your Answers”, ICDE Workshops, Turkey, 2007, pp. 755-759.

[3] Luiz Augusto Pizzato and Diego Molla, “Question Prediction Language Model”, Proceedings of the Australasian Language Technology Workshop, Manchester UK, 2007, pp. 74-81.

[4] Farah BENAMARA and Patrick SAINT-DIZIER: “Lexicalizations Strategies in Cooperative Question-Answering Systems”, In Proceedings of the 20th international conference on Computational Linguistics, Geneva, Switzerland, 2004, pp. 9-16.

[5] Green, B.F., Chomsky, C., & Laughery, K. (1961). “BASEBALL: An automatic question answerer.” Proceedings of the Western Joint Computer Conference, New York: Institute of Radio Engineers, 9-11 May 1961, pp. 219-224.

[6] Andras Kornai, “Evaluating Geographic Information retrieval”, Available at http://www.kornai.com

[7] Farah Benamara and P. Saint Dizier, “Advanced relaxation for cooperative question answering: New Directions in Question Answering”, In Mark T. May bury, AAAI/MIT Press, Geneva, Switzerland, 2004.

[8] Ferres, D., Kanaan, S., Gonzalez, E., Ageno, A., Rodriguez, H., Surdeanu, M. and Turmo, J., “Talp QA system at TREC-2004: Structural and hierarchical relaxation over semantic constraints”, Gaithersburg, MD, Proceedings of the Text Retrieval Conference ,TREC 2004.

[9] Xu, J., A. Licuanan, J. May, S. Miller, and R.Weischedel, “TREC 2002 QA at BBN: Answer selection and confidence estimation”, In Proceedings of the Eleventh Text Retrieval Conference, Gaithersburg, Maryland, TREC 2002, pp. 500-521.

[10] Question Answering system definition: available at http://en.wikipedia.org/wiki/Question_answering.

[11] Online Word Net definition: available at http://www.wordnetonline.com/events.html.

[12] The Stanford Natural Language Processing Group: URL http://nlp.stanford.edu/software/lex-parser.shtml.

[13] Named Entity Recognition Wikipedia: available at http://en.wikipedia.org/wiki/Named_entity_recognition.

[14] Wikipedia : URL http://www.wikipedia.org/

[15] Kanada, Y. “A method of medical name extraction from Japanese text for thematic medical search”, in Proceedings of the 8th International Conference on Information and Knowledge Management, Kansas City, Missouri, 1999, pp. 46-54.

[16] Hermjakob, “Parsing and question classification for question answering”. In Proceedings of the Association for Computational Linguistics 39th Annual Meeting and 10th Conference of the European Chapter Workshop on Open- Domain Question Answering, Toulouse, France, 2001, pp. 17-22.

[17] Strasser, T. C. 1978. The information needs of practicing physicians in northeastern new york state. Bulletin of the Medical Library Association 66:200–209

[18] Smith, R. 1996. What clinical information do doctors need? BMJ 313:1062–1068.

[19] Wilkinson, M. A. 2001. Information sources used by lawyers in problem-solving: An empirical exploration. Library and Information Science Research 23(3):257

[20] Giuse, N. B.; Huber, J. T.; Giuse, D. A.; Brown Jr., C. W.; Bankowitz, R. A.;and Hunt, S. 1994. Information needs of health care professionals in an aids outpatient clinic as determined by chart review. The Journal of the American Medical Informatics Association 1(5):395–403.