IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

XMQAS - An Ontology Based Medical Question Answering System

Midhunlal M[1], Gopika Mangalassery[2]

Computer and Information Science, College of Engineering Cherthala, Cherthala, India[1,2]

Abstract: Designing question answering systems requires efficient and deep analysis of natural language questions. A key process for this task is to translate the semantic relations expressed in the question into a machine-readable representation. This work tackles question analysis in the medical field using an approach based on Ontology. The main computing methods of the question answering system are based on the application of natural language processing technique to infer the focus of the question and process the question based on the meaning inferred from the question. The efficiency of the outcome is mainly depended on extraction of the correct characteristic of the question and the accuracy of the medical documents are using**.**

Keywords: Medical Question Analysis, Information Extraction, Document Retrieval, Natural Language Processing.

I. INTRODUCTION

Question Answering, unlike traditional Information Retrieval, aims to provide inquirers with direct, precise answers to their questions, by employing Information Extraction and Natural Language Processing techniques. The search engines like google provides a large number of documents that are potentially relevant for the questions posed by the inquirers instead of providing accurate answer. The contribution of QA systems is to provide a fast access to the searched knowledge, which is a crucial point in the medical domain for both practitioners and patients. Their performance is often evaluated by measuring their precision and recall over the retrieved answers. Due to the continuous, exponential growth of information produced in the biomedical domain, and due to the crucial impact of such information upon research and upon real world applications, there is a particularly great and growing demand for QA systems that can effectively and efficiently aid biomedical researchers and health care professionals in their information search. QA systems equipped with reasoning capabilities can derive more adequate answers by using inference mechanisms. The question answering task has two reference inputs: the corpora to be used to extract the relevant answers and the question itself. Question answering systems must have the potential to automatically mine relevant knowledge from multiple sources and summarize the results to form answers based on important concepts embedded in the question. And to improve efficiency of retrieval, they can provide compact answers rather than entire documents, which can help users, pinpoint useful information quickly.

II. RELATED WORKS

Question answering can be considered an advanced form of information retrieval. Most research development in the area of QA are, the work proposed by Rafael M Terol et al. [1] in restricted medical domain. The advantages of this QA system include the simple process of defining the question taxonomy answered by the system as well as the possibility of locally or remotely managed document collections. And the system uses wordnet and UMLS uses for infer the knowledge. Yong Gang Cao et al.[2] proposed a online based clinical question answering system named Ask HERMES to perform robust semantic analysis on complex clinical questions. The novel clustering based summarization and presentation of Ask HERMES offers a clear advantage over the long document lists retrieved from Google, with the potential to save busy physicians time in retrieving potentially irrelevant documents. Asma Ben Abacha et al.[3] proposed an approach for translating natural language questions into machine readable representation. This method mainly dealing with translation of YES/NO questions and WH questions. Approach including medical entity recognition, semantic relation extraction and automatic translation to SPARQL queries.

To extract semantic relations between medical entities Asma Ben Abacha et al. [4] present a method with an empirical study on the treatment relation. This allows (i) to extract and annotate medical entities and relationships from medical texts and (ii) to explore semantically the produced RDF annotations. In the healthcare domain, to automatically separate consumer questions from professional questions Feifan Liu et al.[5] developed supervised machine-learning model. The system uses BOW features, Statistical features, and Linguistic category features as learning features.

III. METHODOLOGY

For a medical question answering system, the ultimate focus on the natural language question inputted to the system. Aim is to extract the medical entities and correct relation from the question given, using that information the system retrieve relevant documents and the accurate answer from the documents. The current question answering systems are not much efficient and provide processing overhead. Here proposes a better method for

DOI 10.17148/IJARCCE.2016.54228

medical question answering system which identifies best matching for the question.

The system uses two ontology's, they are UMLS and Word net. The architecture is given below.

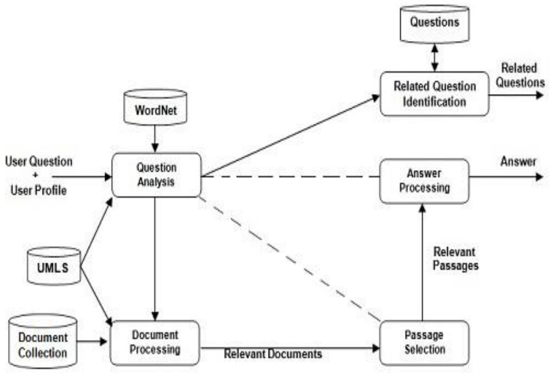


Fig1. Architecture of proposed system

A. Pre-processing
In pre-processing phase there are two steps are carried out. They are

1. Term-Document Matrix Creation: Term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. The TDM is created at the beginning of the process. In the case of a medical question answering system, the keywords should be the medical terms. Creation of a TDM is a time consuming task, but its only need to carry out at the first stage of the system. Once the TDM is created, can use that until there is no change in the documents. Searching using TDM is much faster than the serial search in the documents for the keywords. In TDM row represents the documents and column represents the keywords.
2. Pattern Generation: For medical question answering system, targeting on 7 types of semantic relations chosen according to an analysis of medical question taxonomies[11], they are

1. Treats: Treatment improves or cures medical problem
2. Complicates: Treatment worsens medical problem
3. Prevents: Treatment prevents medical problem
4. Causes: Treatment causes medical problem
5. Diagnoses: Test detects, diagnoses or evaluates medical Problem
6. DhD: Drug has dose
7. PhSS: Problem has signs or symptoms

According to this question taxonomy, the 10 most important question formulated by doctors

1) What is the drug of choice for condition x?
2) What is the cause of symptom x?
3) What test is indicated in situation x?
4) What is the dose of drug x?
5) How should I treat condition x (not limited to drug
6) treatment)?
7) How should I manage condition x (not specifying diagnostic or therapeutic)?
8) What is the cause of physical finding x?
9) What is the cause of test finding x?
10) Can drug x cause (adverse) finding y?
10. Could this patient have condition x?

This pattern generation task consists of the definition of the patterns that identify each generic question. These patterns are composed by the combination of types of medical entities and other elements in the question like verbs. These patterns can be generated according to the automatic generation of the patterns through the processing of questions according to the question taxonomy. Fig.2 shows the main processing carried out in the pattern generation task.

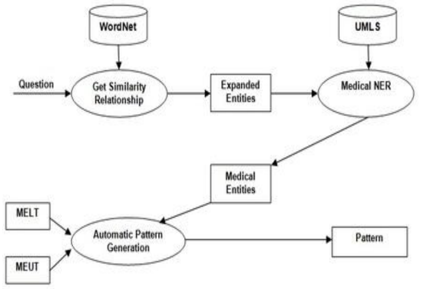


Fig 2. Architecture of pattern Generation phase

The first step in pattern generation phase is the identification of similarity relation of words in the question with other entities(or words) using Word Net[?]. Next step is the recognition of medical named entities whose type is noun (NN) or complex nominal (NNC) including their possible adjective modifiers (JJ). The third step consists of the automatic setting of the MELT (Medical Entities Lower Threshold) whose score is set to the number of medical entities in the logic form minus one, and the automatic setting of the MEUT (Medical Entities Upper Threshold) of which the score is set to the number of medical entities in the logic form. Finally, assigns a possible expected answer type for each pattern.

B. Question Analysis Phase
The question answering task has two reference inputs: the corpora to be used to extract the relevant answers and the question itself. Each of these inputs must be analyzed in a manner that makes the question-answer matching semantically relevant, easy to understand and potentially traceable. The natural language questions formulated to the system are processed initially by the question analysis component. The main steps question analysis phase are given in the fig.3.

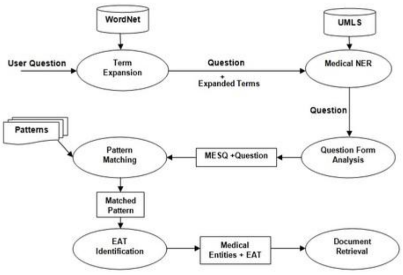


Fig 3. Steps in QA phase

DOI 10.17148/IJARCCE.2016.54228

Once the user enters a question into the system, the system identifies the similarity relationship of words in the question with every other word using wordnet.

In next step, system extract the entities from the question whose type is noun (NN) or complex nominal (NNC) including their possible adjective modifiers (JJ).

And check the extracted words are medical terms or not using UMLS. Once the medical terms are identified, next step is to set the MESQ (medical entities score in question) for the question. MESQ can be defined as the number of medical entities in the logic form of the question.

The next step consists of finding those patterns matched with the user questions and MELT _MESQ _ MEUT.

In this step, sets a entities matching measure (EMM) for every user question and pattern pair. EMM is the number of medical entities that match between the question and the pattern, select the pattern whose difference between EMM and MELT is the lowest one.

Every pattern is associated with a semantic relation, using this identifies the expected answer type(EAT) of the user question. Finally, the identified medical entities and EAT is given to document retrieval phase.

C. Document Retrieval Phase
Once the keywords are received from question analysis phase, keywords are searched in TDM. For every keyword used for document retrieval, need to calculate TFIDF value. The TFIDF value of a term is calculated using following equation

TFIDF = TF*IDF
IDF = log (N/df)

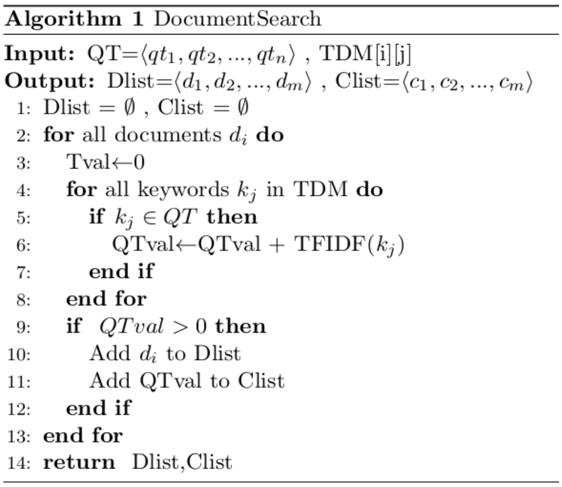
where,
TF - Term Frequency
IDF - Inverse Document Frequency
N - Total no: of documents in the corpus
df - Document frequency

The algorithm used for document retrieval is given in Algorithm 1. Document Search returns a list of documents and its tfidf value. After analysing the list, the system identifies five documents whose tfidf value is greater than all other documents.

These documents are selected for next level of processing.

D. Passage retrieval Phase
Best matching sentences for the given question is identified in this phase. First step performs the identification of correct passage from the retrieved documents.

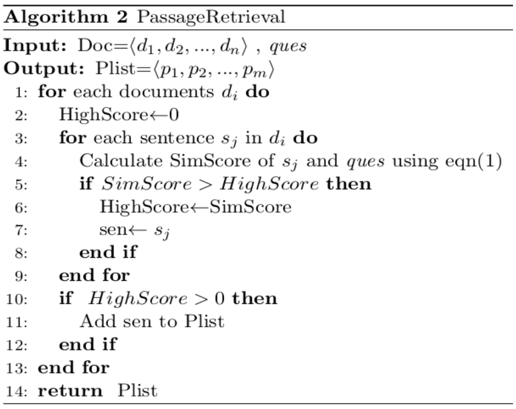
This is done by the pattern matching technique in the documents. The algorithm for passage retrieval phase is given Algorithm 2. The input will be a set of documents and the user question. The output will be the passage identified from each documents. The equation used for pattern matching technique is,



```
Algorithm 1 DocumentSearch
Input: QT=$\langle qt_1, qt_2, ..., qt_n \rangle$ , TDM[i][j]
Output: Dlist=$\langle d_1, d_2, ..., d_m \rangle$ , Clist=$\langle c_1, c_2, ..., c_m \rangle$
Dlist = $\emptyset$ , Clist = $\emptyset$
for all documents $d_i$ do
    Tval←0
    for all keywords $k_j$ in TDM do
        if $k_j \in QT$ then
            QTval←QTval + TFIDF($k_j$)
        end if
    end for
    if $QTval > 0$ then
        Add $d_i$ to Dlist
        Add QTval to Clist
    end if
end for
return Dlist,Clist
```

$$\text{Similarity Score } S_s = \frac{|W_i \cap W_j|}{|log(W_i)|+|log(W_j)|} \text{———(1)}$$

where,
Wi - length of first sentence
Wj - length of second sentence



```
Algorithm 2 PassageRetrieval
Input: Doc=$\langle d_1, d_2, ..., d_n \rangle$ , $ques$
Output: Plist=$\langle p_1, p_2, ..., p_m \rangle$
for each documents $d_i$ do
    HighScore←0
    for each sentence $s_j$ in $d_i$ do
        Calculate SimScore of $s_j$ and $ques$ using eqn(1)
        if $SimScore > HighScore$ then
            HighScore←SimScore
            sen← $s_j$
        end if
    end for
    if $HighScore > 0$ then
        Add sen to Plist
    end if
end for
return Plist
```

E. Answer Extraction Phase
The answer extraction phase analyze the passages extracted in the previous phase to check whether the passage is accurate for the question asked by the user. For each passage extracted in previous phase, system identifies the medical entities and its semantic relation. And identifies the sentences with same semantic relation of question. For these sentences, assign a score depending upon the number of search terms found in each sentence. Finally the passage with maximum score is given as expected answer for the user question.

IV. EXPERIMENTAL RESULTS

A. Performance Analysis of Document Retrieval Phase

The precision and recall values of document retrieval phase is shown in the fig4. The equation used for calculating precision and recall are given below.

$$\text{Precision} = \frac{\text{Accurately retrieved documents}}{\text{Total documents retrieved}}$$

$$\text{Recall} = \frac{\text{Accurately retrieved documents}}{\text{Total relevant documents}}$$

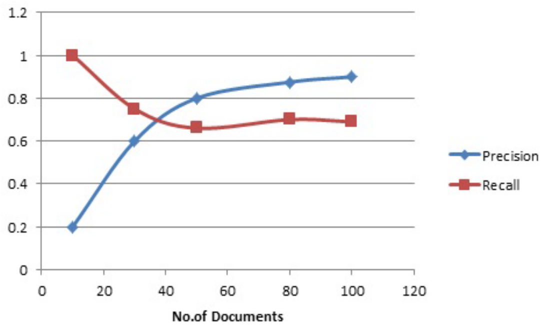


Fig 4: Evaluation

B. Overall System Performance Evaluation

$$\text{Precision} = \frac{\text{Accurately answered questions}}{\text{Total input questions}}$$

$$\text{Recall} = \frac{\text{Accurately answered questions}}{\text{Accurately processes questions}}$$

Questions given	10	20	30	40	50
Accurately processed questions	10	19	28	37	45
Accurately answered questions	8	17	25	34	42
Precision	.8	.85	.83	.85	.84
Recall	.88	.89	.89	.91	.93

V. CONCLUSION

This work presents an approach to question-answering system for medical inquiry. This approach is based on three main steps: (i) analysis of the natural language question (ii) retrieve the relevant documents based on the question and (iii) retrieve the correct passage from documents. The effectiveness of the system is based on the precision of the answer retrieved for a particular question. The performance of the system is improved because of using UMLS and wordnet. And this system is useful for doctors as well as common users.

REFERENCES

[1]. Rafael M Terol, Patricio Martinez Barco, Manuel Palomar, A knowledge based method for the medical question answering problem, Computers in Biology and Medicine 37 (2007) 1511 1521, Elsevier january 2007.

[2]. YongGang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, AskHERMES: An online question answering system for complex clinical questions,Journal of Biomedical Informatics 44 (2011) 277288, Elsevier January 2011.

[3]. Asma Ben Abacha, Pierre Zweigenbaum, Medical Question Answering: Translating Medical Questions into SPARQL Queries, IHI12, January 2830, 2012, Miami, Florida, USA, ACM January 2012.

[4]. Asma Ben Abacha, Pierre Zweigenbaum, Automatic extraction of semantic relations between medical entities: a rule based approach, Fourth International Symposium on Semantic Mining in Biomedicine (SMBM) Hinxton, UK. 25-26 October 2010, Journal of Biomedical Semantics 2011.

[5]. Feifan Liu a, Lamont D. Antieau a, Hong Yu, Toward automated consumer question answering: Automatically separating consumer questions from professional questions in the healthcare domain, Journal of Biomedical Informatics 44 (2011) 10321038, Elsevier August 2011.

[6]. Xiaohua Zhou, Hyoil Han, Isaac Chankai, Ann A. Prestrud and Ari Brooks, Converting Semi-structured Clinical Medical Records into Information and Knowledge, 21st International Conference on Data Engineering, pp.1162, April 2005

[7]. Xiaohua Zhou and Hyoil Han, Isaac Chankai, Ann Prestrud and Ari Brooks, Approaches to Text Mining for Clinical Medical Records, SAC06, 23-27, April -2006, Dijon, France, Copyright 2006 ACM

[8]. George A Miller, WordNet: A Lexical Database for English

[9]. UMLS Reference Manual Bethesda (MD): National Library of Medicine,(US)-2014AB release http://www. nlm.nih. gov/ research/umls/

[10]. Ely, J.W, Osheroff, J.A, Gorman, P.N, Ebell, M.H, Chambliss, M.L, Pifer, A taxonomy of generic clinical questions: classification study, BMJ, 321 (2000)

[11]. Sofia J. Athenikos a, Hyoil Han, Biomedical question answering: Asurvey, computer methods and programs in biomedicine 99 1-4,Elsevier October 2009.

[12]. Python Programming Language - http://www.python.org/ - (Python 2.7 : July 3rd, 2010 release).

BIOGRAPHIES

Midhunlal M received the M-tech post graduation in Computer and information science from College of engineering Cherthala in the duration 2013-2015. During the year 2014-15 he collected and conducted analysis on medical question and answering.

Gopika Mangalassery received the M-tech post graduation in Computer and information science from College of engineering Cherthala in the duration 2013-2015. During the year 2014-15 she did literature survey. After that information extraction was done.