Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-215

CliniQA : Highly Reliable Clinical Question Answering System

Yuan NI[a1], Huijia ZHU[a], Peng CAI[a], Lei ZHANG[a], Zhaoming QUI[a], Feng CAO[a]
[a]*IBM China Research Lab*

Abstract. Evidence-based medicine (EBM) aims to apply the best available evidences gained from scientific method to clinical decision making. From the computer science point of view, the current bottleneck of applying EBM by a decision maker (either a patient or a physician) is the time-consuming manual retrieval, appraisal, and interpretation of scientific evidences from large volume of and rapidly increasing medical research reports. Patients do not have the expertise to do it. For physicians, study has shown that they usually have insufficient time to conduct the task. CliniQA tries to shift the burden of time and expertise from the decision maker to the computer system. Given a single clinical foreground question, the CliniQA will return a highly reliable answer based on existing medical research reports. Besides this, the CliniQA will also return the analyzed information from the research report to help users appraise the medical evidences more efficiently

Keywords. Evidence based Medicine (EBM), Question & answering system (QA), Natural language processing (NLP)

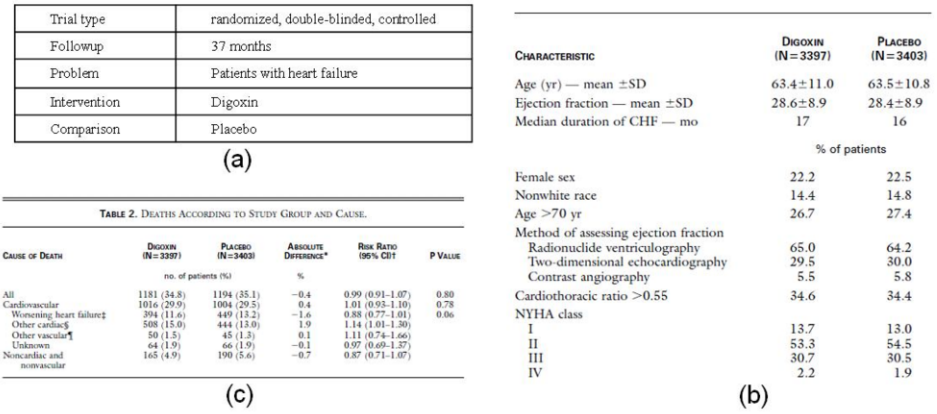
Introduction

Different users may encounter different clinical questions, and they require clinical evidences to answer their questions and to help decision making. For physician, it is likely that they have some clinical questions and want to find the clinical evidences to help them diagnose or treat the patients. For patients, as they are becoming smarter, they take more responsibility on their own health, thus they may ask clinical questions to the clinical evidence that could help them take part in the decision making for their own health. For medical professionals and policymaker, they may issue clinical questions to find clinical evidence to assess the healthcare technology. However, due to the large volume of rapid increasing medical information, it is very time-consuming for users to manually retrieve the clinical evidences, appraise the evidences and interpret them. CliniQA is proposed to automatically provide highly reliable answers to users' clinical questions. Additionally, besides the answer extracted from the clinical evidence, CliniQA will also conduct deep analysis on the clinical evidence and provide additional information to users to help the clinical evidence appraisal.

Let us see an example as follows. Suppose a physician has a clinical question saying "Is there any evidence for the use of Digoxin to reduce the mortality in patients with heart failure?", the CliniQA would generate an answer saying "an article from The New England Journal of Medicine concluded that 'Digoxin did not reduce overall

[1] Corresponding Author: IBM Research - China. Email : niyuan@cn.ibm.com

mortality, but it reduced the rate of hospitalization both overall and for worsening heart failure'.". Besides this, the CliniQA will give the extracted information, which is called *evidence object*, as shown in Figure 1. The evidence object contains two kinds of information: (1) the type of trial, the followup, the PICO (Problem, Intervention, Comparison, Outcome) elements [1] as in Figure 1 (a); (2) tables about the population and the trial results as in Figure 1 (b) & (c)



Trial type	randomized, double-blinded, controlled
Followup	37 months
Problem	Patients with heart failure
Intervention	Digoxin
Comparison	Placebo

(a)

TABLE 2. DEATHS ACCORDING TO STUDY GROUP AND CAUSE.

CAUSE OF DEATH	DIGOXIN (N=3397)	PLACEBO (N=3403)	ABSOLUTE DIFFERENCE*	RISK RATIO (95% CI)†	P VALUE
	no. of patients (%)		%		
All	1181 (34.8)	1194 (35.1)	−0.4	0.99 (0.91–1.07)	0.80
Cardiovascular	1016 (29.9)	1004 (29.5)	0.4	1.01 (0.93–1.10)	0.78
Worsening heart failure‡	394 (11.6)	449 (13.2)	−1.6	0.88 (0.77–1.01)	0.06
Other cardiac§	508 (15.0)	444 (13.0)	1.9	1.14 (1.01–1.30)	
Other vascular¶	50 (1.5)	45 (1.3)	0.1	1.11 (0.74–1.66)	
Unknown	64 (1.9)	66 (1.9)	−0.1	0.97 (0.69–1.37)	
Noncardiac and nonvascular	165 (4.9)	190 (5.6)	−0.7	0.87 (0.71–1.07)	

(c)

CHARACTERISTIC	DIGOXIN (N=3397)	PLACEBO (N=3403)
Age (yr) — mean ±SD	63.4±11.0	63.5±10.8
Ejection fraction — mean ±SD	28.6±8.9	28.4±8.9
Median duration of CHF — mo	17	16
	% of patients	
Female sex	22.2	22.5
Nonwhite race	14.4	14.8
Age >70 yr	26.7	27.4
Method of assessing ejection fraction		
Radionuclide ventriculography	65.0	64.2
Two-dimensional echocardiography	29.5	30.0
Contrast angiography	5.5	5.8
Cardiothoracic ratio >0.55	34.6	34.4
NYHA class		
I	13.7	13.0
II	53.3	54.5
III	30.7	30.5
IV	2.2	1.9

(b)

Figure 1: Content of the Evidence Object

Dina and Lin [2] have proposed a knowledge based and statistical approach for answering clinical questions. However, their approach only provides the answer text while the evidence object to facilitate the critical appraisal is not considered. The CliniQA prototype has made use of IBM's DeepQA framework, which is used to develop the supercomputer Watson who has beaten the human champions in the Jeopardy! game. In the prototype, we have focused on the therapy question and the clinical trials as the evidences. While our approach could be extended to other types of question such as diagnose and other types of clinical evidences.

1. System Architecture

Figure 2 illustrates the overall architecture of the CliniQA prototype. The system consists of six main components, which could be divided into offline part and online part. This section describes the functions of each component and the technique details will be elaborated in Section 3. The offline part includes the evidence analyzer, which is used to preprocess the clinical evidences including the structure and semantic annotations, indexing, and stores the results into the **Evidence base**.

For the online components, firstly, given a clinical question in natural language, the **question analyzer** performs semantic annotations on the question and identifies the PICO elements from the question. It also determines whether a predefined question template could match this question, and generates the corresponding search query for the evidence retrieval component. Then, the **evidence retrieval** conducts a search on the evidence base. The top N matched clinical evidences will be considered as the candidate answers. The next step is a set of **candidate scorers**, which are used to

measure the probability on how the candidate answer satisfies the clinical question. For example, one dimension could be whether the patients' information mentioned in the question satisfied in the clinical trial of the evidence. Then the **answer generator** will integrate the various scores for one candidate into the final score and identify the set of evidences that contain the answers to the question. The machine learning technique is used here to determine the final answer, and then the answer paragraphs are extracted. Finally, the **result displayer** retrieves the evidence object of the answer article from the evidence base and displays the results as shown in Figure 1.

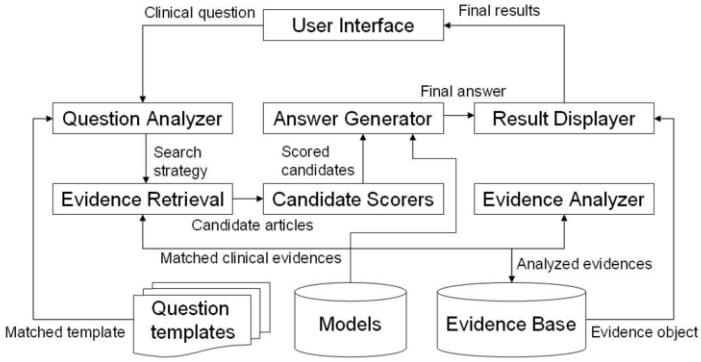


Figure 2. CliniQA Architecture

2. Methods

In the CliniQA prototype, the natural language processing technique is used to perform the question and evidence analysis and the machine learning technique is used to merge different scores and determine the final answers. The following of this section discusses these techniques in details.

Question analyzer. We have designed a set of question templates to cover most therapy questions. If some question matches one of the templates, the specific search strategy will be used to achieve higher retrieval precision. In terms of the PICO framework, the clinical questions always contain four kinds of elements: Problem & population (P), Intervention (I), Compared intervention (C) and Outcome (O). The therapy question is about the treatment relationship between I and P. These questions could be represented by the following three templates: (1) <I?, P>. This template asks for the intervention for a given problem, e.g. "how should I treat polymenorrhea in a 14-year-old girl?" (2)?<I, P>. This kind of questions asks for the evidences or the effectiveness of using the Intervention on the Problem. Sometime, the compared intervention is also indicated. E.g. "what is the evidence for using Metformin in people with type 1 diabetes who are obese and poorly controlled?" (3) <I?>. This kind of questions asks for the usage of some intervention, e.g. "Is melatonin good for anything? I do not know anything about melatonin. I need to know the dose".

Given a clinical question in natural language, firstly we will identify the PICO elements. We use the medical concept annotator MetaMap [3] to identify the medical concepts from the question. Then in terms of the semantic type of each concept, we classify them into PICO using the mapping in [4]. Note that there exist the common semantic types between P and I/C, e.g. [Treatment & Drug]. We make use of the NLP

parser results to distinguish them. If the phrase has a dependency relationship on a noun for people such as patients, child, etc., it is classified as P; otherwise, it is classified as I/C. Secondly, we use the SemRap [5] tool to detect whether a TREAT relationship exists. If it exists, it is a therapy question. Thirdly, the analyzed question will be matched with the templates. If a template is matched, the corresponding search query will be generated which indicates the keywords on different fields; otherwise, the whole question will be used as the keywords for searching.

Clinical evidence preprocessing. The clinical evidence preprocessing includes two parts, i.e. semantic analysis to generate evidence object and the indexing for searching. The semantic analysis tries to extract the following information from the evidence document: (1) PICO elements; (2) randomization and followup for the clinical trial; (3) useful tables such as the table to describe patients' information. Considering the well structured clinical trial articles, we make use of the linguistic and structure characteristics of the text. For each kind of information, a set of rules are created for candidate retrieval. Different weights are assigned to different rules. The candidate with the highest score will be the target information. Due to the space limitation, we use the extraction of Intervention as an example to illustrate our method. Three rules are designed for Intervention: (1) if the title satisfies the pattern "[XXX] of [YYY] on/for/in [ZZZ]", then the [YYY] part contains the intervention; (2) if there exists a sentence in the objective part of the abstract section, and the sentence satisfies the pattern "the evaluation/effectiveness/impact/effect(s) of [YYY]", then the [YYY] part contains the intervention; (3) the most frequently appeared medical concept of semantic type [TREATMENT&DRUG] is also considered as the candidates. The candidate generated by different rules is given different weights. The same candidate will be merged and the final ranked candidate list will be generated. The top candidate is considered as the Intervention for the article.

Given the analyzed clinical evidences, we use the Lucene indexer to build the indexes for evidence retrieval. Each evidence article corresponds to one document, which has four fields: (1) P field is to index the problem and population information; (2) I/C field is to index the intervention and comparison information; (3) O field is to index the outcome information; (4) article field is to index the text for the whole article. Given the search query from the question analyzer, if the question has the requirements on the specific field, the retrieval will be conducted at the corresponding field.

Candidate scorers. The candidate scorers are used to compute a score to measure the probability on how likely the candidate article contains the answer. Considering different characteristics of different questions, we could design different scorers to address the probability at different dimensions. We have a set of scorers to measure the semantic similarity between the question and the evidence article, for example the PersonAgeScorer and PersonGenderScorer are to measure whether the patients' age/gender mentioned in the question are similar with the population in the article. We also have the scorers to measure the quality of the article, for example the double-blinded, random controlled trial has higher score than a non double-blinded trial. In the CliniQA prototype, we have developed 15 scorers already. We plan to design more scorers to achieve a better accuracy.

Answer generator. Given a list of candidates and each candidate has a vector of scores, we apply a model on them to integrate different scores for one candidate into one confidence score for the candidate. The machine learning technique is used to train the model by using some existing question and answer pairs. Then we sort the candidates according to their confidence scores. A threshold is set such that all

candidates with confidence scores higher than the threshold would be considered as containing the answers. Then, we extract the answer paragraphs from the whole article. Considering the therapy questions, the conclusion section in the abstract part is extracted as the answer paragraphs for users. Finally, for each answer, the result displayer gets the corresponding evidence object from the evidence base, and shows the results to users.

3. Results

To evaluate the performance of CliniQA system, we need to find a set of clinical questions with existing answers to build our training dataset and test dataset. We use the Trip Answers website [6]. Currently, there are 6382 questions in the repository. We use all answers for 6382 questions to create the evidence base, and each answer is treated as one evidence article. We have randomly selected 1000 questions as our training datasets and 500 questions as our test datasets. The logistic regression is used to train the model. If the answer passage generated by the CliniQA is the same as the answer passage in Trip answers, the question is marked as correctly answered. The metric we use is the precision at top-k recall, which is widely used in information retrieval. The experimental results show that the CliniQA achieves the 71% precision at top-5 recall and 46% precision at top-1 recall on the test dataset.

4. Discussion

In this paper, we have introduced the CliniQA, which is an automatic clinical question answering system. The CliniQA has the following features: (1) the natural language processing techniques are widely used for question and article analysis; (2) many scorers are integrated together by machine learning techniques to determine the final answer; (3) besides the answer paragraph, the additional information, i.e. evidence object, is shown to users to help the evidence appraisal. As future work, firstly we plan to handle other types of questions and integrate other types of clinical evidences; secondly, we plan to develop more scorers to achieve better precision; thirdly, we will try to invite domain experts to conduct an extensive evaluation.

References

[1] Richardson WS, Wilson MC, Nishikawa J, and Hayward RS. The Well-Built Clinical Question: A Key to Evidence-based Decisions. American College of Physicians Journal Club, 123(3):A12-A13
[2] Dina Demner-Fushman and Jimmy Lin. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. Association for Computational Linguistics, 33(1), 2007.
[3] Aronson, Alan R. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. AMIA Annu Symp Proc. 2001, 17-21
[4] Xiaoli Huang, Jimmy Lin and Dina Demner-Fushman. Evaluation of PICO as s Knowledge Representation for Clinical Question. AMIA Annu Symp Proc, 2006, 359-363
[5] Rindflesch, Thomas C. and Marcelo Fiszman. The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text. J Biomed Inform, 36(6):462-477
[6] www.tripanswers.org