Question Answering in Biomedicine – An evaluation of third party search engines

Student :Andreea Tutos
Id 41064739
Supervisor: Diego Molla

Abstract

The internet has become a source of wealth of medical information which is made available to practitioners mainly through search engines. Every day patient treatment decisions are impacted by the ability to locate the information available on different web sites. This paper describes the evaluation of the answerability of a set of clinical questions posed by physicians. The clinical questions have been identified as belonging to two categories of the five leaf high level hierarchical Evidence Taxonomy created by Ely and his colleagues: Intervention and Non Intervention. This study employs the services of five search engines (PubMed, Google, MedQA, Answers.com and OneLook) for locating the answers to the question corpus. The evaluation criteria include quality of answers, ranked according to the link's position in the response list and also relevance to the question. The results show that Non Intervention questions seem to be easier to answer than Intervention questions.

1 Introduction

Medical professionals face today an overwhelming amount of information they need to be aware of in order to provide high quality medical care. Almost every aspect of patient care relies on searches of published medical articles and evidence from previous work of other physicians. Latest clinical guidelines urge physicians to practice Evidence Based Medicine when providing care for their patients (Yu et al, 2005). Evidence Based Medicine implies referring to the best evidence from scientific and medical research that can assist in making decisions about patient care. Lowering the barriers to the use of evidence based knowledge has the potential of improving the quality of patient consultation at the point of care.

A US study shows that 49% of family physicians errors are due to a lack of knowledge about medical aspects of the case (Ely et al, 1995). Another study revealed that medical practitioners have two questions for every three patients seen and 40% of their clinical questions were on medical facts (Covell et al, 1985). The research goes even further discovering that family physicians do not pursue answers for 64% of their clinical questions. Out of those questions, 80% have answers that could be located (Ely et al, 1995).

Our project aims to determine the answerability of a set of 40 medical questions sourced from the Parkhurst Exchange[1] website. This is one of the first steps towards creating a medical question answering system. Question answering is an advanced form of information retrieval in which answers are generated in response to ad-hoc questions. Recognizing the importance of Evidence Based Medicine, our project goes further than other referenced studies by not limiting the corpus of questions to definitional questions. Definitional questions have been described as having the format of "What is a/an X?".

The questions and answers provided by Parkhurst Exchange were used as a benchmark in measuring the relevance of answers located through the five selected search engines: PubMed[2], Google, MedQA[3], Answers.com[4] and OneLook[5].

The structure of this document is as follows: Section 2 describes studies and concepts related to our project, Section 3 introduces the evalua-

[1] See http://www.parkhurstexchange.com
[2] See http://www.ncbi.nlm.nih.gov/pubmed/
[3] See http://monkey.ims.uwm.edu:8080/MedQA/query_qa.cgi
[4] See http://www.answers.com
[5] See http://www.onelook.com

tion methodology employed in the study. The methodology details on the corpus of questions and how it has been selected and on the classification of candidate questions according to the evidence node in the evidence taxonomy. It also describes the selected search engines and the reasons behind their selection. Question processing description follows together with answer extraction section. The discussion of results presents the findings of our research, followed by the conclusions section that analyzes the results.

2 Background

The corpus of questions of our study has been constructed from the questions and answers list available on the Parkhurst Exchange website. Parkhurst Exchange is a highly regarded medical publishing website based in Canada that includes a collection of over 4800 clinical questions and their answers provided by physicians. Since 1983 when it first started, it continues to develop strong relationships with top physicians across many medical disciplines.

Our project refers to the Evidence taxonomy created by Ely and his colleagues (Yu and Sable, 2005). This high level, five leaf hierarchy categorizes medical questions that are potentially answerable with evidence. The hierarchy is presented in Figure 1 below.

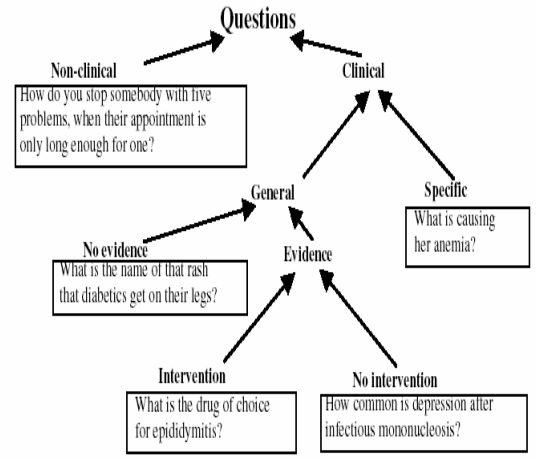

Figure 1: "Evidence Taxonomy" created by Ely and his colleagues, with examples.

Ely and his colleagues have concluded that the Non-clinical, Specific and No Evidence questions are not answerable with evidence, while both categories of Evidence (Intervention and No Intervention) are potentially answerable. Non-clinical questions do not address the specific medical domain and Specific questions require information from patient's personal record.

We have focused on the two evidence categories confirmed by Ely's study as being answerable with evidence (Yu and Sable, 2005): Intervention and Non-Intervention questions. According to the Evidence taxonomy, intervention questions are scenario-based, quite complex and they require complex answers that provide descriptions of possible treatments or recommended drugs. Non-intervention questions usually enquire about medical conditions or drugs, without asking for directions in managing a disease. They generally belong to the family of factoid questions for which short answers are usually expected.

Our work is related to the study of Hong and Kaufman (Yu and Kaufman, 2007) who conducted a cognitive evaluation of four online engines on answering definitional questions. Our study does not limit the input questions to definitional questions only, exploring further by including questions belonging to the Evidence node in the Evidence Taxonomy. Yu and Kaufman's evaluation criteria included quality of answers, ease of use, time spent and number of actions taken to locate an answer. Their results showed that PubMed performed poorly, Google was the preferred system for quality of answer and ease of use and MedQA surpassed Google in time spent and number of actions. The findings of this study contradict the work of Berkowitz (Berkowitz, 2002) who concluded in his research that Google performed poorly in regards to quality of answers as it referred to consumer-oriented sites.

PubMed is a search engine that accesses a reputable medical repository (MEDLINE) maintained by the US National Library of Medicine (Demner-Fushman and Lin, 2007). The MEDLINE database includes over 15 million medical articles and has proven to be a well recognized knowledge source across medical question answering studies.

Google is a widely used web search engine that uses text matching techniques to locate web pages relevant to a user's search. Google's architecture includes a list of features that make it a high-precision search engine. First to be mentioned is the ability to determine quality rankings or PageRanks for each web page based on the link structure of the Web. Another important characteristic of Google is that it establishes a relation between the text of links and the pages the links point to (Brin and Page, 1998).

MedQA (Yu et al, 2007) is one of the first developed medical answering systems that respond to definitional questions accessing the MEDLINE records and other World Wide Web collections. It automatically analyzes a large number of electronic documents in order to generate short and coherent answers in response to the input questions. The reason behind deciding on definitional questions is that they are 'more clear-cut' as opposed to other types of clinical questions that can have large variations in their expected answers.

Answers.com is a website that can offer useful answers to categories of questions like business, health, travel, technology, science, entertainment, arts, etc. Their collection includes over four million answers drawn from over 180 titles from brand-name publishers, together with content created by their own editorial team.

OneLook is a dictionary and translation metasearch engine that accesses more than 900 online dictionaries in order to locate the desired definition. It offers the ability to decide on the dictionary to focus on, with choices of domains as medical, art, business, etc.

3 Evaluation Methodology

3.1 Questions Corpus

Constructing the corpus of questions has proved to be a relatively complex process mainly due to the lack of medical background of our project team. To overcome this obstacle, we have decided to select those clinical questions that address relatively simple health issues and have no complicated medical language. The fact that Parkhurst Exchange website, our source of questions, provides both the questions and their answers had a positive impact on the research carried out for our project as it partially compensated the limited medical knowledge in our team.

The website's medical questions are grouped in over 30 categories such as Psychiatry, Oncology, Pediatrics, Endocrinology, etc. In our selection process we have opted for the 'Browse All' categories option which lists all questions sorted descending based on the date they have entered the collection. We have then picked questions that addressed areas that presented relatively straightforward enquiries. A list of examples is included in Table 1. We admit that the question selection process might have introduced bias in our corpus of questions and there is no guarantee that the proportion of questions reflects real life statistics.

Our source of question corpus, Parkhurst Exchange, contains mainly clinical questions asked by family doctors. We assumed we should be able to map the selected questions onto the evidence taxonomy tree. We have classified them as belonging to the 'Intervention' and 'Non-Intervention' categories.

Table 1 shows a snapshot of our corpus of questions and categories resulted from the classification process.

Question	Category
Is watermelon allergenic	No Intervention
When to introduce solids to infants	Intervention
Should family doctors be immunized with Pneumovax and Menactra or Menjugate	Intervention
Can cell phones cause cancer	No Intervention
How much folic acid — 400 µg, 1 mg, 5 mg — is recommended before conception and during pregnancy	Intervention
How to beat recurrent UTIs	Intervention
How to recognize autism in adults	No Intervention
Does skin colour affect vitamin D requirements	No Intervention

Table 1: Example of questions classified according to the Evidence taxonomy.

After completing the question classification process, the resulted structure of our question corpus was 33% Intervention questions and 67% No Intervention questions.

3.2 Search Engines and Question Answering Systems

We have selected the search engines to include in our project based on a few guidelines. They needed to be available online and free of charge and also able to accept natural language questions. Although the initial project plan considered the possibility of transforming the questions into PICO format, this idea was later postponed due to the lack of resources. The PICO format (Niu et al, 2003) has four components that reflect key aspects of patient care: primary problem, main intervention, main intervention comparison and outcome of intervention. Without the option of mapping the input questions to the PICO format, selecting search engines that accepted natural language questions became a must.

Google was included in our study as two different entities: the standard Google and Google pointed towards the PubMed database. Our justification for this approach was the observation that Google returned quite often information from consumer-oriented web sites, as opposed to scientific articles and publications. To make results interpretation more accurate, Google was

pointed to search for information against PubMed (MEDLINE) database, ensuring compatibility with the results provided by the PubMed search engine itself.

The MedQA system proved to be quite unstable, producing parse errors or simply becoming frozen during an answer search cycle. As a result, the evaluation of its performance is not entirely relevant.

3.3 Question processing

Turning knowledge into specific requests for information is not always an easy task. Some information needs are difficult to express and when they can be expressed, the way the question is interpreted influences the delivered answers. Yu et al (2005) mentions that Ely and his colleagues have calculated an average of 2.7 different ways of expressing generic General practitioner's clinical questions. The same study mentions the difficult step of explaining the context of the questions to the information source.

Query modification was applied to the corpus questions when running the original question through a search engine did not produce any relevant results. We have defined five levels of processing which will be applied to improve search outcomes.

The first level of processing involves introducing synonyms or hypernyms of the medical terms in the attempt to improve the performance of the search. Example: we have replaced "infectious conjunctivitis" with "bacterial conjunctivitis".

The next level is detecting any abbreviations that might decrease search engines ability to find answers. Example: we have replaced "BP" with "blood pressure".

The third level is adding general medical terms such as 'disease', 'syndrome' or 'condition' to help clarify the target of the search query. Example: "What is shoulder frozen" has been replaced with "What is frozen shoulder syndrome".

The fourth level we have defined implies eliminating additional grammatical terms such as adverbs and prepositions from the original question. Example: original question "Are there any contraindications to dental office visits in pregnancy" was modified to "Dental office visits in pregnancy".

The fifth level in our processing diagram involves using existing knowledge to transform the question in the attempt to express the medical context. Example: "What is the evidence that antibiotics change the course of the disease in infectious conjunctivitis" became "Are antibiotics recommended for bacterial conjunctivitis".

In order to be able to process the questions through the different levels we have used one of the online medical dictionaries: MedLinePlus[6]. MedLinePlus has extensive information from the National Institute of Health and other trusted sources on over 750 diseases and conditions and is a service offered by the US National Library of Medicine.

A summary of the five levels of question processing is shown in Table 2 below.

Processing Level	Description	Original Question/Term	Processed Question/Term
1	introduce synonyms/hypernyms	infectious	bacterial
2	replace abbreviations	BP	blood pressure
3	Introduce general medical terms	What is shoulder frozen	What is shoulder frozen syndrome
4	eliminate additional terms	Are there any contraindications to dental office visits in pregnancy	Dental office visits in pregnancy
5	express medical context	What is the evidence that antibiotics change the course of the disease in infectious conjunctivitis	Are antibiotics recommended for bacterial conjunctivitis

Table 2: Question processing levels

3.4 Answer extraction

In our attempt to locate answers to our corpus of questions, we have established a limit of 10 first links returned in response to a query. Any other links past this limit, relevant or irrelevant, have been ignored. Any relevant link that refers to a scientific article but does not have an abstract available has been ignored. We have set this rule as usually, if the abstract of the article is not available, the attempt of viewing the full text of the publication fails, requesting a registered username and password.

Most of the search engines included in the study will return a list of links that will then need to be evaluated in order to determine their relevance to the query. This is a time consuming process that MedQA, as a question answering system, manages to overcome by providing a summarized and concise answer. For some instances of our searches, when PubMed returned only one link in response to a query, the abstract was automatically displayed and we were able to locate the answer.

[6] See http://medlineplus.gov/

4 Results

In order to evaluate the results of our answers search, we have used a scoring system first referred to in the Text Retrieval Conference (TREC), called Mean Reciprocal Rank (MRR) (Voorhees, 2001). If a link returned by a search was the nth ($n \leq 10$) position in the list of resulted links, and it was evaluated as being relevant to the question using the Parkhurst Exchange answers as a benchmark, it was given a score of 1/n. We have adopted this methodology in order to assess the ranking system of each search engine. The further down the list, the more effort required from the user to locate the answer. Our evaluation includes the "ease of use" component in our scoring system.

The results of our evaluation are presented in Figures 2 and 3, for the two evidence categories our corpus of questions was mapped to. They have been calculated as an average of scores, per question category and search engine.

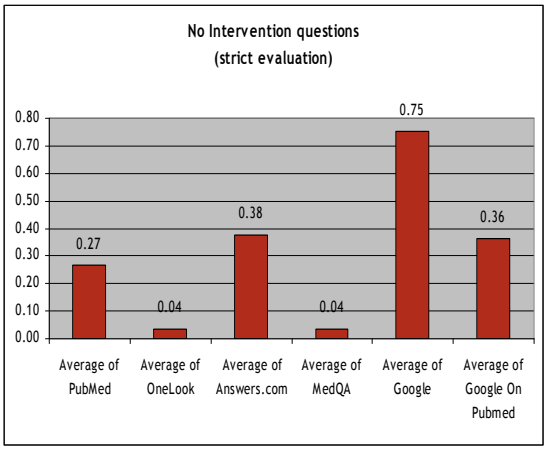

Figure 2: No Intervention questions scores

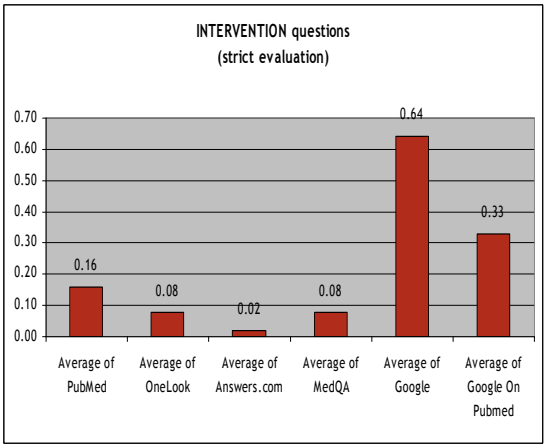

Figure 3: Intervention questions scores

After processing the 40 medical questions through all the selected search engines, we have obtained a total of 113 answers.

The results of the actual location of the answer in a scientific article are shown in Figure 4. Our results show that the answer can be located in one of the sections: abstract, results, conclusions, recommendations, purpose or methods.

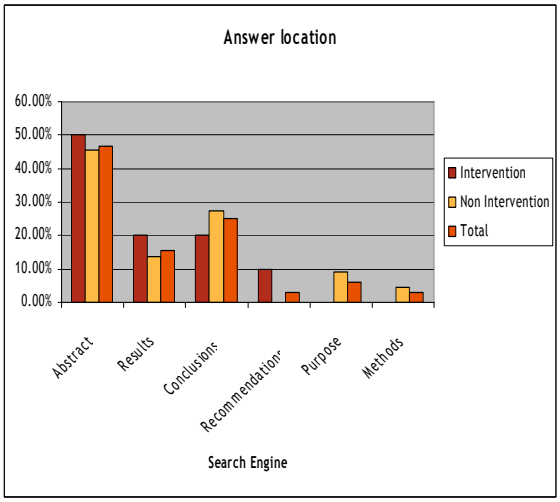

Figure 4: Answer location in scientific articles

The results above do not refer to answers located in consumer oriented websites which do not follow a set document structure. They have been obtained after analyzing the answers extracted from medical scientific articles which represent 28% of our total number of answers.

5 Discussion and Conclusions

As represented in Table 3, our results show that Google has the best performance for both Intervention and Non-Intervention questions. The disadvantage is that the quality of answers provided could be contestable as Google often points to consumer oriented articles rather than scientific publications.

Intervention	Source	Position
	Google	1
	Google On Pubmed	2
	PubMed	3
	OneLook	4
	MedQA	4
	Answers.com	5
No Intervention	Google	1
	Answers.com	2
	Google On Pubmed	3
	PubMed	4
	OneLook	5
	MedQA	5

Table 3: Overall scores

Google on PubMed has the second place for Intervention questions, proving that Google still seems to be one of the best search engines. PubMed was outperformed by Google on PubMed which is sort of a surprise. Analysing the detailed results, this is mainly due to PubMed returning the relevant links further down in the list and consequently obtaining a lower score than Google on PubMed. The conclusion is that the ranking algorithm adopted by PubMed is not performing as well as Google's. This is in line with the observation of Plikus et al (2006) that concluded that PubMed does not produce well classified search outputs and proposed PubFocus as a web server that helps ranking by adding publication quality attributes.

Answers.com performed quite well on No Intervention questions as opposed to Intervention questions. The unbalanced results prove that a non-medical oriented search engine struggles to produce answers for scenario-based, complex medical questions.

MedQA obtained one of the worst scores, but as mentioned earlier, this was mainly due to the fact that the online link was not always up and running.

Analyzing the overall results, 4 out of 5 search engines managed to handle No Intervention questions better than Intervention question, confirming the claim that the complexity of the query has an impact on the results. We have also found out that all the questions in our corpus of questions are answerable with current technology. Going further to the actual location of the answer in medical articles, we have determined that the probability of the answer to be located in the Abstract section of an article is 50%, Conclusions section 25% and Results section 15%. This gives a good indication on the areas a question answering medical system should look most of the time for answers to ad-hoc queries.

Our study results have been compiled on a small set of 40 questions and we admit this might introduce some bias in our process. Our results will have to be confirmed and compared to the performance obtained on a larger corpus of questions (over 200).For a more confident evaluation, we recommend having medically trained teams performing the answers search and evaluation.

References

Berkowitz L, 2002. *Review and evaluation of Internet-based clinical reference tools for physicians*. White Paper commissioned by UpToDate

Brin Sergey, Page Lawrence, 1998. *An Anatomy of a Large-scale Hypertextual Web Search Engine.* Proceedings of the 7th International World Wide Web Conference, Page 107 – 117.

Covell D G, Uman G C, Manning P R. 1985. *Information needs in office practice: are they being met?* Annual Intern Medicine, vol 103, p596-599

Demner-Fushman Dina, Lin J Jimmy. 2007. *Answering clinical questions with Knowledge-based and Statistical Techniques.* Computational Linguistics, 33(1):63-103(2007)

Ely J W, Levinson W, Elder N C, Mainous A G,Vinson D C. 1995 , *Perceived causes of family physician's errors*. Journal of family practice, 40(4):337--44

Niu Yun, Hirst Graeme, McArthur Gregory, Rodriguez-Gianolli Patricia, 2003. *Answering Clinical Questions with Role identification;* Proc. ACL, Workshop on Natural Language Processing in Biomedicine

Plikus V Maxim, Zhang Zina, Chong Cheng-Ming. 2006. *PubFocus:semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithms.* BMC Bioinformatics.

Voorhees Ellen, 2001. *"The TREC question answering track",* Natural Language Engineering, 7(4):361-378, Cambridge University Press.

Yu Hong, Kaufman David. 2007. *A cognitive evaluation of four online search engines for answering definitional questions posed by physicians*. Pacific Symposium on Biocomputing, page 328-339

Yu Hong, Sable Carl, Zhu Hai Ran. 2005. *Classifying Medical Questions based on an Evidence Taxonomy.*AAAI Workshop on Question Answering in Restricted Domains

Yu Hong, Sable Carl. 2005. *Being Erlang Shen: Identifying Answerable Questions.* International Joint Conference of Artificial Intelligence (IJCAI'05), Workshop on Knowledge and Reasoning for Answering Questions

Yu Hong, Lee Minsuk, Kaufman David, Ely John, Osheroff Jerome A., Hripcsak George and Cimino James J. 2007; *Development, implementation and a cognitive evaluation of a definitional question answering system for physicians*; Journal of Biomedical Informatics, 40(3):236-251, Elsevier Science, San Diego, USA