A Study on the Use of Search Engines for Answering Clinical Questions

Andreea Tutos[1] **Diego Mollá**[2]

Department of Computing, Faculty of Science
Macquarie University,
Sydney, NSW 2109

[1] Email: andreea.tutos@students.mq.edu.au
[2] Email: diego.molla-aliod@mq.edu.au

Abstract

This paper describes an evaluation of the answerability of a set of clinical questions posed by physicians. The clinical questions belong to two categories of the five-leaf high-level hierarchical Evidence Taxonomy created by Ely and his colleagues: Intervention and Non Intervention. The questions are passed to two search engines (PubMed, Google), two question-answering systems (MedQA, Answers.com's BrainBoost), and a dictionary (OneLook) for locating the answers to the question corpus. The output of the systems is judged by a human and scored according to the Mean Reciprocal Rank (MRR). The results show the need for question modification and analyse the impact of specific types of modifications. The results also show that No Intervention questions are easier to answer than Intervention questions. Further, generic search engines like Google obtain higher MRR than specialised systems and even higher than a version of Google based on specialised literature (PubMed) only. In addition, an analysis of the location of the answer in the returned documents is provided.

Keywords: Question Answering, Evidence Based Medicine, Search, Evaluation.

1 Introduction

Latest clinical guidelines urge physicians to practise Evidence Based Medicine (EBM) 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 (Sackett et al. 1996). However, current practise of EBM is challenged by the large amounts of external evidence information available to the medical practitioner. The number of biomedical publications is increasing to the point where thousands of new articles are published daily world wide, and no human can keep up to date without help. 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 study about the major obstacles to answering doctor's questions about patient care with evidence (Ely et al. 2002) highlighted, among other factors, the excessive time required to find the information, the difficulty in formulating an adequate question according to recommended practise in EBM, and the difficulty of synthesising multiple bits of evidence into a clinically useful statement. All of these issues are targets of current research in text-based question answering. We envision a scenario whereby the practitioner would ask a question using his or her words, and the system would search for the evidence and present it in the most effective way.

Our project is a step towards assessing the potential of the use of question-answering technology to access external evidence stored in the Internet by studying the answerability of a set of 50 medical questions sourced from the Parkhurst Exchange[1] website. We study the relevance of answers located through two selected search engines: PubMed[2] and Google, two question-answering systems: MedQA[3] and Answers.com's BrainBoost[4], and a dictionary: OneLook.[5] In the process we perform an initial study of the modifications required for the questions to facilitate the retrieval of the answers by the above tools.

Our work is related to the study by Yu & Kaufman (2007) who conducted a cognitive evaluation of four online engines on answering definitional questions. 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. Our study does not limit the input questions to definitional questions only. We use a wider range of questions belonging to the 'Evidence' node in the Evidence Taxonomy introduced by Ely et al. (2002). Further, we study the ability of freely available systems to provide documents containing the answer, and the relative position of the answer-bearing documents in the ranked list presented to the user. In addition, we explore specific types of query modifications that can be made to find the documents.

The structure of this document is as follows: Section 2 describes studies and concepts related to our project. Section 3 introduces the evaluation methodology employed in the study. The methodology details the corpus of questions and how it has been selected. It details the classification of candidate questions according to the 'Evidence' node in the Evidence Taxonomy. It also describes the selected systems and

This work forms part of a student project in Macquarie University's masters unit ITEC810.

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

the reasons behind their selection. A description of question processing follows together with a section on answer extraction. Section 4 presents the results of the evaluation. Section 5 analyses the results. Finally, Section 6 provides a summary and an indication of lines of future work.

2 Background

2.1 Question Answering

There has been considerable research in the area of open-domain Question Answering (QA). This research has been mainly driven by the Text REtrieval Conference (TREC) (Voorhees 2001), and more recently by the Cross Language Evaluation Forum (CLEF) (Vallin et al. 2005), the workshops by the NII Text Collection for IR Project (NTCIR) (Kando 2005), the Document Understanding Conference (DUC) (Dang 2006), and the Text Analysis Conference (TAC) (Dang 2008). Open-domain QA initially focused on fact-based questions that expected short answers, but more recently (e.g. in DUC and TAC) questions allowed more complex answers that are the result of combining information from multiple documents. This is the sort of questions that are applicable to the biomedical domain.

The biomedical domain is a specialised domain that presents challenges and opportunities that make it a very useful area for researchers, together with the potential of being very beneficial to the users (Zweigenbaum 2003, Zweigenbaum et al. 2007, Mollá & Vicedo 2007). In particular, there are collections of documents which can be used as corpora for searching the answers. For example, MEDLINE is a collection of abstracts maintained by the US National Library of Medicine (NLM) that contains more than 17 million records dating back to 1966. There are also terminological resources such as NLM's Medical Subject Headings (MeSH), which contains an extensive list of diseases, drugs and treatments. And there are tools like PubMed Central[6] which provides an interface to MEDLINE and incorporates query expansion using MeSH in an attempt to find documents that are related to the question. The time is ripe for the development of question-answering technology for the biomedical domain.

There have been some attempts to integrate question-answering technology to the medical domain. Some methods are based on the so-called PICO format to formulate the questions. The PICO format (Niu et al. 2003) has four components that reflect key aspects of patient care and which are recommended for the practise of Evidence Based Medicine: primary **P**roblem, main **I**ntervention, main intervention **C**omparison, and **O**utcome of intervention. Current systems presume a preliminary stage that converts the question to the PICO format that can be easily processed by the computer (Niu et al. 2003, Demner-Fushman & Lin 2007). However, not all clinical questions (even among those that are strictly evidence-based questions) can be mapped in terms of PICO elements (Huang et al. 2006). There is also evidence that even doctors may find it difficult to formulate the question in terms of the PICO format (Ely et al. 2002). Therefore, research focusing on the PICO format will first need to show that it is possible to automatise the analysis of questions into the PICO format, or at least to provide tools that would help the practitioner to formulate PICO questions. This work falls outside the scope of this paper and therefore we do not use PICO in our experiments.

[6] http://www.ncbi.nlm.nih.gov/pmc/

I. Clinical (n=193)
  A. General (n=141)
    1. Evidence (n=106)
      a. Intervention (n=71)
        What is the drug of choice for epididymitis?
      b. No Intervention (n=35)
        How common is depression after infectious mononucleosis?
    2. No Evidence (n=35)
      What is the name of that rash that diabetics get on their legs?
  B. Specific (n=52)
    What is causing her anaemia?
II. Non-clinical (n=7)
  How do you stop somebody with five problems, when their appointment is only long enough for one?

Figure 1: Evidence Taxonomy used to classify 200 questions from family doctors

MedQA (Yu et al. 2007) is a recent medical answering system that responds to definitional questions by accessing the MEDLINE records and other World Wide Web collections. It automatically analyses a large number of electronic documents in order to generate short and coherent answers in response to the input questions. The reason behind using 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. MedQA relies on the IMRAD (Introduction, Methods, Results and Discussion) structure of biomedical articles to determine the relevance of an article to the search query. MedQA is the first system to integrate end-to-end QA technology including question analysis, information retrieval, answer extraction and summarisation techniques (Lee et al. 2006). The system includes a Web demo, but unfortunately the demo was often not functional when the experiments reported in the present study were carried out.

2.2 Evidence Taxonomy

Our work uses the Evidence Taxonomy created by Ely et al. (2002). This high-level, five-leaf hierarchy categorises medical questions that are potentially answerable with evidence. The hierarchy is presented in Figure 1, with the examples given in the original paper.

Ely et al. (2002) 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 the patient personal record.

We have focused on the two evidence categories confirmed as being answerable with evidence according to Ely et al. (2002): 'Intervention' and 'No 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. 'No 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.

TITLE: Is watermelon allergenic?

QUESTION: "A 16-year-old female patient had an urticarial reaction from watermelon. She now avoids eating it," writes ABDULRAHEM LAFTAH, MD, of Watson Lake, Yukon. "What substance in watermelon would have caused the attack, and are there other related foods she should now stay away from?"

ANSWER: Watermelon does contain allergenic proteins that could provoke an IgE-dependent urticarial response. You can refer the patient for allergy skin testing to determine if this fruit was indeed the culprit. Watermelon belongs to a family of foods associated with ragweed pollen. These include cantaloupe, honeydew, zucchini, banana, cucumber and chamomile tea. Individuals suffering from ragweed allergic rhinitis may develop symptoms, often mild, after eating these foods. This is particularly true during or following hay fever season, when their IgE to ragweed is the highest. PK

Figure 2: Sample of question and answer

3 Evaluation Methodology

3.1 Corpus of Questions

The corpus of questions of our study has been constructed from the question and answer list available on the Parkhurst Exchange website.[1] Parkhurst Exchange is a medical publishing website based in Canada that includes a collection of over 4,800 clinical questions and their answers provided by physicians. Since 1983 when it first started, it reportedly continues to develop strong relationships with top physicians across many medical disciplines.

To determine whether the output of a system contains the answer we relied on human judgement (the first author of this paper). To facilitate this judgement, we have selected clinical questions that address relatively simple health issues and have no complicated medical language. Figure 2 shows an example of a question and answer that we used in our study. As the figure shows, the answers are not simple factoids typical of current QA systems.

The website's medical questions appear grouped in over 30 categories such as Psychiatry, Oncology, Pediatrics, Endocrinology, etc. In our selection process we have opted for the 'Browse All' 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 therefore the results presented in this study are of a preliminary manner and need to be verified with a larger set of questions.

Parkhurst Exchange contains mainly clinical questions asked by family doctors. We have mapped our selection to the Evidence Taxonomy tree. All of the questions were classified as belonging to the 'Intervention' (46%) and 'No Intervention' (54%) categories.[7] This distribution is relatively close to the percentages of the study by Ely et al. (2000).

[7] Whereas all the questions we looked at were classified as either 'Intervention' or 'No Intervention', we didn't check whether *all* questions in Parkhurst Exchange could be classified this way.

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 g, 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

Is watermelon allergenic?

("citrullus"[MeSH terms] OR "citrullus"[All Fields] OR "watermelon"[All Fields]) AND allergenic[All Fields]

Figure 3: A simple PubMed query and its expanded form

3.2 Search Engines and Question Answering Systems

We have selected the systems to test based on a few guidelines. They needed to be available online and free of charge and also be able to accept natural language questions. We initially considered the possibility of transforming the questions into PICO format. However, this idea was later postponed due to the intrinsic problems of formulating a query into PICO as described in Section 2.1. Without the option of mapping the input questions to the PICO format, selecting systems that accepted natural language questions became a must.

PubMed is a search engine that accesses a reputable medical repository (MEDLINE) maintained by the US National Library of Medicine (Demner-Fushman & Lin 2007). The MEDLINE database includes over 19 million medical articles and is a well recognised knowledge source across medical question answering studies. PubMed uses MeSH to expand the query with related terms. Figure 3 shows an example of a query and its expanded form.

Google is a popular 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 an effective 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 characteristic of Google is that it establishes a relation between the text of links and the pages the links point to (Brin & Page 1998).

Two variants of Google were included in our study:

the standard Google and Google pointed towards the PubMed database. The reason for the second variation was the observation that quite often Google returned information from consumer-oriented web sites rather than scientific articles and publications. To make the results easier to compare, Google was pointed to search for information against PubMed (MEDLINE) database, ensuring compatibility with the results provided by the PubMed search engine itself. Using Google on PubMed only also addresses any possible concerns about the quality of the information provided by user-oriented sites indexed by Google.

MedQA (Yu et al. 2007) is one of the first developed end-to-end medical answering systems and responds to definitional questions accessing the MEDLINE records and other World Wide Web collections. It automatically analyses a large number of electronic documents in order to generate short and coherent answers in response to the input questions.

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. A subsequent attempt to rerun all questions through the answering system proved even more unsuccessful, as we were unable to obtain any answers due to a 404 HTTP error ("the requested resource is not available").

Answers.com is a website that offers 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. Apart from its repository of questions and answers, Answers.com also hosts BrainBoost, a generic end-to-end question answering system that highlights the answer to the user's questions. In our experiments we used BrainBoost[4] rather than the general answers.com site.

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, though we did not use this feature in our experiments.

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) calculated an average of 2.7 different ways of expressing generic General Practitioner's clinical questions. The same study mentions the difficulty of explaining the context of the questions to the information source.

Question processing is therefore an important and difficult task in QA. The specific task of automated classification of clinical questions still has room for improvement, as illustrated by the results reported by Yu et al. (2005) on the classification of questions according to the Evidence Taxonomy (less than 60% accuracy for the five-category classification), and the results by Yu & Cao (2008) on a different taxonomy by the National Library of Medicine (76% F-score). These results are below those of generic question classification systems such as the one by Li & Roth (2002) (up to 98.80% accuracy for a six-category classification). Question analysis is the largest source of errors in generic question-answering systems, with over 50% of the errors attributed to this stage by Moldovan et al. (2003). We therefore expect an even larger impact of question processing for medical question answering.

To mitigate the difficulties of question processing in our study, we applied query modification to every question that did not produce any relevant results when run unmodified through all systems. Then, exactly the same (possibly modified) question was sent to all systems. We did this to obtain results that are comparable across all systems.

We applied five levels of processing, in this order, until a system returned relevant results:

1. Introduce synonyms, hyponyms, and hypernyms of the medical terms in an attempt to improve the performance of the search. Example: we replaced *infectious conjunctivitis* with *bacterial conjunctivitis*.
2. Expand any abbreviations that might decrease the system's ability to find answers. Example: we replaced *BP* with *blood pressure*.
3. Add general medical terms such as *disease*, *syndrome* or *condition* to help clarify the target of the search query. Example: *What is shoulder frozen* was replaced with *What is frozen shoulder syndrome*.
4. Eliminate additional grammatical terms such as adverbs and prepositions from the original question. Example: the original question *Are there any contraindications to dental office visits in pregnancy* was modified to *Dental office visits in pregnancy*.
5. Use external knowledge to transform the question as an 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*.

The query modifications were made manually. To source relevant words we used the online dictionary MedLinePlus.[8] MedLinePlus has extensive information from the National Institutes 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.

In order to evaluate the efficiency of our question processing and the degree to which each defined level of transformation had a positive impact on the search results, we have analysed the questions that did not produce any relevant answers when run through the systems in their original form. We have then determined which level of transformation has been applied in order to get a relevant answer. If after applying a particular level of processing, we have obtained a relevant answer or link, we have flagged that question as being improved by Level x of transformation. In order to quantify if there was an improvement, we did not consider the position of the relevant link on the results page and did not try to improve the relevant link position in the list by applying a subsequent level of processing. After computing the results we have observed that Level 4 of processing "Eliminate additional terms" was applied with the highest frequency

[8] http://medlineplus.gov/

Level	Description	Original Question	Processed Question
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

(45.95% of total successful transformations), followed by Level 5 "Express medical context" (27.03% of total successful transformations). The results are presented in Table 3.

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 links that refer to a scientific article but do not have an abstract available have 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 systems included in the study would return a list of links that 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 summarised 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.

4 Results

In order to evaluate the results of our retrieval systems, we have used the Mean Reciprocal Rank (MRR), an evaluation measure frequently used in question-answering evaluations and first introduced in TREC (Voorhees 2001). If a link returned by a search was in the nth ($n \leq 10$) position in the list of results, 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 system. 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.

In order to evaluate whether a summarised answer or a link returned in response to a search query is relevant, a human judge (the first author of this paper) has referred to the answer provided by the Parkhurst Exchange website. We initially opted for a lenient evaluation, in the sense that a link or summarised answer that was relatively relevant to the question received a score that was giving them a credit lower that the 10th position of a relevant answer in the top 10 list: 1/11. However we have later revised this scoring system as we came to the conclusion that it was possible that this methodology was introducing bias in our evaluation. We have decided to stick with the strict evaluation that only gives credit to links or summarised answers that express the same ideas as the Parkhurst Exchange benchmark answer. This decision was also supported by the limited medical knowledge of the human judge, which difficulted a comprehensive evaluation and judgement of diagnosis, drugs and treatments that are related to the search question.

After processing the 50 medical questions through all the selected systems, we have obtained a total of 119 answers.

The results of our evaluation are presented in Table 4, 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 system.

The results of the actual location of the answer in a scientific article are shown in Table 6. This table shows the percentage of occurrences of the answer in a specific section (note that a document did not necessarily have all the sections listed in the table). Our results show that the answer can be located in one of the following sections: abstract, results, conclusions, recommendations, purpose or methods. The abstract was the section that most likely contained the answer.

The results of Table 6 do not refer to answers located in consumer oriented websites which do not follow a set document structure. They have been obtained after analysing the answers extracted from medical scientific articles which represent 34% of our total number of answers.

5 Discussion of Results

The results are summarised in Table 5 and show the following:

Google performed better than the other systems tested for both Intervention and Non-Intervention questions. Google on PubMed also has the second place for both Intervention and No Intervention questions, showing that Google still seems to be a comparatively good system.

PubMed was outperformed by Google on PubMed. Analysing the detailed results, we concluded that Google's advantage was mainly due to PubMed returning the relevant links further down in the list and consequently obtaining a lower score than Google on

Processing level	1	2	3	4	5
How often level was applied	5.41%	10.81%	10.81%	45.95%	27.03%

Table 3: Question Processing Results

	PubMed	OneLook	Answers.com	MedQA	Google	Google on PubMed
No Intervention	0.27	0.04	0.38	0.04	**0.80**	0.41
Intervention	0.24	0.04	0.10	0.04	**0.54**	0.35

Table 4: Question Scores (MRR@10)

	Source	**Position**
Intervention	Google	1
	Google on PubMed	2
	PubMed	3
	OneLook	4
	MedQA	4
	Answers.com	6
No Intervention	Google	1
	Google on PubMed	2
	Answers.com	3
	PubMed	4
	OneLook	5
	MedQA	5

Table 5: Overall scores

PubMed. Apparently the ranking algorithm adopted by PubMed is not performing as well as Google's. This is in line with the observation by Plikus et al. (2006) that concluded that PubMed does not produce well classified search outputs and proposed PubFocus to help ranking by adding publication quality attributes. Table 7 provides some examples of rankings for the same link in PubMed as opposed to Google on PubMed. It is quite obvious that in those instances Google assigned a better score than PubMed for the same relevant link.

The systems were generally bad at detecting acronyms. Apparently any possible advantages of PubMed's automated query expansion were offset by our manual modification of the query. We indeed observed that PubMed did not detect acronyms, and often questions presenting acronyms were processed satisfactorily only after manual acronym expansion. Analysing the PubMed search engine behaviour we noticed that even if MEDLINE benefits from MeSH, the controlled vocabulary thesaurus, by expanding the query with related terminology, it still underperforms in retrieving relevant answers for questions using acronyms. An example is the original question *Is it a good idea to take ASA before an extended period of air travels* in which *ASA* is the medical acronym for *acetylsalicylic acid*. We expected that Google, as a generic search engine, would not be able to handle the acronym, but PubMed was also not able to translate it and could not provide any answers until we manually processed the question and replaced the abbreviation in the question with the explicit chemical substance name.

A rather surprising observation was that Google outperformed Google on PubMed. We attribute this to the much larger corpus of text indexed by Google as compared with Google on PubMed. Google's search system and ranking of results is designed for large volumes of highly hyperlinked data and therefore the reduced data in PubMed may affect its ability to rank the results effectively. However, even though Google obtained the best MRR scores, we still need to evaluate whether Google's returned text would be acceptable to a medical practitioner.

Generally, systems were better on 'No Intervention' questions. Analysing the performance of the generic search engines and question answering systems, we observed that Google performed better on 'No Intervention' questions with an average score of 0.80 as opposed to 0.54 for 'Intervention' questions. Answers.com also proved better on 'No Intervention' questions than 'Intervention' questions. The results show that the systems have more difficulties on producing answers for scenario-based, complex medical questions.

Overall Answers.com performed much better than OneLook and this could be explained by the fact that Answers.com is designed to answer questions and incorporates question-answering techniques (BrainBoost), whereas OneLook is basically a dictionary and therefore only able to handle definitions. In particular, OneLook only managed to answer two questions out of the 50 included in our corpus questions: one 'Intervention' question and one 'No Intervention' question. These results show that OneLook is currently not suitable as a potential technology for medical answering systems.

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.

We have also found out that, after manual query modification, all the questions in our corpus of questions were answerable with current technology, which we consider to be an important finding for future medical question answering systems as it indicates the potential benefit of developing question answering systems. The most effective query transformation consisted of eliminating noise from the query (45.98% of questions), followed by using expert knowledge to transform the query (27.03%). An example is the original question *What's the best antihistaminic for mild acute urticaria in infants and children?* for which PubMed could not locate an answer until we have transformed it into *antihistaminic for mild acute urticaria in children*. The impact of introducing synonyms was relatively low (5.41%).

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 26.19% and Results section 14.29%. 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 50 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. For a more confident evaluation, we recommend random sampling of the Parkhurst Exchange or another corpus that more closely reflects the characteristics of questions asked

	Abstract	Results	Conclusions	Recommendations	Purpose	Methods
Non Intervention	43.48%	17.39%	26.09%	0.00%	8.70%	4.35%
Intervention	57.89%	10.53%	26.32%	5.26%	0.00%	0.00%

Table 6: Percentage of answer location in scientific articles

Question No.	*Question*	*Category*	*PubMed*	*Google on PubMed*
19	*When should moles be removed?*	Intervention	7	2
43	*What can be done for a patient with persistent (non-typhoid) Salmonella in stool, despite 2 antibiotics*	Intervention	7	4
48	*Is it a good idea to take an ASA before an extended period of air travel?*	Intervention	5	4

Table 7: Google on PubMed vs. PubMed ranking for the same relevant link

by medical doctors.

6 Summary and Future Work

We have presented a study of the answerability of documents returned by current freely available technology within the domain of clinical question answering. Our study shows that current technology is able to find the answers to the asked questions, though the questions need to be transformed. We have applied a set of question transformations and evaluated the impact of transformation on the answerability of results returned. It is our intention to explore methods to automatically perform such transformations to increase recall.

Our study also includes an analysis of the location of the answer. This analysis needs to be extended by considering the type of question and other factors and narrow down the actual zoning that could be done to find the answer.

We obtained the surprising result that Google performs best than any other systems, including PubMed which is specialised on medical text, and even including Google on PubMed documents only.

Future work includes an evaluation of PubMed enhanced with an optimised ranking system such as the one provided by PubFocus. We would then compare these results with the previous PubMed performance and of Google on PubMed.

Another line of future work is determining the actual quality of the answers returned by a particular search engine or question answering system. The study presented was based on MRR as the measure to compare all systems. Note, however, that MRR is only concerned with the location of the document containing the answer but it does not measure the quality of the answers presented or the impact that the erroneous answers may produce in the judgement of the medical doctor. It is therefore desirable to evaluate the acceptance of the answer returned by the medical practitioner, and any possible errors of judgement that non-relevant texts could introduce. The study of answer quality will also help to determine the best technology to extract the answer and present it to the user.

References

Brin, S. & Page, L. (1998), The anatomy of a large-scale hypertextual web search engine, *in* 'Proc. WWW-7', Brisbane, Australia.

Dang, H. T. (2006), DUC 2005: Evaluation of question-focused summarization systems, *in* 'Proceedings of the Workshop on Task-Focused Summarization and Question Answering', Association for Computational Linguistics, Sydney, pp. 48–55.

Dang, H. T. (2008), Overview of the tac 2008 opinion question answering and summarization tasks, *in* 'Proc. TAC 2008'.

Demner-Fushman, D. & Lin, J. J. (2007), 'Answering clinical questions with knowledge-based and statistical techniques.', *Computational Linguistics* **33**(1), 63–103.

Ely, J., Osheroff, J. A., Ebell, M. H., Chambliss, M. L., Vinson, D., Stevermer, J. J. & Pifer, E. A. (2002), 'Obstacles to answering doctors' questions about patient care with evidence: Qualitative study', *BMJ* **324**(7339), 710.

Ely, J., Osheroff, J. A., Gorman, P. N., Ebell, M. H., Chambliss, M. L., Pifer, E. A. & Stavri, P. Z. (2000), 'A taxonomy of generic clinical questions: Classification study', *British Medical Journal* **321**(7258), 429–432.

Huang, X., Lin, J. & Demner-Fushman, D. (2006), Evaluation of PICO as a knowledge representation for clinical questions, *in* 'AMIA Annu Symp Proc.', pp. 359–363.

Kando, N. (2005), Overview of the fifth NTCIR workshop, *in* 'Proceedings NTCIR 2005'.

Lee, M., Cimino, J., Zhu, H. R., Sable, C., Shanker, V., Ely, J. & Yu, H. (2006), Beyond information retrieval — medical question answering, *in* 'Proc. AMIA 2006', p. 6 pages.

Li, X. & Roth, D. (2002), 'Learning question classifiers', *Proc. COLING 02* .

Moldovan, D., Pas ca, M., Harabagiu, S. & Surdeanu, M. (2003), 'Performance issues and error analysis in an open-domain question answering system', *ACM Transactions on Information Systems* **21**(2), 133–154.

Mollá, D. & Vicedo, J. L. (2007), 'Question answering in restricted domains: An overview', *Computational Linguistics* **33**(1), 41–61.

Niu, Y., Hirst, G., McArthur, G. & Rodriguez-Gianolli, P. (2003), Answering clinical questions with role identification, *in* 'Proc. ACL, Workshop on Natural Language Processing in Biomedicine'.

Plikus, M., Zhang, Z. & Chuong, C. M. (2006), 'PubFocus: Semantic MEDLINE/PubMed citations analysis through integration of controlled biomedical dictionaries and ranking algorithm', *BMC Bioinformatics* **7**(1), 424.

Sackett, D. L., Rosenberg, W. M., Gray, J., Haynes, R. B. & Richardson, W. S. (1996), 'Evidence based medicine: What it is and what it isn't', *BMJ* **312**(7023), 71–72.

Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., nas, A. P., de Rijke, M., Sacaleanu, B., Santos, D. & Sutcliffe, R. (2005), Overview of the CLEF 2005 multilingual question answering track, *in* 'Proceedings CLEF 2005'. Working note.

Voorhees, E. M. (2001), 'The TREC question answering track', *Natural Language Engineering* **7**(4), 361–378.

Yu, H. & Cao, Y.-g. (2008), Automatically extracting information needs from ad hoc clinical questions, *in* 'AMIA Annu Symp Proc.', pp. 96–100.

Yu, H. & Kaufman, D. (2007), A cognitive evaluation of four online search engines for answering definitional questions posed by physicians, *in* 'Proc. Pacific Symposium on Biocomputing', pp. 328–339.

Yu, H., Lee, M., Kaufman, D., Ely, J., Osheroff, J. A., Hripcsak, G. & Cimino, J. J. (2007), 'Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians.', *Journal of Biomedical Informatics* **40**(3), 236–251.

Yu, H., Sable, C. & Zhu, H. R. (2005), Classifying medical questions based on an evidence taxonomy, *in* 'Proc. AAAI'05 Workshop on Question Answering in Restricted Domains'.

Zweigenbaum, P. (2003), Question answering in biomedicine, *in* 'Proc. EACL2003, workshop on NLP for Question Answering', Budapest.

Zweigenbaum, P., Demner-Fushman, D., Yu, H. & Cohen, K. B. (2007), 'Frontiers of biomedical text mining: current progress.', *Briefings in Bioinformatics* **8**(5), 358–375.