Development of a Corpus for Evidence Based Medicine Summarisation

Diego Molla
Department of Computing
Macquarie University
Sydney, Australia
diego.molla-aliod@mq.edu.au

Maria Elena Santiago-Martinez
Department of Computing
Macquarie University
Sydney, Australia
maria.santiago-martinez@mq.edu.au

Abstract

In this paper we introduce some of the key NLP-related problems related to the practice of Evidence Based Medicine and propose the task of multi-document query-focused summarisation as a key approach to solve these problems. We have completed a corpus for the development of such multi-document query-focused summarisation task. The process to build the corpus combined the use of automated extraction of text, manual annotation, and crowdsourcing to find the reference IDs. We perform a statistical analysis of the corpus for the particular use of single-document summarisation and show that there is still a lot of room for improvement from the current baselines.

1 Introduction

An important form of medical practice is based on Evidence Based Medicine (EBM). (Sackett et al., 1996; Sackett et al., 2000). Within the EBM paradigm, the physician is urged to consider the best available evidence that is relevant to the patient at point of care. However, the physician is currently overwhelmed with the large volumes of published text available. For example, the US National Library of Medicine offers PubMed[1], a database of medical publications that comprises more than 19 million abstracts. The median time spent to conduct a clinical systematic review is 1,139 hours (Allen and Olkin, 1999). In contrast, the average time that a physician spends searching for a topic is two minutes (Ely et al., 1999). In practice, the physician would typically try to keep up to date by reading systematic reviews. However, systematic reviews are generic studies that may or may not be applicable to the particular case that the physician is concerned with. When there are no appropriate systematic reviews, the physician will need to search over the research literature, find the relevant information, and appraise it in terms of quality of the results and applicability to the patient (Sackett et al., 2000).

There is a range of NLP tasks that have been attempted on this area, but so far not much work has been done on multi-document query-based summarisation. We argue that this task would greatly help the physician but the lack of appropriate corpora has hindered the development and testing of such query-based summarisers for this domain. In this paper we present such a corpus, show some characteristics of the corpus, and advance some specific tasks that the corpus is suited for.

Section 2 introduces EBM and its connection with tasks related to multi-document query-based summarisation. Section 3 describes the corpus. Section 4 details how the corpus was built. Section 5 gives an indication of the use of the corpus for the specific task of single-document summarisation. Finally, Section 6 concludes the paper.

2 Evidence Based Medicine and Summarisation

In this section we introduce EBM and present work related to the use of NLP for EBM.

[1] http://www.ncbi.nlm.nih.gov/pubmed

Diego Mollá and Maria Elena Santiago-Martinez. 2011. Development of a Corpus for Evidence Based Medicine Summarisation. In *Proceedings of Australasian Language Technology Association Workshop*, pages 86−94

2.1 Evidence Based Medicine

There are two key components in EBM: clinical expertise and external clinical evidence (Sackett et al., 1996). Clinical expertise is gained through clinical experience and clinical practice, whereas external clinical evidence needs to be obtained by consulting external sources. Systematic reviews enable physicians to quickly acquire the best evidence for a selection of topics. Such reviews are written by domain experts and are found at libraries such as the Cochrane Library[2] and UpToDate[3], to name two of the better known ones. However, EBM guides are quick to point out that there is not always a systematic review that addresses the specific topic at hand (Sackett et al., 2000) and then a search on the primary literature becomes necessary.

Ely *et al.* (Ely et al., 2002) highlight the following six obstacles for investigators and physicians to search and find the evidence: (1) the excessive time required to find information; (2) difficulty to modify the original question; (3) difficulty selecting an optimal strategy to search for information; (4) failure of a seemingly appropriate resource to cover the topic; (5) uncertainty about how to know when all the relevant evidence has been found; and (6) inadequate synthesis of multiple bits of evidence into a clinically useful statement. In this paper we will address the specific NLP technologies that can be used to overcome these obstacles, with special emphasis on summarisation technology.

The standard recommendation within EBM is to search the literature by determining specific information according to the PICO mnemonic (Armstrong, 1999). PICO highlights four components that reflect key aspects of patient care: primary **P**roblem or population, main **I**ntervention, main intervention **C**omparison, and **O**utcome of intervention.

PICO helps determining what terms are important in a query and therefore it helps building the query, which is sent to the search repositories. Once the documents are found, they need to be read by a person who eliminates irrelevant documents.

The retrieved documents need then to be appraised according to the strength of the evidence of the information reported in them. A number of guidelines for appraisal have been established. The Strength of Recommendation Taxonomy (SORT) (Ebell et al., 2004) is one of the better known ones and it specifies a scale of three grades based on the quality and type of evidence:

A Grade Consistent and good-quality patient-oriented evidence.

B Grade Inconsistent or limited-quality patient-oriented evidence.

C Grade Consensus, usual practice, opinion, disease-oriented evidence, or case series for studies of diagnosis, treatment, prevention, or screening.

Patient-oriented evidence relates to the impact in the patient (e.g. effect in mortality or in their quality of life), as opposed to disease-oriented evidence (e.g. lowering of blood pressure or blood sugar). Quality of evidence is assessed by the type of study (diagnosis, treatment, prevention, prognosis) and relevant variables for assessing the quality of evidence are the size and randomisation of the subjects and the consistency of the results.

As a final step, the physician still needs to locate the specific information presented in the documents. Current resources offer an array of presentation methods ranging from a list of bibliographic data (title, authors, publication details) sorted by date in PubMed to the clustering of information according to fields such as treatments, causes of condition, complications of condition, and pros & cons of treatment in HealthBase.[4]

2.2 Summarisation for Evidence Based Medicine

An important amount of research has been carried out on many aspects of medical support systems (Demner-Fushman et al., 2009; Zweigenbaum et al., 2007). In this section we present some of the NLP research that is relevant to EBM, with special emphasis on tasks that are related to multi-document query-based summarisation.

Much of the current work in NLP for EBM can be categorised as aiming to retrieve the evidence.

[2] http://www.thecochranelibrary.com/

[3] http://www.uptodateonline.com/

[4] http://healthbase.netbase.com

Recent studies aiming at increasing recall show that both Boolean and ranked retrieval have their limitations (Karimi et al., 2009). Using the Cochrane systematic reviews and their queries as sample data, Karimi *et al.* (Karimi et al., 2009) show that a combination of Boolean and ranked retrieval methods outperforms each of them individually but recall is still under 80% and precision is as low as 2.7% (Karimi et al., 2009).

The evidence found needs to be ranked by order of importance. A problem of PubMed is that the results are not presented in order of relevance or of importance. It is telling that, for example, generic search engines often find and present the correct information in a more prominent rank than specialised search engines like PubMed do, though the source of the information from where the answer is found is often questionable (Berkowitz, 2002; Tutos and Mollá, 2010). This has been addressed by PubFocus, which incorporates ranking functionality based on bibliometric data (Plikus et al., 2006).

Judging the quality of the evidence is one of the principal steps in EBM practice, and we advance that a good EBM summariser should provide information about the quality of the evidence summarised. Berkowitz (2002) mentioned that Google did "surprisingly well [in his study], but [it showed] low validity overall." If the information given is not from a reliable source it is not usable. PubMed abstracts contain meta-data information including the study type (e.g. "meta-analysis", "review") that can be used to filter the search results. This information is used by published search strategies (e.g. (Shojania and Bero, 2001; Haynes et al., 1994; Haynes et al., 2005)). Current implementations incorporating appraisal of the quality use information based on word co-occurrences (Goetz and von der Lieth, 2005) and bibliometrics (Plikus et al., 2006). More closely related to EBM are attempts to grade papers according to SORT or similar taxonomies (Tang et al., 2009; Sarker et al., 2011).

Question Answering (QA) technology is naturally suitable for the task of finding the required information, and in fact Zweigenbaum (2003) has argued for the use of the resources available in the medical domain to implement QA systems. However, the questions addressed by current QA technology seek simple answers. Whereas QA technology has traditionally focused on seeking names, lists, and definitions, EBM seeks more complex information that includes the type and quality of evidence.

Some QA systems for clinical answers are based on the PICO information. Those question-answering systems presume a preliminary processing stage that clearly identifies each component of PICO so that it can be processed by the computer, such as EPoCare's QA system (Niu et al., 2003) and CQA-1.0 (Demner-Fushman and Lin, 2007). Both EPoCare and CQA-1.0 follow specialised strategies to identify information addressing each field of the PICO query.

Some QA systems focus on specific kinds of questions. MedQA[5] (Yu et al., 2007) focuses on definitional questions. It accepts unstructured questions and integrates technology including question analysis, information retrieval, answer extraction and summarisation techniques (Lee et al., 2006). The work by Leonhard (2009), in contrast, focuses on comparison questions.

It has been shown that physicians want help to locate the information quickly by using lists, tables, bolded subheadings and by avoiding lengthy, uninterrupted prose (Ely et al., 2005). One of the findings by Ely et al. (2002) is the difficulty to synthesise the multiple bits of evidence into a clinically useful statement, which is the task of summarisation technology. The survey by Afantenos et al. (2005) presents various approaches to summarisation, including multi-document summarisation, from medical documents. Of particular interest are the context-based multi-document summarisation approaches such as CENTRIFUSER (Elhadad et al., 2005), which builds structured representations of the documents as source for the summaries.

SemRep (Fiszman et al., 2004) provides abstractive summarisation of biomedical research literature by producing a semantic representation based on the UMLS concepts and their relations as found in the text. The semantic representation is a set of predications (concept)-relation-(concept) that is presented graphically to the user.

Clustering methods can also help present the information. The Trip database,[6] for example, clus-

[5]This system is integrated in AskHERMES, http://www.askhermes.org/

[6]http://www.tripdatabase.com/

ters the search results by publication type and incorporates a sliding control to filter out publication types associated with lesser quality. The system by Demner-Fushman and Lin (2006) clusters the results by the intervention component of PICO. Using UMLS as a resource, interventions mentioned in the text are grouped into common categories and the clusters are presented labelled with the intervention type. The resulting system outperformed PubMed in their evaluations.

All of the techniques mentioned above are related to summarisation technology in one or another form, or are actual summarisation systems. By working on query-based multi-document summarisation for EBM we are contributing to some of the above research areas, and we are aiming at helping the physician practice EBM efficiently.

3 Source and Structure of the Corpus

Mollá (2010) argues that there is no corpus available for the development and testing of summarisation techniques in the EBM domain. We are providing such a corpus. The corpus is sourced from the Journal of Family Practice (JFP)[7] and uses the "Clinical Inquiries" section. A key advantage of using the "Clinical Inquiries" section of JFP instead of full systematic reviews such as the Cochrane Reviews[8] is that the text in each inquiry is much more compact but it still has the links to the references in case the physician needs more information. In other words, the text looks very much like what a summariser should deliver.

For each question, the corpus contains the following information:

1. The URL of the clinical inquiry from which the information has been sourced.

2. The question, e.g. *What is the most effective treatment for tinea pedis athlete's foot?*

3. The evidence-based answer. The answer may contain several parts, since a question may be answered according to distinct pieces of evidence. For each part, the corpus includes a short description of the answer, the Strength of Recommendation (SOR) grade of the evidence related to the answer, and a short description that explains the reasoning behind allocating such a SOR grade.

4. The answer justifications. For each of the parts of the evidence-based answer there is one or more justifications describing the actual findings reported in the research papers supporting the answer.

5. The references. Each answer justification includes one or more references to the source research paper. Each reference includes the PubMed ID and the full abstract information as encoded in PubMed, if available.

4 Creation of the Corpus

The conversion of the corpus from the original text in JFP to the machine-processable form followed several steps involving automatic extraction and conversion of text, manual annotation, and crowdsourcing annotation.

4.1 Extracting Questions and Answers

The process to extract the questions and answers was relatively straightforward. We obtained permission from the publishers to download all the freely available clinical inquiries. All of the inquiries were downloaded in their original HTML format, and a Python script was used to take advantage of the relatively uniform format that marks up the questions and answers in the source. We found that the markup had changed several times (the documents date from 2001 to 2010), so we had to accommodate all changes of format. The resulting information was stored in a local database.

The question corresponds with the title of the clinical inquiry, which is formulated as a question.

The answer parts are clearly marked in the original text. Each part (called "snip" in the corpus) contains the text, SOR grade, and criteria for the SOR grade.

4.2 Annotating Answer Justifications

The answer justifications were detected automatically. However, the source text did not match each

[7] http://jfponline.com/

[8] http://www.cochrane.org/cochrane-reviews

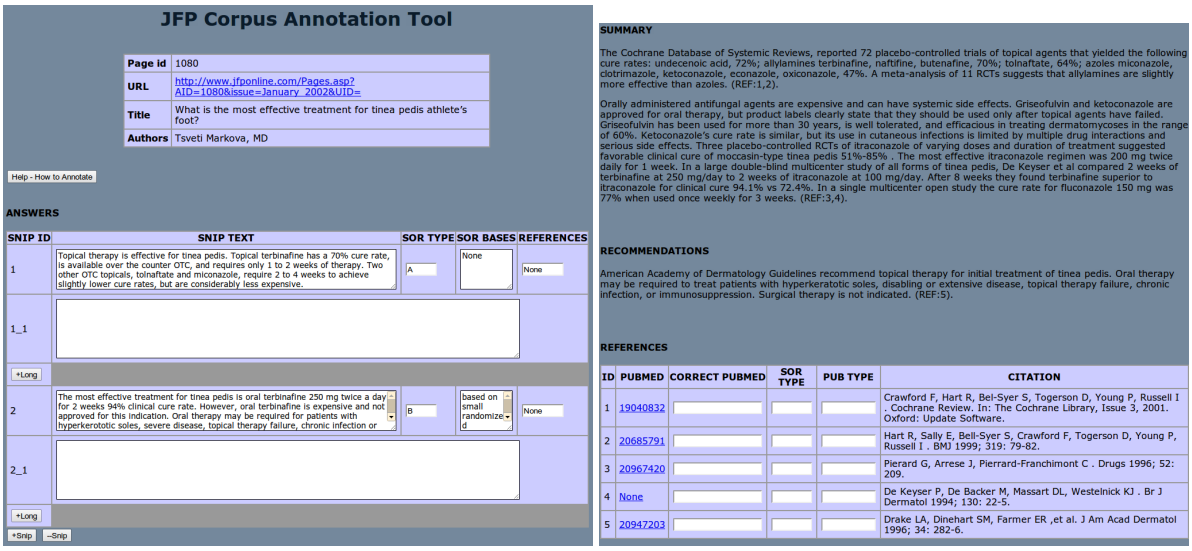

Figure 1: Screen shots of the annotation tool

justification to the specific answer snip. We therefore had to do the matching manually.

We created a web-based annotation tool that displays the question and each of the answer parts. Each answer part has associated empty slots where the annotator could copy and paste the answer justification. Figure 1 shows screen-shots of the annotation tool.

The total number of pages to annotate was distributed among three annotators. The annotators were members of the research team. A small percentage of the pages was annotated by all annotators (the annotators did not know beforehand which of the pages were annotated by all), to check for inconsistencies. The annotation process was done in several stages, with periodic checks on the common pages to detect and solve systematic inconsistencies in the annotation criteria. During those checks the annotators agreed on a set of criteria, an extract of which is:

1. Remove phrases connecting to text outside the answer justification and modify anaphora to make the text self-contained. For example, change *In another study* to *In a study* or *The second study* to *A study*.

2. Remove all general, introductory text.

3. If a justification has several references, split it into separate justifications whenever possible. In the process, some of the text may need to be copied so that each justification is self-contained.

4. If a paragraph does not have any references, check if it can be added to the previous or the next paragraph.

These criteria mostly addressed the need for each answer justification to be self-contained, and to match an answer justification to one reference only whenever possible. After inspection of a random sample of the common pages, the annotators agreed that the variations in the annotations are acceptable.

4.3 Crowdsourcing for Extracting Reference Information

Text formatting in the source text allowed the easy detection of references. To improve the usefulness of these references, we added the PubMed ID of those references found in PubMed.

We first tried to identify the PubMed ID automatically by searching on PubMed using information extracted from the reference text. The text was preprocessed by removing all the information about authors and pagination. We noted that if the authors or pagination items are present in the reference, they rarely appear in any other positions than first and last

respectively. We also noted that authors and pagination are easy to find and ignore: authors contain initials and capital case surnames; while pagination always contains numbers and punctuation such as semi-colon, colon or hyphen.

Publication names such as the names of journals and books were more difficult to detect and to normalise. We decided, instead of trying to detect them, to run a list of searches containing all combinations of remaining sentences. For example, if after removing author and pagination information there are three sentences $S1, S2, S3$, the following searches were made: $S1$-$S2$-$S3$, $S1$-$S2$, $S1$-$S3$, $S2$-$S3$, $S1$, $S2$, $S3$. These individual searches were sent to PubMed via its "Entrez Utilities" interface. The ID of the search whose returned title had the largest substring overlap with the original string was selected. As a last resort, if no searches returned an ID, a final search was made with the complete reference text.

Manual inspection of a small random sample revealed, however, that this method often did not find the correct ID. We therefore created a crowdsourcing task using Amazon Mechanical Turk.

An initial pilot experiment was made with 30 references grouped in sets ("hits") of 10 references. Each hit was allocated to three Turkers. The Turkers were asked to check the ID using PubMed, and correct it if necessary. If no ID was available, the Turkers were asked to enter "nf". We later checked the Turkers' annotations by searching PubMed using the provided IDs and found an error rate of 18% (17 out of the total of 90 were incorrect). We examined the errors and concluded that:

1. Most workers got straight to work without reading the instructions provided. For example, they typically used the ID code "0" instead of "nf" when they could not find an ID.

2. We needed an automatic (or semi-automatic) way of judging whether the workers were cheating: manual checks were too time consuming.

3. There should be a threshold for approval of work. We decided to set the threshold to 2/10 wrong annotations per page to reject cheaters.

With these findings we performed the final Mechanical Turk task. Each hit had 10 references and was sent to five Turkers. The Turkers were asked to read the instructions and were asked to do an automated test with three references. After they passed the test they were given a passcode that was required to submit the work. Each hit included two "trick" questions with known answers. The following automated tests were done on each hit:

1. Did the user answer the known references correctly?

2. Is the ID valid? A script sent each ID to PubMed and checked whether it existed.

3. Is the ID correct? The automated test checked whether the percentage of matching between the reference title and the title returned by ID was beyond a threshold of 50%.

4. Did the Turker agree with the majority? Majority was 3 or more Turkers. This test was cancelled if the ID of majority was wrong or invalid (as determined by the other tests), or in the specific case that three Turkers agreed on one ID and two Turkers agreed on another ID (we just thought that this was too a close call).

The output of the automated test was visually inspected, and those Turker jobs with two or more errors were rejected. This was done by scrolling through the errors reported by the automatic tests, finding the disputed PubMed ids, manually checking the PubMed database to decide which one is "correct" and which one is "wrong" and then changing the tags if necessary.

The final accuracy of the annotation task was manually checked on a random sample of 100 references and double-checking them. No errors were detected.

Finally, once all IDs were found, the abstracts were automatically downloaded from PubMed and added to the corpus. We chose to download the XML format, which contains useful metadata that markups the bibliography details, the abstract text, and additional annotations such as classification tags and MeSH terms.

5 Utility of the Corpus

The final statistics of the corpus are: 456 questions (called "record" in the corpus), 1,396 answer parts

(called "snip"), 3,036 answer justifications (called "long"), and 2,908 references. There is an average of 3.06 answer parts per question, 2.17 answer justifications per answer part, and 1.22 references per answer justification. There is an average of 6.57 references per question.

The distribution of SOR grades is: 345 for A, 535 for B, 330 for C, 15 for D,[9] and 171 without grade.

We envisage the use of this corpus for the following tasks:

Evidence-based summarisation. This is the main use of the corpus. It can be used to develop and test single-document summarisation by using the questions and original abstracts as the input source, and the answer justifications as the target summaries. Alternatively, it can be used to develop and test multiple-document summarisation by using the answer parts as the target summaries. Parts of the corpus have already been used for this purpose (Mollá, 2010).

Appraisal. The SOR grades can be used to test the ability to appraise the quality of the system. Appraisal can be done in the ranking component of a retrieval system, or as a separate classification task. Parts of the corpus have already been used for this purpose (Sarker et al., 2011).

Clustering. Given the natural grouping of references to form parts of the answer, the corpus can be used to develop query-focused clustering of the retrieved references.

Retrieval. The corpus references can be used as the target results of an information retrieval system. The usefulness of this corpus for assessing retrieval, however, is likely to be limited, given the findings by Dickersin et al. (1994) that between 20% and 30% of relevant literature present in MEDLINE is not present in systematic reviews.

In the remainder of this section we focus on the task of query-focused single-document summarisation, where the task is to summarise the abstract of a paper within the context of the question. The target summary is the answer justification, and the evaluation metric is ROUGE-L with stemming (Lin, 2004), a very popular metric used in the evaluation of summarisation systems.

For every answer justification/reference pair, we extracted all combinations of three sentences from the abstract and computed their ROUGE-L scores against their answer justification. With this information we computed the ROUGE-L boundary points of the document deciles. For example, the boundary points of the first decile of a document indicate the minimum and maximum values of the 10% proportion of combinations of 3-sentences with lowest ROUGE-L scores. Then we aggregated the decile boundaries of all documents to create the set of document decile boundaries according to the formula

$$Boundary[i] = \{boundary[i](x) | x \in D\}$$

where $boundary[0](x)$ is the minimum ROUGE-L score of the first decile of document x, $boundary[1](x)$ is the maximum ROUGE-L score of the first decile of document x, and so on. The resulting boxplot is shown in Figure 2. The means and standard deviations are listed in Table 1. This information shows that, in order to perform better than simple random choice of sentences, we need to obtain a ROUGE-L score of at least 0.188. For reference, a simple baseline that returns the last three sentences obtains a ROUGE-L score of 0.193, and the best system configuration that uses information of the abstract structure of those described by Mollá (2010)[10] achieves a ROUGE-L score of 0.196 when applied to our corpus. We can see that these baselines are in the range between 50% and 60% percentiles.

6 Conclusions

We have presented a corpus for the development of research in NLP in medical texts. The corpus was sourced from the Clinical Inquiries section of the Journal of Family Practice, and the process involved a set of manual and automatic methods for the extraction and annotation of information. We also describe a process of crowdsourcing that was used to find the PubMed IDs of the references.

[9] SORT has only grades A, B, and C, but apparently some authors used one more level D to indicate very poor evidence.

[10] This is the system configuration that uses abstract structure but does not use question information.

Boundary	0	1	2	3	4	5	6	7	8	9	10
Mean	0.094	0.136	0.153	0.164	0.176	0.188	0.200	0.213	0.229	0.249	0.299
Std Dev	0.060	0.062	0.065	0.067	0.070	0.073	0.076	0.081	0.087	0.094	0.112

Table 1: Statistics of the decile boundaries of ROUGE-L data

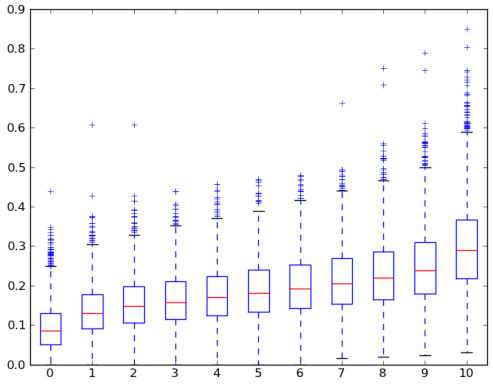

Figure 2: ROUGE-L boxplots for all decile boundaries

The emphasis of this corpus is the development and testing of query-focused multi-document summarisation systems for Evidence Based Medicine, but we envisage its use in other tasks such as text classification, and clustering.

We have shown a set of statistics of the ROUGE-L scores of the abstracts within the context of document summarisation. The data show that current baselines do not perform much better than simple random choice and there is still much room for improvement. The challenge is up for researchers to take.

Further work includes the use of this corpus for some of the tasks described above. We are also studying the possibility of including additional annotation of the specific abstract sentences that are found to be most relevant to the answer justifications. This information could be used to perform pyramidal-style evaluation such as the one described by Dang and Lin (2007).

References

Stergos Afantenos, Vangelis Karkaletsis, and Panagiotis Stamatopoulos. 2005. Summarization from medical documents: a survey. *Artificial Intelligence in Medicine*, 33(2):157–177, February. PMID: 15811783.

I. Elaine Allen and Ingram Olkin. 1999. Estimating time to conduct a meta-analysis from number of citations retrieved. *JAMA: The Journal of the American Medical Association*, 282(7):634–635, August. PMID: 10517715.

E. C. Armstrong. 1999. The well-built clinical question: the key to finding the best evidence efficiently. *WMJ*, 98(2):25–28.

Lyle Berkowitz. 2002. Review and evaluation of internet-based clinical reference tools for physicians. Technical report, UpToDate.

Hoa Dang and Jimmy Lin. 2007. Different structures for evaluating answers to complex questions: Pyramids won't topple, and neither will human assessors. In *Proceedings ACL*.

Dina Demner-Fushman and Jimmy Lin. 2006. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings ACL*. The Association for Computer Linguistics.

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

Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*. Online uncorrected proof.

K. Dickersin, R. Scherer, and C. Lefebvre. 1994. Identifying relevant studies for systematic reviews. *BMJ (Clinical Research Ed.)*, 309(6964):1286–1291, November. PMID: 7718048.

Mark H. Ebell, Jay Siwek, Barry D. Weiss, Steven H. Woolf, Jeffrey Susman, Bernard Ewigman, and Marjorie Bowman. 2004. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *Am Fam Physician*, 69(3):548–556, Feb.

N. Elhadad, M.-Y. Kan, J. L. Klavans, and K. R. McKeown. 2005. Customization in a unified framework for summarizing medical literature. *Artificial Intelligence in Medicine*, 33(2):179–198, February. PMID: 15811784.

John W. Ely, Jerome A. Osheroff, Mark H. Ebell, George R. Bergus, Barcey T. Levy, M. Lee Chambliss, and Eric R. Evans. 1999. Analysis of questions asked by family doctors regarding patient care. *BMJ*, 319(7206):358–361, Aug.

John Ely, Jerome A Osheroff, Mark H Ebell, M. Lee Chambliss, DC Vinson, James J. Stevermer, and Eric A. Pifer. 2002. Obstacles to answering doctors' questions about patient care with evidence: Qualitative study. *BMJ*, 324(7339):710.

John W. Ely, Jerome A. Osheroff, M. Lee Chambliss, Mark H Ebell, and Marcy E. Rosenbaum. 2005. Answering physicians' clinical questions: Obstacles and potential solutions. *J Am Med Inform Assoc.*, 12(2):217–224.

Marcelo Fiszman, Thomas C. Rindflesch, and Halil Kilicoglu. 2004. Abstraction summarization for managing the biomedical research literature. In *Procs. HLT-NAACL Workshop on Computational Lexical Semantics*, pages 76–83.

T. Goetz and C.-W. von der Lieth. 2005. PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts. *Nucleic Acids Research*, 33(Web Server):W774–W778.

R. Brian Haynes, Nancy L. Wilczynski, K. Ann McKibbon, Cynthia J. Walker, and John C. Sinclair. 1994. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *Journal of the American Medical Informatics Association: JAMIA*, 1(6):447–458, December. PMID: 7850570.

R. Brian Haynes, K. Ann McKibbon, Nancy L. Wilczynski, Stephen D. Walter, and Stephen R. Werre. 2005. Optimal search strategies for retrieving scientifically strong studies of treatment from medline: analytical survey. *BMJ (Clinical Research Ed.)*, 330(7501):1179, May. PMID: 15894554.

Sarvnaz Karimi, Justin Zobel, Stefan Pohl, and Falk Scholer. 2009. The challenge of high recall in biomedical systematic search. In *Proc. DTMBIO*, pages 89–92, Honk Kong.

Minsuk Lee, James Cimino, Hai Ran Zhu, Carl Sable, Vijay Shanker, John Ely, and Hong Yu. 2006. Beyond information retrieval — medical question answering. In *Proc. AMIA 2006*.

Annette Leonhard. 2009. Towards retrieving relevant information for answering clinical comparison questions. In *Proceddings BioNLP 2009*, pages 153–161.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10.

Diego Mollá. 2010. A corpus for evidence based medicine summarisation. In *Proceedings of the Australasian Language Technology Workshop*, volume 8, pages 76–80.

Yun Niu, Graeme Hirst, Gregory McArthur, and Patricia Rodriguez-Gianolli. 2003. Answering clinical questions with role identification. In *Proc. ACL, Workshop on Natural Language Processing in Biomedicine*.

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

David L. Sackett, William M. Rosenberg, Jamuir Gray, R. Brian Haynes, and W. Scott Richardson. 1996. Evidence based medicine: What it is and what it isn't. *BMJ*, 312(7023):71–72.

David L. Sackett, Sharon E. Straus, W. Scott Richardson, William Rosenberg, and R. Brian Haynes. 2000. *Evidence-Based Medicine: How to Practice and Teach EBM*. Churchill Livingstone, 2 edition.

Abeed Sarker, Diego Mollá, and Cécile Paris. 2011. Towards automatic grading of evidence. In *Proceedings of the Third International Workshop on Health Document Text Mining and Information Analysis (LOUHI 2011)*, pages 51–58, Bled, Slovenia.

Kaveh G. Shojania and Lisa A. Bero. 2001. Taking advantage of the explosion of systematic reviews: an efficient MEDLINE search strategy. *Effective Clinical Practice: ECP*, 4(4):157–162, August. PMID: 11525102.

Thanh Tang, David Hawking, Ramesh Sankaranarayana, Kathleen M. Griffiths, and Nick Craswell. 2009. Quality-oriented search for depression portals. In *ECIR '09 Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, Berlin, Heidelberg. Springer.

Andreea Tutos and Diego Mollá. 2010. A study on the use of search engines for answering clinical questions. In *Proceedings HIKM 2010*.

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

Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen. 2007. Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, 8(5):358–375.

Pierre Zweigenbaum. 2003. Question answering in biomedicine. In *Proc. EACL2003, workshop on NLP for Question Answering*, Budapest.