Situated Question Answering in the Clinical Domain: Selecting the Best Drug Treatment for Diseases

Dina Demner-Fushman[1,3] **and Jimmy Lin**[1,2,3]
[1]Department of Computer Science
[2]College of Information Studies
[3]Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA
demner@cs.umd.edu, jimmylin@umd.edu

Abstract

Unlike open-domain factoid questions, clinical information needs arise within the rich context of patient treatment. This environment establishes a number of constraints on the design of systems aimed at physicians in real-world settings. In this paper, we describe a clinical question answering system that focuses on a class of commonly-occurring questions: "What is the best drug treatment for *X*?", where *X* can be any disease. To evaluate our system, we built a test collection consisting of thirty randomly-selected diseases from an existing secondary source. Both an automatic and a manual evaluation demonstrate that our system compares favorably to PubMed, the search system most commonly-used by physicians today.

1 Introduction

Over the past several years, question answering (QA) has emerged as a general framework for addressing users' information needs. Instead of returning "hits", as information retrieval systems do, QA systems respond to natural language questions with concise, targeted information. Recently, research focus has shifted away from so-called factoid questions such as "What are pennies made of?" and "What country is Aswan High Dam located in?" to more complex questions such as "How have South American drug cartels been using banks in Liechtenstein to launder money?" and "What was the Pentagon panel's position with respect to the dispute over the US Navy training range on the island of Vieques?"—so-called "relationship" and "opinion" questions, respectively.

These complex information needs differ from factoid questions in many important ways. Unlike factoids, they cannot be answered by named-entities and other short noun phrases. They do not occur in isolation, but are rather embedded within a broader context, i.e., a "scenario". These complex questions set forth parameters of the desired knowledge, which may include additional facts about the motivation of the information seeker, her assumptions, her current state of knowledge, etc. Presently, most systems that attempt to tackle such complex questions are aimed at serving intelligence analysts, for activities such as counter-terrorism and war-fighting.

Systems for addressing complex information needs are interesting because they provide an opportunity to explore the role of semantic structures in question answering, e.g., (Narayanan and Harabagiu, 2004). Opportunities include explicit semantic representations for capturing the content of questions and documents, deep inferential mechanisms (Moldovan et al., 2002), and attempts to model task-specific influences in information-seeking environments (Freund et al., 2005).

Our own interest in question answering falls in line with these recent developments, but we focus on a different type of user—the primary care physician. The need to answer questions related to patient care at the point of service has been well studied and documented (Gorman et al., 1994; Ely et al., 1999; Ely et al., 2005). However, research has shown that existing search systems, e.g., PubMed, are often unable to supply clinically-relevant answers in a timely manner (Gorman et al., 1994; Chambliss and Conley, 1996). Clinical question answering represents a high-impact application that has the potential to improve the quality of medical care.

Proceedings of the Workshop on Task-Focused Summarization and Question Answering, pages 24–31,
Sydney, July 2006.

From a research perspective, the clinical domain is attractive because substantial medical knowledge has already been codified in the Unified Medical Language System (UMLS) (Lindberg et al., 1993). This large ontology enables us to explore knowledge-rich techniques and move beyond question answering methods primarily driven by keyword matching. In this work, we describe a paradigm of medical practice known as evidence-based medicine and explain how it can be computationally captured in a semantic domain model. Two separate evaluations demonstrate that semantic modeling yields gains in question answering performance.

2 Considerations for Clinical QA

We begin our exploration of clinical question answering by first discussing design constraints imposed by the domain and the information-seeking environment. The practice of evidence-based medicine (EBM) provides a well-defined process model for situating our system. EBM is a widely-accepted paradigm for medical practice that involves the explicit use of current best evidence, i.e., high-quality patient-centered clinical research reported in the primary medical literature, to make decisions about patient care. As shown by previous work (De Groote and Dorsch, 2003), citations from the MEDLINE database maintained by the National Library of Medicine serve as a good source of evidence.

Thus, we conceive of clinical question answering systems as fulfilling a decision-support role by retrieving highly-relevant MEDLINE abstracts in response to a clinical question. This represents a departure from previous systems, which focus on extracting short text segments from larger sources. The implications of making potentially life-altering decisions mean that all evidence must be carefully examined in context. For example, the efficacy of a drug in treating a disease is always framed in the context of a specific study on a sample population, over a set duration, at some fixed dosage, etc. The physician simply cannot recommend a particular course of action without considering all these complex factors. Thus, an "answer" without adequate support is not useful. Given that a MEDLINE abstract—on the order of 250 words, equivalent to a long paragraph—generally encapsulates the context of a clinical study, it serves as a logical answer unit and an entry point to the information necessary to answer the physician's question (e.g., via drill-down to full text articles).

In order for a clinical QA system to be successful, it must be suitably integrated into the daily activities of a physician. Within a clinic or a hospital setting, the traditional desktop application is not the most ideal interface for a retrieval system. In most cases, decisions about patient care must be made by the bedside. Thus, a PDA is an ideal vehicle for delivering question answering capabilities (Hauser et al., 2004). However, the form factor and small screen size of such devices places constraints on system design. In particular, since the physician is unable to view large amounts of text, precision is of utmost importance.

In summary, this section outlines considerations for question answering in the clinical domain: the necessity of contextualized answers, the rationale for adopting MEDLINE abstract as the response unit, and the importance of high precision.

3 EBM and Clinical QA

Evidence-based medicine not only supplies a process model for situating question answering capabilities, but also provides a framework for codifying the knowledge involved in retrieving answers. This section describes how the EBM paradigm provides the basis of the semantic domain model for our question answering system.

Evidence-based medicine offers three facets of the clinical domain, that, when taken together, describe a model for addressing complex clinical information needs. The first facet, shown in Table 1 (left column), describes the four main tasks that physicians engage in. The second facet pertains to the structure of a well-built clinical question. Richardson et al. (1995) identify four key elements, as shown in Table 1 (middle column). These four elements are often referenced with a mnemonic PICO, which stands for Patient/Problem, Intervention, Comparison, and Outcome. Finally, the third facet serves as a tool for appraising the strength of evidence, i.e., how much confidence should a physician have in the results? For this work, we adopted a system with three levels of recommendations, as shown in Table 1 (right column).

By integrating these three perspectives of evidence-based medicine, we conceptualize clinical question answering as "semantic unification" between information needs expressed in a

Clinical Tasks	PICO Elements	Strength of Evidence
Therapy: Selecting effective treatments for patients, taking into account other factors such as risk and cost. **Diagnosis:** Selecting and interpreting diagnostic tests, while considering their precision, accuracy, acceptability, cost, and safety. **Prognosis:** Estimating the patient's likely course with time and anticipating likely complications. **Etiology:** Identifying the causes for a patient's disease.	**Patient/Problem:** What is the primary problem or disease? What are the characteristics of the patient (e.g., age, gender, co-existing conditions, etc.)? **Intervention:** What is the main intervention (e.g., diagnostic test, medication, therapeutic procedure, etc.)? **Comparison:** What is the main intervention compared to (e.g., no intervention, another drug, another therapeutic procedure, a placebo, etc.)? **Outcome:** What is the effect of the intervention (e.g., symptoms relieved or eliminated, cost reduced, etc.)?	**A-level evidence** is based on consistent, good quality patient-oriented evidence presented in systematic reviews, randomized controlled clinical trials, cohort studies, and meta-analyses. **B-level evidence** is inconsistent, limited quality patient-oriented evidence in the same types of studies. **C-level evidence** is based on disease-oriented evidence or studies less rigorous than randomized controlled clinical trials, cohort studies, systematic reviews and meta-analyses.

Table 1: The three facets of evidence-based medicine.

PICO-based knowledge structure and corresponding structures extracted from MEDLINE abstracts. Naturally, this matching process should be sensitive to the clinical task and the strength of evidence of the retrieved abstracts. As conceived, clinical question answering is a knowledge-intensive endeavor that requires automatic identification of PICO elements from MEDLINE abstracts.

Ideally, a clinical question answering system should be capable of directly performing this semantic match on abstracts, but the size of the MEDLINE database (over 16 million citations) makes this approach currently unfeasible. As an alternative, we rely on PubMed,[1] a boolean search engine provided by the National Library of Medicine, to retrieve an initial set of results that we then postprocess in greater detail—this is the standard two-stage architecture commonly-employed by many question answering systems (Hirschman and Gaizauskas, 2001).

The complete architecture of our system is shown in Figure 1. The query formulation module converts the clinical question into a PubMed search query, identifies the clinical task, and extracts the appropriate PICO elements. PubMed returns an initial list of MEDLINE citations, which is analyzed by the knowledge extractor to identify clinically-relevant elements. These elements serve as input to the semantic matcher, and are compared to corresponding elements extracted from the question. Citations are then scored and the top ranking ones are returned as answers.

[1] http://www.ncbi.nih.gov/entrez/

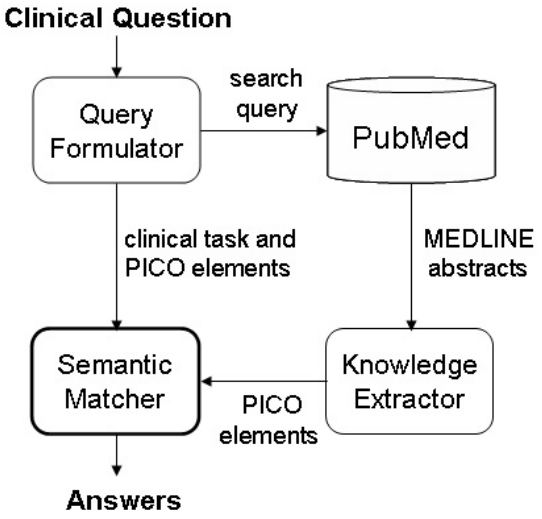

Figure 1: Architecture of our clinical question answering system.

Although we have outlined a general framework for clinical question answering, the space of all possible patient care questions is immense, and attempts to develop a comprehensive system is beyond the scope of this paper. Instead, we focus on a subset of therapy questions: specifically, questions of the form "What is the best drug treatment for *X*?", where *X* can be any disease. We have chosen to tackle this class of questions because studies of physicians' question-asking behavior in natural settings have revealed that this question type occurs frequently (Ely et al., 1999). By leveraging the natural distribution of clinical questions, we can make the greatest impact with the least amount

of development effort. For this class of questions, we have implemented a working system with the architecture described in Figure 1. The next three sections detail each module.

4 Query Formulator

Since our system only handles one question type, the query formulator is relatively simple: the task is known in advance to be therapy and the Problem PICO element is the disease asked about in the clinical question. In order to facilitate the semantic matching process, we employ MetaMap (Aronson, 2001) to identify the concept in the UMLS ontology that corresponds to the disease; UMLS also provides alternative names and other expansions.

The query formulator also generates a query to PubMed, the National Library of Medicine's boolean search engine for MEDLINE. As an example, the following query is issued to retrieve hits for the disease "meningitis":

> (Meningitis[mh:noexp]) AND drug therapy[sh] AND hasabstract[text] AND Clinical Trial[pt] AND English[Lang] AND humans[mh] AND (1900[PDAT] : 2003/03[PDAT])

In order to get the best possible set of initial citations, we employ MeSH (Medical Subject Headings) terms when available. MeSH terms are controlled vocabulary concepts assigned manually by trained medical librarians in the indexing process (based on the full text of the article), and encode a substantial amount of knowledge about the contents of the citation. PubMed allows searches on MeSH headings, which usually yield highly accurate results. In addition, we limit retrieved citations to those that have the MeSH heading "drug therapy"and those that describe a clinical trial (another metadata field). By default, PubMed orders citations chronologically in reverse.

5 Knowledge Extractor

The knowledge extraction module provides the basic frame elements used in the semantic matching process, described in the next section. We employ previously-implemented components (Demner-Fushman and Lin, 2005) that identify PICO elements within a MEDLINE citation using a combination of knowledge-based and statistical machine-learning techniques. Of the four PICO elements prescribed by evidence-based medicine practitioners, only the Problem and Outcome elements are relevant for this application (there are no Interventions and Comparisons for our question type). The Problem is the main disease under consideration in an abstract, and outcomes are statements that assert clinical findings, e.g., efficacy of a drug or a comparison between two drugs. The ability to precisely identify these clinically-relevant elements provides the foundation for semantic question answering capabilities.

6 Semantic Matcher

Evidence-based medicine identifies three different sets of factors that must be taken into account when assessing citation relevance. These considerations are computationally operationalized in the semantic matcher, which takes as input elements identified by the knowledge extractor and scores the relevance of each PubMed citation with respect to the question. After matching, the top-scoring abstracts are presented to the physician as answers. The individual score of a citation is comprised of three components:

$$S_{\text{EBM}} = S_{\text{PICO}} + S_{\text{SoE}} + S_{\text{MeSH}} \quad (1)$$

By codifying the principles of evidence-based medicine, our semantic matcher attempts to satisfy information needs through conceptual analysis, as opposed to simple keyword matching. In the following subsections, we describe each of these components in detail.

6.1 PICO Matching

The score of an abstract based on PICO elements, S_{PICO}, is broken up into two separate scores:

$$S_{\text{PICO}} = S_{\text{problem}} + S_{\text{outcome}} \quad (2)$$

The first component in the above equation, S_{problem}, reflects a match between the primary problem in the query frame and the primary problem identified in the abstract. A score of 1 is given if the problems match exactly, based on their unique UMLS concept id (as provided by MetaMap). Matching based on concept ids addresses the issue of terminological variation. Failing an exact match of concept ids, a partial string match is given a score of 0.5. If the primary problem in the query has no overlap with the primary problem from the abstract, a score of -1 is given.

The outcome-based score S_{outcome} is the value assigned to the highest-scoring outcome sentence,

as determined by the knowledge extractor. Since the desired outcome (i.e., improve the patient's condition) is implicit in the clinical question, our system only considers the inherent quality of outcome statements in the abstract. Given a match on the primary problem, most clinical outcomes are likely to be of interest to the physician.

For the drug treatment scenario, there is no intervention or comparison, and so these elements do not contribute to the semantic matching.

6.2 Strength of Evidence

The relevance score of a citation based on the strength of evidence is calculated as follows:

$$S_{\text{SoE}} = S_{\text{journal}} + S_{\text{study}} + S_{\text{date}} \quad (3)$$

Citations published in core and high-impact journals such as Journal of the American Medical Association (JAMA) get a score of 0.6 for S_{journal}, and 0 otherwise. In terms of the study type, S_{study}, clinical trials receive a score of 0.5; observational studies, 0.3; all non-clinical publications, -1.5; and 0 otherwise. The study type is directly encoded as metadata in a MEDLINE citation.

Finally, recency factors into the strength of evidence score according to the formula below:

$$S_{\text{date}} = (year_{\text{publication}} - year_{\text{current}})/100 \quad (4)$$

A mild penalty decreases the score of a citation proportionally to the time difference between the date of the search and the date of publication.

6.3 MeSH Matching

The final component of the EBM score reflects task-specific considerations, and is computed from MeSH terms associated with each citation:

$$S_{\text{MeSH}} = \sum_{t \in \text{MeSH}} \alpha(t) \quad (5)$$

The function $\alpha(t)$ maps MeSH terms to positive scores for positive indicators, negative scores for negative indicators, or zero otherwise.

Negative indicators include MeSH headings associated with genomics, such as "genetics" and "cell physiology". Positive indicators for therapy were derived from the clinical query filters used in PubMed searches (Haynes et al., 1994); examples include "drug administration routes" and any of its children in the MeSH hierarchy. A score of ± 1 is given if the MeSH descriptor or qualifier is marked as the main theme of the article (indicated via the star notation by indexers), and ± 0.5 otherwise.

7 Evaluation Methodology

Clinical Evidence (CE) is a periodic report created by the British Medical Journal (BMJ) Publishing Group that summarizes the best treatments for a few dozen diseases at the time of publication. We were able to mine the June 2004 edition to create a test collection to evaluate our system. Note that the existence of such secondary sources does not obviate the need for clinical question answering because they are perpetually falling out of date due to rapid advances in medicine. Furthermore, such reports are currently created by highly-experienced physicians, which is an expensive and time-consuming process. From *CE*, we randomly extracted thirty diseases, creating a development set of five questions and a test set of twenty-five questions. Some examples include: acute asthma, chronic prostatitis, community acquired pneumonia, and erectile dysfunction.

We conducted two evaluations—one automatic and one manual—that compare the original PubMed hits and the output of our semantic matcher. The first evaluation is based on ROUGE, a commonly-used summarization metric that computes the unigram overlap between a particular text and one or more reference texts.[2] The treatment overview for each disease in *CE* is accompanied by a number of citations (used in writing the overview itself)—the abstract texts of these cited articles serve as our references. We adopt this approach because medical journals require abstracts that provide factual information summarizing the main points of the studies. We assume that the closer an abstract is to these reference abstracts (as measured by ROUGE-1 precision), the more relevant it is. On average, each disease overview contains 48.4 citations; however, we were only able to gather abstracts of those that were contained in MEDLINE (34.7 citations per disease, min 8, max 100). For evaluation purposes, we restricted abstracts under consideration to those that were published before our edition of *CE*. To quantify the performance of our system, we computed the average ROUGE score over the top one, three, five, and ten hits of our EBM and baseline systems.

To supplement our automatic evaluation, we also conducted a double-blind manual evaluation

[2] We ran ROUGE-1.5.5 with DUC 2005 settings.

	PubMed	EBM	PICO	SoE	MeSH
1	0.160	0.205 (+27.7%)$^{\triangle}$	0.186 (+16.1%)$^{\circ}$	0.192 (+20.0%)$^{\circ}$	0.166 (+3.6%)$^{\circ}$
3	0.162	0.202 (+24.6%)$^{\blacktriangle}$	0.192 (+18.0%)$^{\blacktriangle}$	0.204 (+25.5%)$^{\blacktriangle}$	0.172 (+6.1%)$^{\circ}$
5	0.166	0.198 (+19.5%)$^{\blacktriangle}$	0.196 (+18.0%)$^{\blacktriangle}$	0.201 (+21.3%)$^{\blacktriangle}$	0.168 (+1.2%)$^{\circ}$
10	0.170	0.196 (+15.5%)$^{\blacktriangle}$	0.191 (+12.5%)$^{\blacktriangle}$	0.195 (+15.1%)$^{\blacktriangle}$	0.174 (+2.8%)$^{\circ}$

Table 2: Results of automatic evaluation: average ROUGE score using cited abstracts in *CE* as references. The EBM column represents performance of our complete domain model. PICO, SoE, and MeSH represent performance of each component. ($^{\circ}$ denotes n.s., $^{\triangle}$ denotes sig. at 0.95, $^{\blacktriangle}$ denotes sig. at 0.99)

PubMed results	**EBM-reranked results**
Effect of vitamin A supplementation on childhood morbidity and mortality.	A comparison of ceftriaxone and cefuroxime for the treatment of bacterial meningitis in children.
Intrathecal chemotherapy in carcinomatous meningitis from breast cancer.	Randomised comparison of chloramphenicol, ampicillin, cefotaxime, and ceftriaxone for childhood bacterial meningitis.
Isolated leptomeningeal carcinomatosis (carcinomatous meningitis) after taxane-induced major remission in patients with advanced breast cancer.	The beneficial effects of early dexamethasone administration in infants and children with bacterial meningitis.

Table 3: Titles of the top abstracts retrieved in response to the question "What is the best treatment for meningitis?", before and after applying our semantic reranking algorithm.

of the system. The top five citations from both the original PubMed results and the output of our semantic matcher were gathered, blinded, and randomized (see Table 3 for an example of top results obtained by PubMed and our system). The first author of this paper, who is a medical doctor, manually evaluated the abstracts. Since the sources of the abstracts were hidden, judgments were guaranteed to be impartial. All abstracts were evaluated on a four point scale: not relevant, marginally relevant, relevant, and highly relevant, which corresponds to a score of zero to three.

8 Results

The results of our automatic evaluation are shown in Table 2: the rows show average ROUGE scores at one, three, five, and ten hits, respectively. In addition to the PubMed baseline and our complete EBM model, we conducted a component-level analysis of our semantic matching algorithm. Three separate ablation studies isolate the effects of the PICO-based score, the strength of evidence score, and the MeSH-based score (columns "PICO", "SoE", and "MeSH").

At all document cutoffs, the quality of the EBM-reranked hits is higher than that of the original PubMed hits, as measured by ROUGE. The differences are statistically significant, according to the Wilcoxon signed-rank test, the standard non-parametric test employed in IR.

Based on the component analysis, we can see that the strength of evidence score is responsible for the largest performance gain, although the combination of all three components outperforms each one individually (for the most part). All three components of our semantic model contribute to the overall QA performance, which is expected because clinical relevance is a multi-faceted property that requires a multitude of considerations. Evidence-based medicine provides a theory of these factors, and we have shown that a question answering algorithm which operationalizes EBM yields good results.

The distribution of human judgments from our manual evaluation is shown in Figure 2. For the development set, the average human judgment of the original PubMed hits is 1.52 (between "marginally relevant" and "relevant"); after semantic matching, 2.32 (better than "relevant"). For the test set, the averages are 1.49 before ranking and 2.10 after semantic matching. These results show that our system performs significantly better than the PubMed baseline.

The performance improvement observed in our experiments is encouraging, considering that we were starting off with a strong state-of-the-art

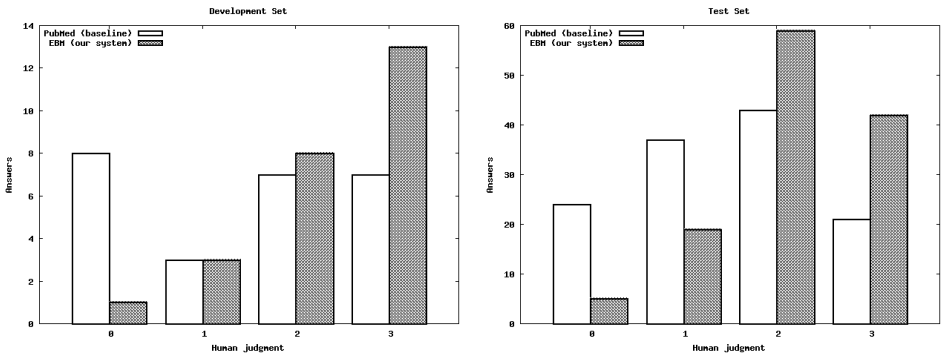

Figure 2: Results of our manual evaluation: distribution of judgments, for development set (left) and test set (right). (0=not relevant, 1=marginally relevant, 2=relevant, 3=highly relevant)

PubMed baseline that leverages MeSH terms. All initial citations retrieved by PubMed were clinical trials and "about" the disease in question, as determined by human indexers. Our work demonstrates that principles of evidence-based medicine can be codified in an algorithm.

Since a number of abstracts were both automatically evaluated with ROUGE and manually assessed, it is possible to determine the degree to which automatic metrics predict human judgments. For the 125 human judgments gathered on the test set, we computed a Pearson's *r* score of 0.544, which indicates moderate predictiveness. Due to the structure of our PubMed query, the keyword content of retrieved abstracts are relatively homogeneous. Nevertheless, automatic evaluation with ROUGE appears to be useful.

9 Discussion and Related Work

Recently, researchers have become interested in restricted-domain question answering because it provides an opportunity to explore the use of knowledge-rich techniques without having to tackle the commonsense reasoning problem. Knowledge-based techniques dependent on rich semantic representations contrast with TREC-style factoid question answering, which is primarily driven by keyword matching and named-entity detection.

Our work represents a successful case study of how semantic models can be employed to capture domain knowledge (the practice of medicine, in our case). The conception of question answering as the matching of knowledge frames provides us with an opportunity to experiment with semantic representations that capture the content of both documents and information needs. In our case, PICO-based scores were found to have a positive impact on performance. The strength of evidence and the MeSH-based scores represent attempts to model user requirements by leveraging meta-level information not directly present in either questions or candidate answers. Both contribute positively to performance. Overall, the construction of our semantic model is enabled by the UMLS ontology, which provides an enumeration of relevant concepts (e.g., the names of diseases, drugs, etc.) and semantic relations between those concepts.

Question answering in the clinical domain is an emerging area of research that has only recently begun to receive serious attention. As a result, there exist relatively few points of comparison to our own work, as the research space is sparsely populated.

The idea that information systems should be sensitive to the practice of evidence-based medicine is not new. Many researchers have studied MeSH terms associated with basic clinical tasks (Mendonça and Cimino, 2001; Haynes et al., 1994). Although originally developed as a tool to assist in query formulation, Booth (2000) pointed out that PICO frames can be employed to structure IR results for improving precision; PICO-based querying is merely an instance of faceted querying, which has been widely used by librarians since the invention of automated retrieval systems. The feasibility of automatically identifying outcome statements in secondary sources has been demonstrated by Niu and Hirst (2004), but our work differs in its focus on the primary medical literature. Approaching clinical needs from a different perspective, the PERSIVAL system leverages patient records to rerank search results (McKeown et al., 2003). Since the primary focus is on person-

alization, this work can be viewed as complementary to our own.

The dearth of related work and the lack of a pre-existing clinical test collection to a large extent explains the *ad hoc* nature of some aspects of our semantic matching algorithm. All weights were heuristically chosen to reflect our understanding of the domain, and were not optimized in a principled manner. Nevertheless, performance gains observed in the development set carried over to the blind held-out test collection, providing confidence in the generality of our methods. Developing a more formal scoring model for evidence-based medicine will be the subject of future work.

10 Conclusion

We see this work as having two separate contributions. From the viewpoint of computational linguistics, we have demonstrated the effectiveness of a knowledge-rich approach to QA based on matching questions with answers at the semantic level. From the viewpoint of medical informatics, we have shown how principles of evidence-based medicine can be operationalized in a system to support physicians. We hope that this work paves the way for future high-impact applications.

11 Acknowledgments

This work was supported in part by the National Library of Medicine. The second author wishes to thank Esther and Kiri for their loving support.

References

A. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proceeding of the AMIA 2001*.

A. Booth. 2000. Formulating the question. In A. Booth and G. Walton, editors, *Managing Knowledge in Health Services*. Facet Publishing.

M. Chambliss and J. Conley. 1996. Answering clinical questions. *The Journal of Family Practice*, 43:140–144.

S. De Groote and J. Dorsch. 2003. Measuring use patterns of online journals and databases. *Journal of the Medical Library Association*, 91(2):231–240, April.

D. Demner-Fushman and J. Lin. 2005. Knowledge extraction for clinical question answering: Preliminary results. In *Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains*.

J. Ely, J. Osheroff, M. Ebell, G. Bergus, B. Levy, M. Chambliss, and E. Evans. 1999. Analysis of questions asked by family doctors regarding patient care. *BMJ*, 319:358–361.

J. Ely, J. Osheroff, M. Chambliss, M. Ebell, and M. Rosenbaum. 2005. Answering physicians' clinical questions: Obstacles and potential solutions. *Journal of the American Medical Informatics Association*, 12(2):217–224, March-April.

L. Freund, E. Toms, and C. Clarke. 2005. Modeling task-genre relationships for IR in the Workplace. In *Proceedings of SIGIR 2005*.

P. Gorman, J. Ash, and L. Wykoff. 1994. Can primary care physicians' questions be answered using the medical journal literature? *Bulletin of the Medical Library Association*, 82(2):140–146, April.

S. Hauser, D. Demner-Fushman, G. Ford, and G. Thoma. 2004. PubMed on Tap: Discovering design principles for online information delivery to handheld computers. In *Proceedings of MEDINFO 2004*.

R. Haynes, N. Wilczynski, K. McKibbon, C. Walker, and J. Sinclair. 1994. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *Journal of the American Medical Informatics Association*, 1(6):447–458.

L. Hirschman and R. Gaizauskas. 2001. Natural language question answering: The view from here. *Natural Language Engineering*, 7(4):275–300.

D. Lindberg, B. Humphreys, and A. McCray. 1993. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291, August.

K. McKeown, N. Elhadad, and V. Hatzivassiloglou. 2003. Leveraging a common representation for personalized search and summarization in a medical digital library. In *Proceedings JCDL 2003*.

E. Mendonça and J. Cimino. 2001. Building a knowledge base to support a digital library. In *Proceedings of MEDINFO 2001*.

D. Moldovan, M. Paşca, S. Harabagiu, and M. Surdeanu. 2002. Performance issues and error analysis in an open-domain question answering system. In *Proceedings of ACL 2002*.

S. Narayanan and S. Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of COLING 2004*.

Y. Niu and G. Hirst. 2004. Analysis of semantic classes in medical text for question answering. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*.

W. Richardson, M. Wilson, J. Nishikawa, and R. Hayward. 1995. The well-built clinical question: A key to evidence-based decisions. *American College of Physicians Journal Club*, 123(3):A12–A13, November-December.