Answer Extraction, Semantic Clustering, and Extractive Summarization for Clinical Question Answering

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

This paper presents a hybrid approach to question answering in the clinical domain that combines techniques from summarization and information retrieval. We tackle a frequently-occurring class of questions that takes the form "What is the best drug treatment for *X*?" Starting from an initial set of MEDLINE citations, our system first identifies the drugs under study. Abstracts are then clustered using semantic classes from the UMLS ontology. Finally, a short extractive summary is generated for each abstract to populate the clusters. Two evaluations—a manual one focused on short answers and an automatic one focused on the supporting abstracts—demonstrate that our system compares favorably to PubMed, the search system most widely used by physicians today.

1 Introduction

Complex information needs can rarely be addressed by single documents, but rather require the integration of knowledge from multiple sources. This suggests that modern information retrieval systems, which excel at producing ranked lists of documents sorted by relevance, may not be sufficient to provide users with a good overview of the "information landscape".

Current question answering systems aspire to address this shortcoming by gathering relevant "facts" from multiple documents in response to information needs. The so-called "definition" or "other" questions at recent TREC evaluations (Voorhees, 2005) serve as good examples: "good answers" to these questions include interesting "nuggets" about a particular person, organization, entity, or event.

The importance of cross-document information synthesis has not escaped the attention of other researchers. The last few years have seen a convergence between the question answering and summarization communities (Amigó et al., 2004), as highlighted by the shift from generic to query-focused summaries in the 2005 DUC evaluation (Dang, 2005). Despite a focus on document ranking, different techniques for organizing search results have been explored by information retrieval researchers, as exemplified by techniques based on clustering (Hearst and Pedersen, 1996; Dumais et al., 2001; Lawrie and Croft, 2003).

Our work, which is situated in the domain of clinical medicine, lies at the intersection of question answering, information retrieval, and summarization. We employ answer extraction to identify short answers, semantic clustering to group similar results, and extractive summarization to produce supporting evidence. This paper describes how each of these capabilities contributes to an information system tailored to the requirements of physicians. Two separate evaluations demonstrate the effectiveness of our approach.

2 Clinical Information Needs

Although the need to answer questions related to patient care has been well documented (Covell et al., 1985; Gorman et al., 1994; Ely et al., 1999), studies have shown that existing search systems, e.g., PubMed, the U.S. National Library of Medicine's search engine, are often unable to supply physicians with clinically-relevant answers in a timely manner (Gorman et al., 1994; Chambliss and Conley, 1996). Clinical information

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 841–848,
Sydney, July 2006.

Disease: Chronic Prostatitis

► **anti-microbial**

1. [temafloxacin] Treatment of chronic bacterial prostatitis with temafloxacin. Temafloxacin 400 mg b.i.d. administered orally for 28 days represents a safe and effective treatment for chronic bacterial prostatitis.

2. [ofloxacin] Ofloxacin in the management of complicated urinary tract infections, including prostatitis. In chronic bacterial prostatitis, results to date suggest that ofloxacin may be more effective clinically and as effective microbiologically as carbenicillin.

3. ...

► **Alpha-adrenergic blocking agent**

1. [terazosine] Terazosin therapy for chronic prostatitis/chronic pelvic pain syndrome: a randomized, placebo controlled trial. CONCLUSIONS: Terazosin proved superior to placebo for patients with chronic prostatitis/chronic pelvic pain syndrome who had not received alpha-blockers previously.

2. ...

Table 1: System response to the question "What is the best drug treatment for chronic prostatitis?"

systems for decision support represent a potentially high-impact application. From a research perspective, the clinical domain is attractive because substantial knowledge has already been codified in the Unified Medical Language System (UMLS) (Lindberg et al., 1993). The 2004 version of the UMLS Metathesaurus contains information about over 1 million biomedical concepts and 5 million concept names. This and related resources allow us to explore knowledge-based techniques with substantially less upfront investment.

Naturally, physicians have a wide spectrum of information needs, ranging from questions about the selection of treatment options to questions about legal issues. To make the retrieval problem more tractable, we focus on a subset of therapy questions taking the form "What is the best drug treatment for *X*?", where *X* can be any number of diseases. We have chosen to tackle this class of questions because studies of physicians' behavior in natural settings have revealed that such questions occur quite frequently (Ely et al., 1999). By leveraging the natural distribution of clinical information needs, we can make the greatest impact with the least effort.

Our research follows the principles of evidence-based medicine (EBM) (Sackett et al., 2000), which provides a well-defined model to guide the process of clinical question answering. 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 (Cogdill and Moore, 1997; De Groote and Dorsch, 2003), citations from the MEDLINE database (maintained by the U.S. National Library of Medicine) serve as a good source of clinical evidence. As a result of these findings, our work focuses on MEDLINE abstracts as the source for answers.

3 Question Answering Approach

Conflicting desiderata shape the characteristics of "answers" to clinical questions. On the one hand, conciseness is paramount. Physicians are always under time pressure when making decisions, and information overload is a serious concern. Furthermore, we ultimately envision deploying advanced retrieval systems in portable packages such as PDAs to serve as tools in bedside interactions (Hauser et al., 2004). The small form factor of such devices limits the amount of text that can be displayed. However, conciseness exists in tension with completeness. For physicians, 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 is always framed in the context of a specific sample population, over a set duration, at some fixed dosage, etc. A physician simply cannot recommend a particular course of action without considering all these factors.

Our approach seeks to balance conciseness and completeness by providing hierarchical and inter-

active "answers" that support multiple levels of drill-down. A partial example is shown in Figure 1. Top-level answers to "What is the best drug treatment for *X*?" consist of categories of drugs that may be of interest to the physician. Each category is associated with a cluster of abstracts from MEDLINE about that particular treatment option. Drilling down into a cluster, the physician is presented with extractive summaries of abstracts that outline the clinical findings. To obtain more detail, the physician can pull up the complete abstract text, and finally the electronic version of the entire article (if available). In the example shown in Figure 1, the physician can see that two classes of drugs (anti-microbial and alpha-adrenergic blocking agent) are relevant for the disease "chronic prostatitis". Drilling down into the first cluster, the physician can see summarized evidence for two specific types of anti-microbials (temafloxacin and ofloxacin) extracted from MEDLINE abstracts.

Three major capabilities are required to produce the "answers" described above. First, the system must accurately identify the drugs under study in an abstract. Second, the system must group abstracts based on these substances in a meaningful way. Third, the system must generate short summaries of the clinical findings. We describe a clinical question answering system that implements exactly these capabilities (answer extraction, semantic clustering, and extractive summarization).

4 System Implementation

Our work is primarily concerned with synthesizing coherent answers from a set of search results—the actual source of these results is not important. For convenience, we employ MEDLINE citations retrieved by the PubMed search engine (which also serves as a baseline for comparison). Given an initial set of citations, answer generation proceeds in three phases, described below.

4.1 Answer Extraction

Given a set of abstracts, our system first identifies the drugs under study; these later become the short answers. In the parlance of evidence-based medicine, drugs fall into the category of "interventions", which encompasses everything from surgical procedures to diagnostic tests.

Our extractor for interventions relies on MetaMap (Aronson, 2001), a program that automatically identifies entities corresponding to UMLS concepts. UMLS has an extensive coverage of drugs, falling under the semantic type PHARMACOLOGICAL SUBSTANCE and a few others. All such entities are identified as candidates and each is scored based on a number of features: its position in the abstract, its frequency of occurrence, etc. A separate evaluation on a blind test set demonstrates that our extractor is able to accurately recognize the interventions in a MEDLINE abstract; see details in (Demner-Fushman and Lin, 2005; Demner-Fushman and Lin, 2006 in press).

4.2 Semantic Clustering

Retrieved MEDLINE citations are organized into semantic clusters based on the main interventions identified in the abstract text. We employed a variant of the hierarchical agglomerative clustering algorithm (Zhao and Karypis, 2002) that utilizes semantic relationships within UMLS to compute similarities between interventions.

Iteratively, we group abstracts whose interventions fall under a common ancestor, i.e., a hypernym. The more generic ancestor concept (i.e., the class of drugs) is then used as the cluster label. The process repeats until no new clusters can be formed. In order to preserve granularity at the level of practical clinical interest, the tops of the UMLS hierarchy were truncated; for example, the MeSH category "Chemical and Drugs" is too general to be useful. This process was manually performed during system development. We decided to allow an abstract to appear in multiple clusters if more than one intervention was identified, e.g., if the abstract compared the efficacy of two treatments. Once the clusters have been formed, all citations are then sorted in the order of the original PubMed results, with the most abstract UMLS concept as the cluster label. Clusters themselves are sorted in decreasing size under the assumption that more clinical research is devoted to more pertinent types of drugs.

Returning to the example in Figure 1, the abstracts about temafloxacin and ofloxacin were clustered together because both drugs are hyponyms of anti-microbials within the UMLS ontology. As can be seen, this semantic resource provides a powerful tool for organizing search results.

4.3 Extractive Summarization

For each MEDLINE citation, our system generates a short extractive summary consisting of three elements: the main intervention (which is usu-

ally more specific than the cluster label); the title of the abstract; and the top-scoring outcome sentence. The "outcome", another term from evidence-based medicine, asserts the clinical findings of a study, and is typically found towards the end of a MEDLINE abstract. In our case, outcome sentences state the efficacy of a drug in treating a particular disease. Previously, we have built an outcome extractor capable of identifying such sentences in MEDLINE abstracts using supervised machine learning techniques (Demner-Fushman and Lin, 2005; Demner-Fushman and Lin, 2006 in press). Evaluation on a blind held-out test set shows high classification accuracy.

5 Evaluation Methodology

Given that our work draws from QA, IR, and summarization, a proper evaluation that captures the salient characteristics of our system proved to be quite challenging. Overall, evaluation can be decomposed into two separate components: locating a suitable resource to serve as ground truth and leveraging it to assess system responses.

It is not difficult to find disease-specific pharmacology resources. We employed *Clinical Evidence (CE)*, a periodic report created by the British Medical Journal (BMJ) Publishing Group that summarizes the best known drugs for a few dozen diseases. Note that the existence of such secondary sources does not obviate the need for automated systems 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.

For each disease, *CE* classifies drugs into one of six categories: beneficial, likely beneficial, trade-offs (i.e., may have adverse side effects), unknown, unlikely beneficial, and harmful. Included with each entry is a list of references—citations consulted by the editors in compiling the resource. Although the completeness of the drugs enumerated in *CE* is questionable, it nevertheless can be viewed as "authoritative".

5.1 Previous Work

How can we leverage a resource such as *CE* to assess the responses generated by our system? A survey of evaluation methodologies reveals shortcomings in existing techniques.

Answers to factoid questions are automatically scored using regular expression patterns (Lin, 2005). In our application, this is inadequate for many reasons: there is rarely an exact string match between system output and drugs mentioned in *CE*, primarily due to synonymy (for example, alpha-adrenergic blocking agent and α-blocker refer to the same class of drugs) and ontological mismatch (for example, *CE* might mention beta-agonists, while a retrieved abstract discusses formoterol, which is a specific representative of beta-agonists). Furthermore, while this evaluation method can tell us if the drugs proposed by the system are "good", it cannot measure how well the answer is supported by MEDLINE citations; recall that answer justification is important for physicians.

The nugget evaluation methodology (Voorhees, 2005) developed for scoring answers to complex questions is not suitable for our task, since there is no coherent notion of an "answer text" that the user reads end–to–end. Furthermore, it is unclear what exactly a "nugget" in this case would be. For similar reasons, methodologies for summarization evaluation are also of little help. Typically, system-generated summaries are either evaluated manually by humans (which is expensive and time-consuming) or automatically using a metric such as ROUGE, which compares system output against a number of reference summaries. The interactive nature of our answers violates the assumption that systems' responses are static text segments. Furthermore, it is unclear what exactly should go into a reference summary, because physicians may want varying amounts of detail depending on familiarity with the disease and patient-specific factors.

Evaluation methodologies from information retrieval are also inappropriate. User studies have previously been employed to examine the effect of categorized search results. However, they often conflate the effectiveness of the interface with that of the underlying algorithms. For example, Dumais et al. (2001) found significant differences in task performance based on different ways of using purely presentational devices such as mouseovers, expandable lists, etc. While interface design is clearly important, it is not the focus of our work.

Clustering techniques have also been evaluated in the same manner as text classification algorithms, in terms of precision, recall, etc. based on some ground truth (Zhao and Karypis, 2002).

This, however, assumes the existence of stable, invariant categories, which is not the case since our output clusters are query-specific. Although it may be possible to manually create "reference clusters", we lack sufficient resources to develop such a data set. Furthermore, it is unclear if sufficient interannotator agreement can be obtained to support meaningful evaluation.

Ultimately, we devised two separate evaluations to assess the quality of our system output based on the techniques discussed above. The first is a manual evaluation focused on the cluster labels (i.e., drug categories), based on a factoid QA evaluation methodology. The second is an automatic evaluation of the retrieved abstracts using ROUGE, drawing elements from summarization evaluation. Details of the evaluation setup and results are preceded by a description of the test collection we created from *CE*.

5.2 Test Collection

We were able to mine the June 2004 edition of *Clinical Evidence* to create a test collection for system evaluation. We randomly selected thirty diseases, generating 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. *CE* listed an average of 11.3 interventions per disease; of those, 2.3 on average were marked as beneficial and 1.9 as likely beneficial. On average, there were 48.4 references associated with each disease, representing the articles consulted during the compilation of *CE* itself. Of those, 34.7 citations on average appeared in MEDLINE; we gathered all these abstracts, which serve as the reference summaries for our ROUGE-based automatic evaluation.

Since the focus of our work is not on retrieval algorithms per se, we employed PubMed to fetch an initial set of MEDLINE citations and performed answer synthesis using those results. The PubMed citations also serve as a baseline, since it represents a system commonly used by physicians.

In order to obtain the best possible set of citations, the first author (an experienced PubMed searcher), manually formulated queries, taking advantage of MeSH (Medical Subject Headings) terms when available. MeSH terms are controlled vocabulary concepts assigned manually by trained medical indexers (based on the full text of the articles), and encode a substantial amount of knowledge about the contents of the citation. PubMed allows searches on MeSH terms, which usually yield accurate results. In addition, we limited retrieved citations to those that have the MeSH heading "drug therapy" and those that describe a clinical trial (another metadata field). Finally, we restricted the date range of the queries so that abstracts published after our version of *CE* were excluded. Although the query formulation process currently requires a human, we envision automating this step using a template-based approach in the future.

6 System Evaluation

We adapted existing techniques to evaluate our system in two separate ways: a factoid-style manual evaluation focused on short answers and an automatic evaluation with ROUGE using *CE*-cited abstracts as the reference summaries. The setup and results for both are detailed below.

6.1 Manual Evaluation of Short Answers

In our manual evaluation, system outputs were assessed as if they were answers to factoid questions. We gathered three different sets of answers. For the baseline, we used the main intervention from each of the first three PubMed citations. For our test condition, we considered the three largest clusters, taking the main intervention from the first abstract in each cluster. This yields three drugs that are at the same level of ontological granularity as those extracted from the unclustered PubMed citations. For our third condition, we assumed the existence of an oracle which selects the three best clusters (as determined by the first author, a medical doctor). From each of these three clusters, we extracted the main intervention of the first abstracts. This oracle condition represents an achievable upper bound with a human in the loop. Physicians are highly-trained professionals that already have significant domain knowledge. Faced with a small number of choices, it is likely that they will be able to select the most promising cluster, even if they did not previously know it.

This preparation yielded up to nine drug names, three from each experimental condition. For short, we refer to these as PubMed, Cluster, and Oracle, respectively. After blinding the source of the drugs and removing duplicates, each short answer was presented to the first author for evaluation. Since

	Clinical Evidence							Physician		
	B	LB	T	U	UB	H	N	Good	Okay	Bad
PubMed	0.200	0.213	0.160	0.053	0.000	0.013	0.360	0.600	0.227	0.173
Cluster	0.387	0.173	0.173	0.027	0.000	0.000	0.240	0.827	0.133	0.040
Oracle	0.400	0.200	0.133	0.093	0.013	0.000	0.160	0.893	0.093	0.013

Table 2: Manual evaluation of short answers: distribution of system answers with respect to *CE* categories (left side) and with respect to the assessor's own expertise (right side). (Key: B=beneficial, LB=likely beneficial, T=tradeoffs, U=unknown, UB=unlikely beneficial, H=harmful, N=not in *CE*)

the assessor had no idea from which condition an answer came, this process guarded against assessor bias.

Each answer was evaluated in two different ways: first, with respect to the ground truth in *CE*, and second, using the assessor's own medical expertise. In the first set of judgments, the assessor determined which of the six categories (beneficial, likely beneficial, tradeoffs, unknown, unlikely beneficial, harmful) the system answer belonged to, based on the *CE* recommendations. As we have discussed previously, a human (with sufficient domain knowledge) is required to perform this matching due to synonymy and differences in ontological granularity. However, note that the assessor only considered the drug name when making this categorization. In the second set of judgments, the assessor separately determined if the short answer was "good", "okay" (marginal), or "bad" based both on *CE* and her own experience, taking into account the abstract title and the top-scoring outcome sentence (and if necessary, the entire abstract text).

Results of this manual evaluation are presented in Table 2, which shows the distribution of judgments for the three experimental conditions. For baseline PubMed, 20% of the examined drugs fell in the beneficial category; the values are 39% for the Cluster condition and 40% for the Oracle condition. In terms of short answers, our system returns approximately twice as many beneficial drugs as the baseline, a marked increase in answer accuracy. Note that a large fraction of the drugs evaluated were not found in *CE* at all, which provides an estimate of its coverage. In terms of the assessor's own judgments, 60% of PubMed short answers were found to be "good", compared to 83% and 89% for the Cluster and Oracle conditions, respectively. From a factoid QA point of view, we can conclude that our system outperforms the PubMed baseline.

6.2 Automatic Evaluation of Abstracts

A major limitation of the factoid-based evaluation methodology is that it does not measure the quality of the abstracts from which the short answers were extracted. Since we lacked the necessary resources to manually gather abstract-level judgments for evaluation, we sought an alternative.

Fortunately, *CE* can be leveraged to assess the "goodness" of abstracts automatically. We assume that references cited in *CE* are examples of high quality abstracts, since they were used in generating the drug recommendations. Following standard assumptions made in summarization evaluation, we considered abstracts that are similar in content with these "reference abstracts" to also be "good" (i.e., relevant). Similarity in content can be quantified with ROUGE.

Since physicians demand high precision, we assess the cumulative relevance after the first, second, and third abstract that the clinician is likely to have examined (where the relevance for each individual abstract is given by its ROUGE-1 precision score). For the baseline PubMed condition, the examined abstracts simply correspond to the first three hits in the result set. For our test system, we developed three different orderings. The first, which we term cluster round-robin, selects the first abstract from the top three clusters (by size). The second, which we term oracle cluster order, selects three abstracts from the best cluster, assuming the existence of an oracle that informs the system. The third, which we term oracle round-robin, selects the first abstract from each of the three best clusters (also determined by an oracle).

Results of this evaluation are shown in Table 3. The columns show the cumulative relevance (i.e., ROUGE score) after examining the first, second, and third abstract, under the different ordering conditions. To determine statistical significance, we applied the Wilcoxon signed-rank test, the

	Rank 1		Rank 2		Rank 3	
PubMed Ranked List	0.170		0.349		0.523	
Cluster Round-Robin	0.181	(+6.3%)$^{\circ}$	0.356	(+2.1%)$^{\circ}$	0.526	(+0.5%)$^{\circ}$
Oracle Cluster Order	0.206	(+21.5%)$^{\triangle}$	0.392	(+12.6%)$^{\triangle}$	0.597	(+14.0%)$^{\blacktriangle}$
Oracle Round-Robin	0.206	(+21.5%)$^{\triangle}$	0.396	(+13.6%)$^{\triangle}$	0.586	(+11.9%)$^{\blacktriangle}$

Table 3: Cumulative relevance after examining the first, second, and third abstracts, according to different orderings. ($^{\circ}$ denotes n.s., $^{\triangle}$ denotes sig. at 0.90, $^{\blacktriangle}$ denotes sig. at 0.95)

standard non-parametric test for applications of this type. Due to the relatively small test set (only 25 questions), the increase in cumulative relevance exhibited by the cluster round-robin condition is not statistically significant. However, differences for the oracle conditions were significant.

7 Discussion and Related Work

According to two separate evaluations, it appears that our system outperforms the PubMed baseline. However, our approach provides more advantages over a linear result set that are not highlighted in these evaluations. Although difficult to quantify, categorized results provide an overview of the information landscape that is difficult to acquire by simply browsing a ranked list—user studies of categorized search have affirmed its value (Hearst and Pedersen, 1996; Dumais et al., 2001). One main advantage we see in our application is better "redundancy management". With a ranked list, the physician may be forced to browse through multiple redundant abstracts that discuss the same or similar drugs to get a sense of the different treatment options. With our cluster-based approach, however, potentially redundant information is grouped together, since interventions discussed in a particular cluster are ontologically related through UMLS. The physician can examine different clusters for a broad overview, or peruse multiple abstracts within a cluster for a more thorough review of the evidence. Our cluster-based system is able to support both types of behaviors.

This work demonstrates the value of semantic resources in the question answering process, since our approach makes extensive use of the UMLS ontology in all phases of answer synthesis. The coverage of individual drugs, as well as the relationship between different types of drugs within UMLS enables both answer extraction and semantic clustering. As detailed in (Demner-Fushman and Lin, 2006 in press), UMLS-based features are also critical in the identification of clinical outcomes, on which our extractive summaries are based. As a point of comparison, we also implemented a purely term-based approach to clustering PubMed citations. The results are so incoherent that a formal evaluation would prove to be meaningless. Semantic relations between drugs, as captured in UMLS, provide an effective method for organizing results—these relations cannot be captured by keyword content alone. Furthermore, term-based approaches suffer from the cluster labeling problem: it is difficult to automatically generate a short heading that describes cluster content.

Nevertheless, there are a number of assumptions behind our work that are worth pointing out. First, we assume a high quality initial result set. Since the class of questions we examine translates naturally into accurate PubMed queries that can make full use of human-assigned MeSH terms, the overall quality of the initial citations can be assured. Related work in retrieval algorithms (Demner-Fushman and Lin, 2006 in press) shows that accurate relevance scoring of MEDLINE citations in response to more general clinical questions is possible.

Second, our system does not actually perform semantic processing to determine the efficacy of a drug: it only recognizes "topics" and outcome sentences that state clinical findings. Since the system by default orders the clusters based on size, it implicitly equates "most popular drug" with "best drug". Although this assumption is false, we have observed in practice that more-studied drugs are more likely to be beneficial.

In contrast with the genomics domain, which has received much attention from both the IR and NLP communities, retrieval systems for the clinical domain represent an underexplored area of research. Although individual components that attempt to operationalize principles of evidence-based medicine do exist (Mendonça and Cimino, 2001; Niu and Hirst, 2004), complete end–to–end clinical question answering systems are dif-

ficult to find. Within the context of the PERSIVAL project (McKeown et al., 2003), researchers at Columbia have developed a system that leverages patient records to rerank search results. Since the focus is on personalized summaries, this work can be viewed as complementary to our own.

8 Conclusion

The primary contribution of this work is the development of a clinical question answering system that caters to the unique requirements of physicians, who demand both conciseness and completeness. These competing factors can be balanced in a system's response by providing multiple levels of drill-down that allow the information space to be viewed at different levels of granularity. We have chosen to implement these capabilities through answer extraction, semantic clustering, and extractive summarization. Two separate evaluations demonstrate that our system outperforms the PubMed baseline, illustrating the effectiveness of a hybrid approach that leverages semantic resources.

9 Acknowledgments

This work was supported in part by the U.S. National Library of Medicine. The second author thanks Esther and Kiri for their loving support.

References

E. Amigó, J. Gonzalo, V. Peinado, A. Peñas, and F. Verdejo. 2004. An empirical study of information synthesis task. In *ACL 2004*.

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

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

K. Cogdill and M. Moore. 1997. First-year medical students' information needs and resource selection: Responses to a clinical scenario. *Bulletin of the Medical Library Association*, 85(1):51–54.

D. Covell, G. Uman, and P. Manning. 1985. Information needs in office practice: Are they being met? *Annals of Internal Medicine*, 103(4):596–599.

H. Dang. 2005. Overview of DUC 2005. In *DUC 2005 Workshop at HLT/EMNLP 2005*.

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.

D. Demner-Fushman and J. Lin. 2005. Knowledge extraction for clinical question answering: Preliminary results. In *AAAI 2005 Workshop on QA in Restricted Domains*.

D. Demner-Fushman and J. Lin. 2006, in press. Answering clinical questions with knowledge-based and statistical techniques. *Comp. Ling.*

S. Dumais, E. Cutrell, and H. Chen. 2001. Optimizing search by showing results in context. In *CHI 2001*.

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.

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 *MEDINFO 2004*.

M. Hearst and J. Pedersen. 1996. Reexaming the cluster hypothesis: Scatter/gather on retrieval results. In *SIGIR 1996*.

D. Lawrie and W. Croft. 2003. Generating hierarchical summaries for Web searches. In *SIGIR 2003*.

J. Lin. 2005. Evaluation of resources for question answering evaluation. In *SIGIR 2005*.

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

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

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

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

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

E. Voorhees. 2005. Using question series to evaluate question answering system effectiveness. In *HLT/EMNLP 2005*.

Y. Zhao and G. Karypis. 2002. Evaluation of hierarchical clustering algorithms for document datasets. In *CIKM 2002*.