The Role of Knowledge in Conceptual Retrieval: A Study in the Domain of Clinical Medicine

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

ABSTRACT

Despite its intuitive appeal, the hypothesis that retrieval at the level of "concepts" should outperform purely term-based approaches remains unverified empirically. In addition, the use of "knowledge" has not consistently resulted in performance gains. After identifying possible reasons for previous negative results, we present a novel framework for "conceptual retrieval" that articulates the types of knowledge that are important for information seeking. We instantiate this general framework in the domain of clinical medicine based on the principles of evidence-based medicine (EBM). Experiments show that an EBM-based scoring algorithm dramatically outperforms a state-of-the-art baseline that employs only term statistics. Ablation studies further yield a better understanding of the performance contributions of different components. Finally, we discuss how other domains can benefit from knowledge-based approaches.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval Models*

General Terms

Measurement, Experimentation

Keywords

question answering, semantic models, reranking

1. INTRODUCTION

Although the field of information retrieval has made enormous progress in the last half century, virtually all systems are still built on the remarkably simple concept of "counting words". Fundamentally, the vector space [35], probabilistic [33], inference network [26], language modeling [30], and divergence from randomness [1] approaches can be viewed as sophisticated "bookkeeping" techniques for matching words from queries with words in documents, under strong assumptions of term independence. Although these methods have been empirically validated (e.g., in TREC evaluations), it is a simple fact that words alone cannot capture the semantic content of documents and information needs.

This assertion translates naturally into the hypothesis that retrieval systems operating at a level above terms (e.g., concepts, relations, etc.) should outperform purely term-based approaches. Unfortunately, studies along these lines, some dating back nearly two decades, have failed to conclusively support this claim (see Section 2). Here, we provide a novel approach to this age-old problem and demonstrate that large gains in retrieval effectiveness are possible in restricted domains if semantic knowledge is appropriately utilized.

Our work, which lies at the intersection between document retrieval and question answering, has the ambitious goal of developing knowledge-rich "conceptual retrieval" algorithms. This is accomplished in three steps: first, we outline a general framework that identifies the types of knowledge important to information seeking (Section 3). Then, we instantiate this framework in the domain of clinical medicine, mirroring a paradigm of practice known as evidence-based medicine [34] (Sections 4 and 5). Document reranking experiments using a collection of real world clinical questions (Section 6) demonstrate that our approach significantly outperforms a state-of-the-art baseline (Section 7). Finally, we explore the contributions of different knowledge sources (Section 8) and discuss how our ideas can be applied to other domains (Section 9).

2. PREVIOUS WORK

Research on more sophisticated retrieval models can generally be grouped into attempts to go beyond simple term-matching and attempts to relax term independence assumptions. Due to space limitations, we only discuss representative works here.

Deeper linguistic analysis of documents and queries represents a popular avenue of exploration. Typical approaches involve application of NLP techniques such as query expansion (using ontological resources), word-sense disambiguation, and parsing. Previous work has shown that use of lexical semantic relations for query expansion does not increase retrieval performance [38], and neither does indexing syntactic structures [14, 37]; although see [9]. Whether word-

SIGIR'06, August 6–10, 2006, Seattle, Washington, USA.
Copyright 2006 ACM 1-59593-369-7/06/0008 ...$5.00.

sense disambiguation helps retrieval is subject to debate [22, 27, 36, 39], but even positive results show modest improvements at best. More encouraging is recent work on formal models that attempt to capture term dependencies [16, 25]; experiments have yielded gains, suggesting that the problem lies not with the ideas but their implementation. Nevertheless, as Belkin [3] pointed out and Buckley and Harman [4] confirmed empirically, many difficulties surrounding information retrieval are not linguistic in nature. It has been suggested by many researchers that information seeking exists in a much broader context involving real-world tasks, different search strategies, users' cognitive structures, etc. Retrieval models often neglect to account for these important factors.

We believe that little headway has been made in leveraging semantic knowledge in IR because nearly all attempts have occurred in unrestricted domains. Reasoning on anything other than a few lexical relations (e.g., using WordNet) in the open domain is exceedingly difficult because there is a vast amount of world and commonsense knowledge that must be encoded, either manually or automatically. As an example, the massive commonsense knowledge store Cyc [23] was found to have negligible impact on question answering performance in a recent TREC evaluation [6]. A promising approach is the use of abductive inferencing techniques to "justify" candidate answers [28], which, with substantial knowledge engineering, has produced impressive performance on simple fact-based questions. Nevertheless, it is unclear if these methods can be applied to more complex information needs. A possible solution is to sacrifice breadth for depth, as exemplified by recent work on question answering in restricted domains [29], e.g., terrorism. In a more restricted semantic space, it is much easier to explicitly encode the body of knowledge necessary to support conceptual retrieval.

Our approach differs from previous work in two important ways: First, we identify linguistic knowledge as one of three types of knowledge critical to the information-seeking process. Second, within a general framework for conceptual retrieval, we present a case study in the domain of clinical medicine, where existing resources can be effectively leveraged to improve retrieval effectiveness. Through a series of ablation studies, we gain a better understanding of how these different types of knowledge interact.

3. TYPES OF KNOWLEDGE

The idea that information should be retrieved at the conceptual level predates the existence of computers themselves; librarians have been building conceptual structures for organizing information long before the invention of computerized retrieval systems. Even after the development of full-text search engines, it was well known that "bags of words" make poor query representations [3]. The idea that systems for retrieval (computer or otherwise) serve to bring the cognitive representations (of the user and the collection) "into alignment" has been explored within the framework of cognitive information retrieval [20], but this line of work has not resulted in computationally implementable models.

Our attempts to develop a framework for conceptual retrieval begin with an outline of the types of knowledge important to the information-seeking process. In particular, we hypothesize that there are three broad categories of knowledge that should be captured by retrieval algorithms:

- **Knowledge about the problem structure,** or *what* representations are useful for capturing the information need? These representations may reflect cognitive structures of expert information seekers (e.g., the manner in which they decompose the problem and analyze retrieved results) or may be purely computational artifacts (or both).
- **Knowledge about user tasks,** or *why* is this information needed and *how* will it be further used? Typically, a search for information is merely the starting point for other activities (e.g., writing a report, making a decision, etc.). These are what Ingwersen [20] calls "work tasks".
- **Knowledge about the domain,** or what *background knowledge* does the information seeker bring to bear in framing questions and interpreting answers? This includes knowledge of terms used to represent concepts and relationship between concepts.[1]

Based on this framework, we envision retrieval as a process of "semantic unification" between representations that encode user information needs and corresponding representations automatically derived from a text collection. This work describes a specific instantiation of this idea in the domain of clinical medicine.

4. CLINICAL INFORMATION NEEDS

The domain of clinical medicine is an appropriate area in which to explore conceptual retrieval algorithms because the problem structure, task knowledge, and domain knowledge are all relatively well-understood. Furthermore, the need to answer questions related to patient care at the point of service has been well-studied and documented [8, 13, 17]. MEDLINE, the authoritative repository of abstracts from the medical and biomedical primary literature maintained by the U.S. National Library of Medicine (NLM), provides the clinically-relevant sources for answering physicians' questions, and is commonly used in that capacity [7, 10]. However, studies have shown that existing systems for searching MEDLINE (such as PubMed, NLM's online search service) are often unable to provide clinically-relevant answers in a timely manner [5, 17]. Better access to high-quality evidence represents a high-impact decision-support application for physicians.

The centerpiece of our approach is a widely-accepted paradigm for medical practice called evidence-based medicine (EBM), which calls for the explicit use of current best evidence, i.e., the results of high-quality patient-centered clinical research, in making decisions about patient care. Naturally, such evidence, as reported in the primary medical literature, must be suitably integrated with the physician's own expertise and patient-specific factors. It is argued that practicing medicine in this manner leads to better patient outcomes and higher quality health care. One of our goals is to develop accurate retrieval systems that support physicians practicing EBM.

Evidence-based medicine specifies three orthogonal facets of the clinical domain, that, when taken together, describe

[1] Previous work mostly focuses on this. For example, query expansion and word sense disambiguation are approaches that model vocabulary mismatch using domain-independent resources such as WordNet. Phrase-based indexing is an attempt to apply a general model of language to model term dependencies.

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.	**Problem/Population:** 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.

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 [31] identifies four key elements, as shown in Table 1 (middle column). These four elements are often referenced with the mnemonic PICO, which stands for Problem/Population, Intervention, Comparison, and Outcome. Finally, the third facet serves as a tool for appraising the strength of evidence (SoE), i.e., how much confidence should a physician have in the results? For this work, we adopted a taxonomy with three levels of recommendations, as shown in Table 1 (right column).

It should be apparent that evidence-based medicine provides two of the three types of knowledge necessary to support conceptual retrieval. The four clinical tasks ground information needs in broader user activities, and strength of evidence considerations model the pertinence (i.e., non-topical aspects of relevance) in a real-world clinical context. The PICO representation provides a problem structure for capturing clinical information needs. In addition to being a cognitive model for problem analysis (as physicians are trained to decompose complex situations in terms of these elements), PICO frames lend themselves nicely to a computational implementation.

Finally, substantial understanding of the clinical domain has already been codified in the Unified Medical Language System (UMLS) [24]. The 2004 version of the UMLS Metathesaurus contains information about over 1 million biomedical concepts and 5 million concept names from more than 100 controlled vocabularies. In addition, software for utilizing this ontology already exists: MetaMap [2] identifies concepts in free text, while SemRep [32] extracts relations between concepts. In summary, the three types of knowledge identified in the previous section already exist in an accessible form.

Integrating these three perspectives of EBM, we conceptualize retrieval as "semantic unification" between needs expressed in a PICO frame and corresponding structures extracted from MEDLINE abstracts. This matching process, naturally, should be sensitive to task-based considerations. As a concrete example, the question "In children with an acute febrile illness, what is the efficacy of single-medication therapy with acetaminophen or ibuprofen in reducing fever?" might be represented as:

> **Task:** therapy
> **Problem:** acute febrile illness
> **Population:** children
> **Intervention:** acetaminophen
> **Comparison:** ibuprofen
> **Outcome:** reducing fever

This frame representation explicitly encodes the clinical task and the PICO structure of the question. After processing MEDLINE citations, automatically extracting PICO elements from the abstracts, and matching these elements with the query, a system might produce the following answer:

> Ibuprofen provided greater temperature decrement and longer duration of antipyresis than acetaminophen when the two drugs were administered in approximately equal doses.
>
> Strength of Evidence: grade A

Many components are required to realize the above question answering capability: first, knowledge extractors for automatically identifying PICO elements in MEDLINE abstracts; second, a citation scoring algorithm that operationalizes the principles of evidence-based medicine; third, an answer generator that produces responses for physicians. This work focuses on the second: an algorithm that integrates knowledge-based and statistical techniques to assess the relevance of MEDLINE citations with respect to a clinical information need. For identifying PICO frame elements in free text abstracts, we employ previously-developed components, as described in [11, 12]. By leveraging a combination of semantic and lexical features, we demonstrated methods for very precisely extracting clinically-relevant elements: populations, problems, and interventions, which are short phrases, and outcomes, which are sentences that assert clinical findings, e.g., efficacy of a drug for a disease or a comparison between two drugs. The output of these knowledge extractors serves as the input to our algorithm for scoring MEDLINE citations.

5. CITATION SCORING

What is the relevance of a MEDLINE abstract with respect to a clinical question? Evidence-based medicine outlines the need to consider three separate facets, each of which contributes to the total score:

$$S_{\mathrm{EBM}} = \lambda_1 S_{\mathrm{PICO}} + \lambda_2 S_{\mathrm{SoE}} + (1 - \lambda_1 - \lambda_2) S_{\mathrm{MeSH}} \quad (1)$$

The relevance of a particular citation is a weighted linear combination of contributions from matching PICO frames, the strength of evidence of the citation, and associated MeSH terms that are indicative of appropriateness for certain clinical tasks. In the simplest model, each component is equally weighted, but we also experimented with learning optimal λ's from training data. Computing S_{PICO} requires knowledge about the problem structure, while S_{SoE} and S_{MeSH} both reflect knowledge about user tasks. For a more detailed description of the scoring algorithm, see [12].

The following subsections describe how each of these individual scores are computed. We readily concede that our citation scoring algorithm is quite *ad hoc*, since many weights are heuristic reflections of our intuition and domain knowledge. However, we know of no comparable scoring algorithm in the clinical domain, and no suitable data set (of sufficient size) from which to derive model parameters in a more principled fashion. This particular scoring implementation serves as a proof-of-concept, and we leave the development of a more formal model for future work. Furthermore, the primary focus of this paper is not the algorithm itself, but rather an exploration of how different types of knowledge interact in a framework for conceptual retrieval.

5.1 Problem Structure

The score of an abstract based on extracted PICO elements, S_{PICO}, is broken up into individual components based on each frame element:

$$S_{\mathrm{PICO}} = S_{\mathrm{problem}} + S_{\mathrm{population}} + S_{\mathrm{intervention}} + S_{\mathrm{outcome}} \quad (2)$$

The first component in the above equation, S_{problem}, reflects a match between the problem in the query frame and the primary problem identified in the abstract. A score of 1 is given if the problems match based on their UMLS concept id as provided by MetaMap, which essentially performs terminological normalization automatically. 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. Finally, if our problem extractor could not identify a problem (but the query frame does contain a problem), a score of -0.5 is given.

Scores based on population and intervention, $S_{\mathrm{population}}$ and $S_{\mathrm{intervention}}$, respectively, count the lexical overlap between the query frame elements and corresponding elements extracted from abstracts. A point is given to either a matching intervention or a matching population. Our framework collapses the processing of interventions and comparisons because it is often difficult to separate the two (e.g., in an abstract that compares the efficacy of two drugs, which is the "baseline" and which is the comparison?). A single extractor identifies all interventions under consideration.

The outcome-based score, S_{outcome}, is the value assigned to the highest-scoring outcome sentence, as determined by the knowledge extractor. As outcomes are rarely specified explicitly in clinical questions, we decided to omit matching on them. Our citation scoring algorithm simply considers the inherent quality of the outcome statements in an abstract, independent of the query (akin to changing document priors). Given a match on the primary problem, all clinical outcomes are likely to be of interest to the physician.

5.2 Task Knowledge

Two components of the EBM score take into account task knowledge. The first quantifies the strength of evidence:

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

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, such as randomized controlled trials, receive a score of 0.5; observational studies, e.g., case-control studies, 0.3; all non-clinical publications, -1.5; and 0 otherwise. The study type is directly encoded as metadata associated with each MEDLINE citation. Finally, recency factors into the strength of evidence; 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.

The other scoring component that encodes task knowledge is based on MeSH (Medical Subject Headings) terms, which are manually-assigned controlled-vocabulary concepts associated with each MEDLINE citation. For each clinical task, we have gathered a list of terms that are positive or negative indicators of relevance. This score is given by:

$$S_{\mathrm{MeSH}} = \sum_{t \in \mathrm{MeSH}} \alpha(t) \quad (4)$$

The function $\alpha(t)$ maps a MeSH term to a positive score if the term is a positive indicator for that particular task, or a negative score if the term is a negative indicator. For example, genomics-related terms such as "genetics" and "cell physiology" are negative indicators for all tasks, while "drug administration routes" and any of its children are strong positive indicators for the therapy task. We have manually identified several dozen indicators and manually assigned weights; see [12] for more details.

6. EVALUATION METHODOLOGY

Ideally, we would like to apply our scoring algorithm directly to MEDLINE citations. However, this would involve pre-extracting and indexing PICO elements from the 15 plus million entries in the complete MEDLINE database. Unfortunately, we do not have access to the computational resources necessary to accomplish this. As an alternative, we evaluate our EBM-based citation scoring algorithm in an abstract reranking task. This corresponds to a two-stage processing pipeline commonly seen in question answering systems [19]: retrieval of an initial set followed by postprocessing. Our experiments employed PubMed, NLM's gateway to MEDLINE.

Since no suitable test collection for evaluating our algorithm exists, we had to first manually create one. Fortunately, collections of clinical questions (representing real-world information needs of physicians) are available on-line.

Does quinine reduce leg cramps for young athletes? *task:* therapy *primary problem:* leg cramps *co-occurring problems:* muscle cramps, cramps *population:* young adult *intervention:* quinine
How often is coughing the presenting complaint in patients with gastroesophageal reflux disease? *task:* diagnosis *primary problem:* gastroesophageal reflux disease *co-occurring problems:* cough
What's the prognosis of lupoid sclerosis? *task:* prognosis *primary problem:* lupus erythematosus *co-occurring problems:* multiple sclerosis
What are the causes of hypomagnesemia? *task:* etiology *primary problem:* hypomagnesemia

Table 2: Sample clinical questions and frames.

From two sources, the Journal of Family Practice[2] and the Parkhurst Exchange[3], we randomly sampled 50 questions, which were manually classified according to clinical task and coded into PICO-based query frames. Our collection was divided into a development set and a blind held-out test set (24 and 26 questions, respectively). The exact distribution of the questions over the task types is shown in the headings of Table 3; these figures roughly follow the prevalence of question types observed by Ely et al. [13]. One example from each clinical task is shown in Table 2.

For each question, the second author, who is a medical doctor, manually crafted PubMed queries to fetch an initial set of hits. The queries took advantage of PubMed's advanced features and represent "best effort" from an experienced user; it was verified that each hit list contained at least some relevant abstracts. The process of generating queries averaged about forty minutes per question. The top fifty results for each query were retained for our experiments. In total, 2309 citations were retrieved because some queries returned fewer than fifty citations.

All abstracts gathered by the above process were then exhaustively evaluated by the same author. Since all abstracts were judged, we did not have to worry about biases when comparing different systems in a reranking setup. In total, the relevance assessment process took approximately 100 hours, or about an average of 2 hours per question.

Our reranking experiment compared four different conditions: the baseline PubMed results; hits reranked using Indri, a state-of-the-art language modeling toolkit [26] (using the questions verbatim as queries); hits reranked by the EBM-scoring algorithm described in Section 5; and hits reranked by combining Indri and EBM scores, $\lambda S_{\text{EBM}} + (1 - \lambda) S_{\text{Indri}}$. The development questions were extensively used in the crafting of the citation scoring algorithm (especially in the manual determination of weights).

To evaluate retrieval effectiveness, we collected the following metrics: mean average precision (MAP), precision at ten retrieved documents (P10), and mean reciprocal rank (MRR). Mean average precision is the most widely-accepted single-point retrieval metric. Precision at top documents is particularly important in a real-world clinical setting because physicians are often under intense time pressure. Mean reciprocal rank, a metric often used for question answering, quantifies the expected position of the first relevant hit.

7. RESULTS

Results of our reranking experiment are shown in Table 3. For the EBM run, each scoring component was equally weighted (i.e., $\lambda_1 = \lambda_2 = 1/3$). For the EBM+Indri run, we settled on a λ of 0.85, which optimized performance over the development set. The Wilcoxon signed-rank test was employed to determine the statistical significance of the results; significance at the 1% level is indicated by either $^\blacktriangle$ or $^\blacktriangledown$, depending on the direction of change; significance at the 5% level, $^\vartriangle$ or $^\triangledown$; *n.s.* is denoted by $^\circ$.

All three conditions (Indri, EBM, EBM+Indri) significantly outperform the PubMed baseline on all metrics. In many cases, the differences are very dramatic, e.g., the EBM algorithm more than doubles MAP and P10 on the test set (vs. PubMed). There are enough therapy questions to achieve statistical significance in the task-specific results; however, due to a smaller number of questions for the other tasks, those results are not statistically significant.

Are differences in performance between Indri, EBM, and EBM+Indri statistically significant? Results of the Wilcoxon signed-rank test are shown in Table 4. For all but one case (MRR on the development set), our EBM scoring algorithm significantly outperforms Indri alone—which supports our claim that appropriate use of semantic knowledge can yield substantial improvements over state-of-the-art ranking methods based solely on term statistics. Furthermore, combining term-based statistical evidence from Indri with EBM scores results in a small but statistically significant increase in MAP on both the development and test set.

For the above experiments, the PICO, SoE, and MeSH components of the EBM score were weighted equally. Separate experiments reported in [12] attempted to optimize λ_1 and λ_2 using the development set. However, optimal weights did not result in statistically significant differences, suggesting that the performance of the EBM-scoring algorithm is relatively insensitive to specific weight settings. We conclude that retrieval performance can be attributed primarily to the use of different semantic resources, as opposed to a fortunate setting of parameters.

Nevertheless, it is important to determine the performance contributions of each knowledge component within our conceptual retrieval framework. The results of ablation studies that isolate each score component are shown in Table 5. As can be seen, each component contributes significantly to the overall performance, given the fact that using S_{PICO}, S_{SoE}, and S_{MeSH} individually results in significantly lower performance (vs. all three components). In general, the PICO score alone outperforms Indri, but not SoE or MeSH alone.

8. PARTIAL SEMANTIC MODELS

The domain of medicine represents a fortunate confluence of circumstances in which problem structure, task knowledge, and domain knowledge are all readily available. In many domains, one or more components may be missing or

[2] `http://www.jfponline.com/`

[3] `http://www.parkhurstexchange.com/qa/`

Development Set		**Therapy** (10)	**Diagnosis** (6)	**Prognosis** (3)	**Etiology** (5)	**All** (24)
MAP	baseline	0.354	0.421	0.385	0.608	0.428
	Indri	0.706 (+100%)$^{\vartriangle}$	0.521 (+24%)$^{\circ}$	0.502 (+30%)$^{\circ}$	0.686 (+13%)$^{\circ}$	0.630 (+47%)$^{\blacktriangle}$
	EBM	0.819 (+131%)$^{\blacktriangle}$	0.794 (+89%)$^{\vartriangle}$	0.635 (+65%)$^{\circ}$	0.649 (+6.7%)$^{\circ}$	0.754 (+76%)$^{\blacktriangle}$
	EBM+Indri	0.826 (+133%)$^{\blacktriangle}$	0.800 (+90%)$^{\vartriangle}$	0.632 (+64%)$^{\circ}$	0.665 (+9.3%)$^{\circ}$	0.762 (+78%)$^{\blacktriangle}$
P10	baseline	0.300	0.367	0.400	0.533	0.378
	Indri	0.620 (+107%)$^{\blacktriangle}$	0.483 (+32%)$^{\circ}$	0.467 (+17%)$^{\circ}$	0.613 (+15%)$^{\circ}$	0.565 (+50%)$^{\blacktriangle}$
	EBM	0.730 (+143%)$^{\blacktriangle}$	0.800 (+118%)$^{\vartriangle}$	0.633 (+58%)$^{\circ}$	0.553 (+3.7%)$^{\circ}$	0.699 (+85%)$^{\blacktriangle}$
	EBM+Indri	0.740 (+147%)$^{\blacktriangle}$	0.800 (+118%)$^{\vartriangle}$	0.633 (+58%)$^{\circ}$	0.553 (+3.7%)$^{\circ}$	0.703 (+86%)$^{\blacktriangle}$
MRR	baseline	0.428	0.792	0.733	0.900	0.656
	Indri	0.900 (+110%)$^{\blacktriangle}$	0.756 (−4.6%)$^{\circ}$	0.833 (+13.6%)$^{\circ}$	1.000 (+11%)$^{\circ}$	0.876 (+34%)$^{\blacktriangle}$
	EBM	0.933 (+118%)$^{\vartriangle}$	0.917 (+16%)$^{\circ}$	0.667 (−9.1%)$^{\circ}$	1.000 (+11%)$^{\circ}$	0.910 (+39%)$^{\blacktriangle}$
	EBM+Indri	0.933 (+118%)$^{\vartriangle}$	0.917 (+16%)$^{\circ}$	0.667 (−9.1%)$^{\circ}$	1.000 (+11%)$^{\circ}$	0.910 (+39%)$^{\blacktriangle}$

Test Set		**Therapy** (12)	**Diagnosis** (6)	**Prognosis** (3)	**Etiology** (5)	**All** (26)
MAP	baseline	0.421	0.279	0.235	0.364	0.356
	Indri	0.595 (+41%)$^{\blacktriangle}$	0.534 (+92%)$^{\circ}$	0.533 (+127%)$^{\circ}$	0.439 (+20%)$^{\circ}$	0.544 (+53%)$^{\blacktriangle}$
	EBM	0.765 (+82%)$^{\blacktriangle}$	0.637 (+129%)$^{\vartriangle}$	0.722 (+207%)$^{\circ}$	0.701 (+93%)$^{\circ}$	0.718 (+102%)$^{\blacktriangle}$
	EBM+Indri	0.777 (+84%)$^{\blacktriangle}$	0.672 (+141%)$^{\vartriangle}$	0.711 (+203%)$^{\circ}$	0.701 (+92%)$^{\circ}$	0.730 (+105%)$^{\blacktriangle}$
P10	baseline	0.350	0.150	0.200	0.320	0.281
	Indri	0.575 (+64%)$^{\blacktriangle}$	0.500 (+233%)$^{\circ}$	0.367 (+83%)$^{\circ}$	0.400 (+25%)$^{\circ}$	0.500 (+78%)$^{\blacktriangle}$
	EBM	0.783 (+124%)$^{\blacktriangle}$	0.583 (+289%)$^{\vartriangle}$	0.467 (+133%)$^{\circ}$	0.660 (+106%)$^{\circ}$	0.677 (+141%)$^{\blacktriangle}$
	EBM+Indri	0.775 (+121%)$^{\blacktriangle}$	0.617 (+311%)$^{\vartriangle}$	0.433 (+117%)$^{\circ}$	0.660 (+106%)$^{\circ}$	0.677 (+141%)$^{\blacktriangle}$
MRR	baseline	0.579	0.443	0.456	0.540	0.526
	Indri	0.750 (+30%)$^{\circ}$	0.728 (+64%)$^{\circ}$	0.833 (+83%)$^{\circ}$	0.380 (−30%)$^{\circ}$	0.683 (+30%)$^{\blacktriangle}$
	EBM	0.917 (+58%)$^{\vartriangle}$	0.889 (+101%)$^{\circ}$	1.000 (+119%)$^{\circ}$	1.000 (+85%)$^{\circ}$	0.936 (+78%)$^{\blacktriangle}$
	EBM+Indri	0.917 (+58%)$^{\vartriangle}$	0.889 (+101%)$^{\circ}$	1.000 (+119%)$^{\circ}$	1.000 (+85%)$^{\circ}$	0.936 (+78%)$^{\blacktriangle}$

Table 3: Results of reranking experiments. ($^{\blacktriangle}$,$^{\blacktriangledown}$=sig. at 1%; $^{\vartriangle}$,$^{\triangledown}$=sig. at 5%; $^{\circ}$=*n.s.*)

	Development Set			**Test Set**		
	MAP	**P10**	**MRR**	**MAP**	**P10**	**MRR**
EBM vs. Indri	+19.7% $^{\vartriangle}$	+23.6% $^{\vartriangle}$	+3.8% $^{\circ}$	+32.1% $^{\blacktriangle}$	+35.4% $^{\blacktriangle}$	+37.0% $^{\blacktriangle}$
EBM+Indri vs. Indri	+20.9% $^{\blacktriangle}$	+24.3% $^{\blacktriangle}$	+3.8% $^{\circ}$	+34.3% $^{\blacktriangle}$	+35.4% $^{\blacktriangle}$	+37.0% $^{\blacktriangle}$
EBM+Indri vs. EBM	+1.0% $^{\vartriangle}$	+0.60% $^{\vartriangle}$	+0.0% $^{\circ}$	+1.7% $^{\vartriangle}$	+0.0% $^{\circ}$	+0.0% $^{\circ}$

Table 4: Performance differences between various rerankers.

not (yet) computationally accessible. Statistical term-based ranking algorithms have the advantage that minimal effort is required to move from domain to domain. In the cases where only a limited amount of knowledge is available, is it possible to obtain the best of both worlds by combining term-based and knowledge-derived evidence?

Additional experiments with our EBM algorithm shed some light on this question. We conducted a number of runs that combined Indri scores with components of the EBM score by linear weighting, $\lambda S_{\text{Indri}}+(1-\lambda)S_{\text{EBM*}}$, where $S_{\text{EBM*}}$ represents different ablated variants of the EBM scoring algorithm. The weights were tuned using the development set. Results of these experiments are shown in Table 6.

We can see that the availability of any individual source of evidence improves Indri results. In this specific domain, problem structure contributes the greatest, although task knowledge also plays an important role. We can view SoE and MeSH scores as modeling non-uniform priors on the relevance of specific document types, based on the particular task at hand. To conclude, not only can a knowledge-based approach to retrieval yield significant improvements over purely term-based methods, but fragmentary evidence from individual knowledge sources can still be useful.

9. APPLICATIONS TO OTHER DOMAINS

Since the primary thrust of this research is a general framework for conceptual retrieval—with our EBM citation scoring algorithm as an illustrative instantiation—it is important to demonstrate the generality of our ideas. In this section, we briefly discuss how other applications might benefit from similar semantic modeling.

The genomics domain represents a straightforward extension to the work presented here—instead of a physician, the target user would be a biomedical researcher. Domain coverage could be provided by UMLS and other specialized sources, e.g., the Gene Ontology (GO) or Online Mendelian Inheritance in Man (OMIM). As with the clinical domain, there exist generalized categories of information needs, as exemplified by query templates in the TREC 2005 genomics track [18]. An example is "What is the role of [gene] in [disease]", fully instantiated in "What is the role of the gene Transforming growth factor-beta1 (TGF-beta1) in the disease Cerebral Amyloid Angiopathy (CAA)?" Finally, task knowledge is not difficult to obtain, given the existence of well-defined task models, e.g., determining the genetic basis of a disease or drug discovery.

Beyond life sciences, our framework for conceptual re-

	MAP	vs. EBM	vs. Indri	**P10**	vs. EBM	vs. Indri	**MRR**	vs. EBM	vs. Indri
Development Set									
S_{PICO}	0.709	$-6.0\%^{\triangledown}$	$+12.5\%^{\circ}$	0.657	$-6.0\%^{\circ}$	$+16.2\%^{\circ}$	0.903	$-0.8\%^{\circ}$	$+3.0\%^{\circ}$
S_{SoE}	0.512	$-32.2\%^{\blacktriangledown}$	$-18.8\%^{\triangledown}$	0.482	$-31.0\%^{\blacktriangledown}$	$-14.7\%^{\circ}$	0.674	$-25.9\%^{\blacktriangledown}$	$-23.1\%^{\triangledown}$
S_{MeSH}	0.512	$-32.2\%^{\blacktriangledown}$	$-18.8\%^{\circ}$	0.457	$-34.6\%^{\blacktriangledown}$	$-19.2\%^{\triangledown}$	0.714	$-21.6\%^{\triangledown}$	$-18.6\%^{\circ}$
$S_{\text{SoE}}+ S_{\text{MeSH}}$	0.556	$-26.4\%^{\blacktriangledown}$	$-12.0\%^{\circ}$	0.528	$-24.5\%^{\blacktriangledown}$	$-6.6\%^{\circ}$	0.781	$-14.2\%^{\circ}$	$-10.9\%^{\circ}$
Test Set									
S_{PICO}	0.646	$-10.0\%^{\blacktriangledown}$	$+18.8\%^{\triangle}$	0.627	$-7.4\%^{\circ}$	$+25.4\%^{\blacktriangle}$	0.847	$-9.5\%^{\circ}$	$+24.0\%^{\circ}$
S_{SoE}	0.457	$-36.4\%^{\blacktriangledown}$	$-16.0\%^{\triangledown}$	0.427	$-36.9\%^{\blacktriangledown}$	$-14.6\%^{\circ}$	0.644	$-31.1\%^{\blacktriangledown}$	$-5.7\%^{\circ}$
S_{MeSH}	0.504	$-29.8\%^{\blacktriangledown}$	$-7.3\%^{\circ}$	0.435	$-35.8\%^{\blacktriangledown}$	$-13.1\%^{\circ}$	0.663	$-29.2\%^{\blacktriangledown}$	$-3.0\%^{\circ}$
$S_{\text{SoE}}+ S_{\text{MeSH}}$	0.538	$-25.1\%^{\blacktriangledown}$	$-1.1\%^{\circ}$	0.485	$-28.4\%^{\blacktriangledown}$	$-3.1\%^{\circ}$	0.677	$-27.6\%^{\blacktriangledown}$	$-0.9\%^{\circ}$

Table 5: Performance contribution of different EBM score components. ($^{\blacktriangle}$,$^{\blacktriangledown}$=sig. at 1%; $^{\triangle}$,$^{\triangledown}$=sig. at 5%; $^{\circ}$=*n.s.*)

	λ	**MAP**	**P10**	**MRR**
Development Set				
S_{Indri}		0.630	0.565	0.876
$\lambda S_{\text{Indri}}+(1-\lambda)S_{\text{PICO}}$	0.46	0.718 (+13.9%)$^{\triangle}$	0.669 (+18.4%)$^{\triangle}$	0.917 (+4.6%)$^{\circ}$
$\lambda S_{\text{Indri}}+(1-\lambda)S_{\text{SoE}}$	0.77	0.663 (+5.1%)$^{\blacktriangle}$	0.586 (+3.7%)$^{\circ}$	0.946 (+7.9%)$^{\circ}$
$\lambda S_{\text{Indri}}+(1-\lambda)S_{\text{MeSH}}$	0.77	0.657 (+4.2%)$^{\triangle}$	0.603 (+6.6%)$^{\circ}$	0.866 (−1.2%)$^{\circ}$
$\lambda S_{\text{Indri}}+(1-\lambda)(0.5S_{\text{SoE}}+0.5S_{\text{MeSH}})$	0.55	0.679 (+7.7%)$^{\circ}$	0.607 (+7.4%)$^{\circ}$	0.917 (+4.6%)$^{\circ}$
Test Set				
S_{Indri}		0.544	0.500	0.683
$\lambda S_{\text{Indri}}+(1-\lambda)S_{\text{PICO}}$	0.46	0.668 (+22.9%)$^{\blacktriangle}$	0.627 (+25.4%)$^{\blacktriangle}$	0.897 (+31.3%)$^{\blacktriangle}$
$\lambda S_{\text{Indri}}+(1-\lambda)S_{\text{SoE}}$	0.77	0.578 (+6.3%)$^{\triangle}$	0.554 (+10.8%)$^{\triangle}$	0.766 (+12.2%)$^{\circ}$
$\lambda S_{\text{Indri}}+(1-\lambda)S_{\text{MeSH}}$	0.77	0.564 (+3.8%)$^{\triangle}$	0.531 (+6.2%)$^{\circ}$	0.731 (+6.9%)$^{\circ}$
$\lambda S_{\text{Indri}}+(1-\lambda)(0.5S_{\text{SoE}}+0.5S_{\text{MeSH}})$	0.55	0.620 (+14.0%)$^{\blacktriangle}$	0.565 (+13.1%)$^{\triangle}$	0.876 (+28.2%)$^{\blacktriangle}$

Table 6: Impact of using partial semantic knowledge.

trieval is broadly applicable to other domains as well. Here, we briefly discuss three others: patent search, enterprise search, and QA in the terrorism/warfighting domain. For patent search [21], the USPTO maintains an extensive classification system that comprises the core of a domain model. Search tasks and information needs are specific and well-defined, e.g., discovery of prior art. In the realm of enterprise search in workplace settings, Freund et al. [15] have identified four broad categories of information needs ("how to", "why", "what", and "show me") and patterns of association between tasks (e.g., "performance tuning") and genres (e.g., "cookbook" or "demo"). The appropriateness of different genres to different tasks parallels Strength of Evidence considerations in medicine, and categories of information needs translate naturally into template-based problem structures. In summary, existing resources in the patent and enterprise domains also support a knowledge-based treatment.

Question answering in the terrorism/warfighting domain has become a widely-researched topic, given current funding priorities in the United States, as exemplified by research programs such as AQUAINT and GALE. In this domain, the triplet of problem structure, task model, and domain knowledge is available. In terms of problem structure, well-known query prototypes have been studied (paralleling query templates in the genomics track), as well as the representations for reasoning about such problems, e.g., "recipes" for acquiring specific weapons of mass destruction. These information needs can be decomposed into simpler structures, which can serve as the basis for a network of related semantic frames that cover the problem domain (e.g., acquire radiological material, build device, etc.). Task models are relatively well specified and functional boundaries are clearly delineated; for example, the interaction between intelligence and operational planning is well understood. Finally, domain-specific ontologies have already been built. All of these elements provide the foundation for conceptual retrieval algorithms that incorporate rich sources of knowledge.

10. CONCLUSION

The contributions of this paper are a general framework for conceptual retrieval and a concrete instantiation of the approach in the clinical domain. We have identified three types of knowledge that are important in information seeking: problem structure (PICO frames), task knowledge (clinical tasks and SoE considerations), and domain knowledge (UMLS). Experiments show that a citation scoring algorithm which operationalizes the principles of evidence-based medicine dramatically outperforms a state-of-the-art baseline in retrieving MEDLINE citations. In addition, ablation studies help us better understand the performance contributions of each scoring component. This work provides a tantalizing peek at the significant advances that can be made in information retrieval based on appropriate use of semantic knowledge, and hopefully paves the way for future work.

11. ACKNOWLEDGMENTS

This work was supported in part by the National Library of Medicine, where the first author was a visiting scientist during the summer of 2005. We'd like to thank D. Oard and

D. Soergel for helpful comments. The first author would like to thank Esther and Kiri for their loving support.

12. REFERENCES

[1] G. Amati and C. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM TOIS*, 20(4):357–389, 2002.

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

[3] N. Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5:133–143, 1980.

[4] C. Buckley and D. Harman. Reliable information access final workshop report, 2004.

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

[6] J. Chu-Carroll, J. Prager, C. Welty, K. Czuba, and D. Ferrucci. A multi-strategy and multi-source approach to question answering. In *TREC 2002*.

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

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

[9] H. Cui, R. Sun, K. Li, M.-Y. Kan, and T.-S. Chua. Question answering passage retrieval using dependency relations. In *SIGIR 2005*.

[10] S. De Groote and J. Dorsch. Measuring use patterns of online journals and databases. *Journal of the Medical Library Association*, 91(2):231–241, 2003.

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

[12] D. Demner-Fushman and J. Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 2006, in press.

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

[14] J. Fagan. *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparisons of Syntactic and Non-Syntactic Methods*. Ph.D., Cornell, 1987.

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

[16] J. Gao, J.-Y. Nie, G. Wu, and G. Cao. Dependence language model for information retrieval. In *SIGIR 2004*.

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

[18] W. Hersh, A. Cohen, J. Yang, R. Bhupatiraju, P. Roberts, and M. Hearst. TREC 2005 genomics track overview. In *TREC 2005*.

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

[20] P. Ingwersen. Cognitive information retrieval. *ARIST*, 34:3–52, 1999.

[21] N. Kando and M.-K. Leong. Workshop on patent retrieval: SIGIR 2000 workshop report. *SIGIR Forum*, 34(1):28–30, 2000.

[22] S.-B. Kim, H.-C. Seo, and H.-C. Rim. Information retrieval using word sense: Root sense tagging approach. In *SIGIR 2004*.

[23] D. Lenat. CYC: A large-scale investment in knowledge infrastructure. *CACM*, 38(11):33–38, 1995.

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

[25] D. Metzler and W. Croft. A Markov random field model for term dependencies. In *SIGIR 2005*.

[26] D. Metzler and W. Croft. Combining the language model and inference network approaches to retrieval. *Information Processing and Management*, 40(5):735–750, 2004.

[27] R. Mihalcea and D. Moldovan. Semantic indexing using WordNet senses. In *Proc. of ACL 2000 Workshop on Recent Advances in NLP and IR*.

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

[29] S. Narayanan and S. Harabagiu. Question answering based on semantic structures. In *COLING 2004*.

[30] J. Ponte and W. Croft. A language modeling approach to information retrieval. In *SIGIR 1998*.

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

[32] T. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477, 2003.

[33] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *TREC-3*, 1994.

[34] D. Sackett, S. Straus, W. Richardson, W. Rosenberg, and R. Haynes. *Evidence-Based Medicine: How to Practice and Teach EBM*. Churchill Livingstone, second edition, 2000.

[35] G. Salton. A vector space model for information retrieval. *CACM*, 18(11):613–620, 1975.

[36] M. Sanderson. Word-sense disambiguation and information retrieval. In *SIGIR 1994*.

[37] A. Smeaton, R. O'Donnell, and F. Kelledy. Indexing structures derived from syntax in TREC-3: System description. In *TREC-3*, 1994.

[38] E. Voorhees. Query expansion using lexical-semantic relations. In *SIGIR 1994*.

[39] E. Voorhees. Using WordNet to disambiguate word senses for text retrieval. In *SIGIR 1993*.