ABSTRACT

Title of dissertation: COMPLEX QUESTION ANSWERING BASED ON A SEMANTIC DOMAIN MODEL OF CLINICAL MEDICINE

Dina Demner-Fushman,
Doctor of Philosophy, 2006

Dissertation directed by: Professor Douglas W. Oard
Professor Jimmy Lin
Department of Computer Science

Much research in recent years has focused on question answering. Due to significant advances in answering simple fact-seeking questions, research is moving towards resolving complex questions. An approach adopted by many researchers is to decompose a complex question into a series of fact-seeking questions and re-use techniques developed for answering simple questions. This thesis presents an alternative novel approach to domain-specific complex question answering based on consistently applying a semantic domain model to question and document understanding as well as to answer extraction and generation.

This study uses a semantic domain model of clinical medicine to encode (a) a clinician's information need expressed as a question on the one hand and (b) the meaning of scientific publications on the other to yield a common representation. It is hypothesized that this approach will work well for (1) finding documents that contain answers to clinical questions and (2) extracting these answers from the documents.

The domain of clinical question answering was selected primarily because of its unparalleled resources that permit providing a proof by construction for this hypothesis. In addition, a working prototype of a clinical question answering system will support research in informed clinical decision making. The proposed methodology is based on the semantic domain model developed within the paradigm of Evidence Based Medicine. Three basic components of this model – the clinical task, a framework for capturing a synopsis of a clinical scenario that generated the question, and strength of evidence presented in an answer – are identified and discussed in detail.

Algorithms and methods were developed that combine knowledge-based and statistical techniques to extract the basic components of the domain model from abstracts of biomedical articles. These algorithms serve as a foundation for the prototype end-to-end clinical question answering system that was built and evaluated to test the hypotheses.

Evaluation of the system on test collections developed in the course of this work and based on real life clinical questions demonstrates feasibility of complex question answering and high accuracy information retrieval using a semantic domain model.

COMPLEX QUESTION ANSWERING BASED ON A SEMANTIC DOMAIN MODEL OF CLINICAL MEDICINE

by

Dina Demner-Fushman

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2006

Advisory Committee:
Professor Douglas W. Oard, Chair/Advisor
Professor Jimmy Lin, Co-Chair/Co-Advisor
Professor Dagobert Soergel
Professor Ben Shneiderman
Professor Philip Resnik
Dr. Thomas C. Rindflesch

ACKNOWLEDGMENTS

This work would have not been possible without the support and encouragement of my advisors, members of my committee, and my colleagues at NLM.

I am equally indebted to all these remarkable people, and very grateful to have had an opportunity to work with, and be advised by each one of them.

First I would like to thank my advisor, Dr. Doug Oard who never fails to amaze me with the breadth of his knowledge, ability to capture the essential in research and guide his students towards this understanding. I am very grateful to Dr. Oard for introducing me to the field of Information Retrieval, broadening my horizon, and his constant support.

This dissertation and my research benefited immensely from working with Dr. Jimmy Lin, my co-advisor. His ability to see the overall picture, keep me focused and on the right track was instrumental in completion of this work.

I appreciate the insights, constructive criticism, and support of my committee members: Dr. Ben Shneiderman, Dr. Dagobert Soergel, Dr. Tom Rindflesch, and Dr. Philip Resnik, who introduced me to the world of human-computer interaction studies, natural language processing, and information studies, gracefully agreed to serve on my committee, and were very generous with their valuable time in monitoring and guiding my progress.

My thanks go to Dr. Susan Hauser, my supervisor at the National Library of

Medicine who combines an ability to envision futuristic projects with the practicality of an engineer who actually implements these projects.

I am grateful to Dr. George Thoma, Dr. Lan Aronson, and Susanne Humphrey for their interest in my research, support, and thought-stimulating discussions.

It is impossible to name everyone who I am indebted to: people at Hunter College who gave me a second chance, the CLIP lab, with its amalgamation of talents, people who directly helped with this work and encouraged me at the National Library of Medicine, and last, but not least, my patient and supportive family.

Thank you all!

TABLE OF CONTENTS

List of Figures vi

List of Abbreviations vii

1 Introduction 1
1.1 Motivation . . . 1
1.2 Contributions . . . 7
1.3 Outline of Thesis . . . 8

2 Review of the Literature 10
2.1 Information needs of clinicians . . . 10
2.2 The Evidence-Based Medicine domain model . . . 14
2.2.1 Clinical tasks . . . 16
2.2.2 A framework for synopsis of a clinical scenario (PICO) . . . 17
2.2.2.1 PICO frame for Therapy or Prevention . . . 18
2.2.2.2 PICO frame for Diagnosis . . . 19
2.2.2.3 PICO frames for Etiology . . . 20
2.2.2.4 PICO frame for Prognosis . . . 21
2.2.3 Quality of research / Strength of evidence in medical articles . 22
2.3 Prior use of the EBM-based domain model in Information Retrieval and Natural Language Processing . . . 26
2.4 Information Retrieval and Natural Language Processing in the medical domain . . . 29
2.4.1 Document /Discourse structure utilization . . . 30
2.4.2 Domain knowledge utilization . . . 32
2.5 Question Answering . . . 36
2.5.1 Open Domain Question Answering . . . 39
2.5.2 Closed Domain Question Answering . . . 41

3 Resources, Tools and Test Collections 44
3.1 Unified Medical Language System (UMLS) . . . 44
3.2 MEDLINE database . . . 48
3.3 PubMed . . . 51
3.4 Essie . . . 53
3.5 MetaMap and MMTx . . . 54
3.6 SemRep . . . 58
3.7 Online databases of clinical questions and answers . . . 59
3.7.1 Parkhurst Exchange Forum . . . 59
3.7.2 American Family Physicians Inquiries Network . . . 60
3.7.3 Clinical Evidence . . . 61
3.8 Test Collections . . . 62

4 EBM-based Question Answering System 73
4.1 System architecture overview . . . 75
4.2 User Query: PICO frames . . . 77
4.2.1 Identifying PICO elements in a real life clinical question . . . 77
4.2.2 Instantiating input PICO frames . . . 81
4.2.3 Frames to PubMed queries . . . 83
4.3 Document Retrieval using Entrez Utilities . . . 88
4.4 Document frame generation . . . 88
4.4.1 Citation text preprocessing . . . 89
4.4.2 Semantic Processing . . . 93
4.4.2.1 Knowledge (PICO) extraction . . . 93
4.4.2.2 Clinical Task classification . . . 106
4.4.2.3 Strength of Evidence classification . . . 110
4.5 Document scoring and ranking . . . 113
4.5.1 Question–Document frame matching. (PICO score) . . . 115
4.5.2 Document scoring example . . . 119
4.6 Answer generation . . . 123
4.6.1 Semantic clustering . . . 125

5 System evaluation 131
5.1 Evaluation metrics . . . 132
5.2 Evaluation of knowledge extractors . . . 134
5.3 Re-ranking evaluation . . . 141
5.3.1 Pilot document re-ranking experiments . . . 142
5.3.2 Document re-ranking experiments . . . 145
5.3.3 Domain model components contribution to re-ranking . . . 149
5.4 Answer generation evaluation . . . 151
5.4.1 Best answer evaluation . . . 151
5.4.2 Interactive clinical scenario emulation . . . 156
5.4.3 Exploration of automatic evaluation of abstracts . . . 168

6 Conclusions 173
6.1 Recapitulation . . . 173
6.2 Implications . . . 175
6.3 Limitations . . . 176
6.4 Future work . . . 179

A Search strategies for PICO-annotated collection 184

B Questions in the FPIN collection 185

C Diseases in the CE collection 188

D Extracting comparative structures 190

E Coefficients for document ranking and answer generation 195

LIST OF FIGURES

2.1 MEDLINE abstract for PP-ICONS analysis. 23

3.1 The concept *Lou Gehrig's Disease* and some of its hypernyms. 47

3.2 MEDLINE citation in MEDLINE format. 48

3.3 MetaMap machine output example. 56

4.1 EBM-based Question Answering 74

4.2 System architecture 75

4.3 Positions of outcome statements in 275 training abstracts. 103

4.4 PICO elements automatically annotated in the abstract of MEDLINE citation 12874489 119

4.5 Otitis Externa key points in BMJ Clinical Evidence. 128

5.1 Semantic clustering on problems. 157

5.2 Semantic clustering on interventions. 158

5.3 Reference answer in the Parkhurst Database. 159

5.4 Cluster selection for evaluation 160

5.5 Results of answer clustering for the *FPIN* collection. 163

D.1 Drug comparisons identified using SemRep. 191

LIST OF ABBREVIATIONS

BMA	British Medical Academy
BMJ	British Medical Journal
CE	Clinical Evidence
CQA	Clinical Question Answering system
CUI	Concept Unique Identifier
EBM	Evidence Based Medicine
FPIN	Family Practitioner Inquiry Network
FSM	Finite-State Machine
JAMA	the Journal of the American Medical Association
IR	Information Retrieval
MeSH	Medical Subject Headings
MMTx	MetaMap Technology Transfer
NLM	National Library of Medicine
NLP	Natural Language Processing
PICO	Patient/Intervention/Comparison/Outcome framework
PP-ICONS	Patient/Problem-Intervention/Comparison/Number/Statistics
QA	Question Answering
RCT	Randomized Clinical Trial
SoE	Strength of Evidence
TREC	Text REtrieval Conference
URA	Uniform Retrieval Architecture
UMLS	Unified Medical Language System

Chapter 1

Introduction

> *It is vital to remember that information - in the sense of raw data - is not knowledge; that knowledge is not wisdom; and that wisdom is not foresight. But information is the first essential step to all of these.*
>
> Attributed to Arthur C. Clarke

1.1 Motivation

People need information to successfully operate in their environment, support their decisions and solve problems. A gap between an individual's background knowledge and information needed in a certain context motivates a search for information. Not surprisingly therefore, information retrieval systems became essential in everyday life as well as in carrying out professional tasks. An ideal information retrieval system would be capable of understanding a user's information need and delivering exactly the needed information in desired form, complexity, specificity, and completeness. This definition of an information retrieval system differs significantly from the belief that an information retrieval system merely informs its users on the existence (or non-existence) and whereabouts of documents relating to a request (van Rijsbergen 1979), and reflects the rapid development of this area of research in the past decades.

Despite the amazing advances in the search techniques and information re-

trieval, such an information system is still a long-term goal. There are several components that need to be thoroughly understood before such high accuracy information retrieval becomes possible. The first essential task is the understanding of the users' needs and the users' information seeking behavior. The second essential task is the understanding and processing of the raw data available to the information system. And finally, the third task is providing a framework for matching of the user's needs expressed in a request to an information retrieval system, and the processed data.

A not an unreasonable assumption is that many users' requests already are based on underlying questions. The questions could be very specific, or not quite yet clear to the users themselves. In the latter case the question is formulated as *what is known about X?* or even broader, *in what context can X be found?*, and the user's initial behavior amounts to browsing in the hope of consciously focusing the question and further formalizing the information need as a query. This assumption permits modeling high accuracy information retrieval as a question answering problem, with the understanding that an "answer" provided by a system may vary in form and function depending on the complexity and nature of the question, and the nature of user's information needs. Viewing question answering as a form of high accuracy information retrieval allows combining rich resources accumulated in both areas of research.

Question answering is rapidly advancing from identification of simple facts to understanding and reasoning in the context of complex scenarios. The complex questions arising in (or simulating) a realistic situation are interchangeably

called scenario-based questions. The development of automatic approaches to question answering takes place primarily in the framework of large-scale evaluations, such as the yearly cycle of Question Answering track of the Text Retrieval Conference (Voorhees 2003); question answering challenge at the NTCIR (Fukumoto, Kato, & Masui 2004); and the CLEF multilingual question answering track (Magnini *et al.* 2004). TREC, co-sponsored by the National Institute of Standards and Technology and U.S. Department of Defense, was started in 1992 to encourage research in information retrieval and increase the availability of appropriate evaluation techniques. From its beginning in 1999 until 2003 the Question Answering track, with its aim to promote returning information, i.e., answers rather than lists of documents in response to a statement of information need, focused on the so-called factoid questions. Answering closed-class factoid questions such as, for example, *How much fiber should you have per day?* involves either finding an exact short string, often representing an entity well studied in information extraction tasks, such as named entities (person, organization, location), temporal expressions, and number expressions, or somewhat more relaxed, a 250-byte passage of text. A variation on the factoid QA requires finding a list of short answers to questions of the type: *Name 10 autoimmune diseases.*

A first attempt at the large-scale complex question answering was the introduction of the definition questions in 2003 TREC evaluation. Definition questions ask for interesting information about an entity or event, for example, *What is TB?* At present, the most complex questions are the relationship questions introduced in the TREC-2005 evaluation. These questions request evidence in support of a

given answer, or the evidence itself, for example, *Has pressure from China affected America's willingness to sell high-tech weaponry to Taiwan?* Only six groups participated in this task and only three attempted to complete the task automatically. Solutions proposed to address the new challenge were for the most part to reduce it to problems for which the solutions are known: either by treating relationship questions as definition questions, or by decomposing relationship questions into several factoid questions. Both methodologies are somewhat limited: it is not clear if types of factoid questions correspond to the potential types of evidence; and if there is a one-to-one correspondence between interesting information and supporting evidence.

It is tempting to assume that an approach based on understanding of the user's needs, of the questions, and of the documents that potentially contain evidence, or in general, on understanding of the context, will lead to better answers. Testing the hypothesis that a common representation of a question and the documents will provide a means for complex question answering requires resources such as question taxonomies and tools for document understanding and analysis. Additional resources might be needed to provide a meaningful way to unify the question and the document representations. Due to limited resources and the lack of understanding which resources are needed the above hypothesis cannot be tested in open-domain question answering (answering any question that anyone might ask.) Fortunately there exists a domain with abundant freely available resources, deep understanding of users' tasks, needs, and behavior, and a well developed framework for manual complex question answering. This is the do-

main of medical informatics, which provides not only taxonomies of questions (Ely *et al.* 2000) and understanding of the user's needs (Gorman & Helfand 1995; Florance, Giuse, & Ketchell 2002), but also several models of the medical domain ranging from conceptual models for communication in electronic health records (Asp & Petersen 2003) to a comprehensive model of the clinical process (Sackett *et al.* 2000). The semantic domain model most widely accepted by clinicians and adopted in this dissertation is developed within the paradigm of Evidence Based Medicine (EBM). When speaking of the EBM paradigm, clinicians follow Thomas Kuhn, who views scientific paradigms as ways of looking at the world which define both the problems which can legitimately be addressed and the range of admissible evidence which may bear on their solution (Evidence-Based Medicine Working Group 1992). The medical domain provides all resources necessary to define the problems and formally describe the existing evidence, thus enabling testing the hypothesis that success in answering complex questions can be achieved by consistently relying upon an EBM-based semantic domain model at every step of the process. Development of a question answering framework based on the semantic domain model has a potential to assist in clinical decision making.

The knowledge-intensive question answering paradigm originated in artificial intelligence, independent of information retrieval approach. As early as 1959, preliminary experiments demonstrated that if both the question and the text potentially containing an answer could be coded semantically and syntactically, their common canonical form representations then could be matched to determine if the text indeed contains the answer (Russell & Norvig 2003). The most recent development in

this approach is question answering by predictive annotation (Prager *et al.* 2000). Predictive annotation is based on classification of all potential answer phrases by the type of a question that can be answered using the phrase. Questions are classified by the same scheme. Matching and selection of an answer is based on the distance (in words) between the question and each answer phrase. Although predictive annotation is successful in answering factoid questions, there are two potential limitations to this technique when dealing with complex questions: it is impossible to enumerate potential answers to all complex scenario-based questions that could be contained in a given text; and the full answer is most probably distributed among several documents and needs to be generated with respect to the context of the question. A lesser and potentially solvable drawback of this approach is its reliance on surface string matching rather than conceptual matching.

It seems that adding a layer of abstraction and extending the canonical form of the question and the text representation to encompass the user needs and the semantic domain model should address these limitations by advancing from the notion of the phrase capable of answering a simple closed-class question, to the level of creating a coherent semantic representation of the text capable of answering all types of questions characteristic of the domain.

This study uses a semantic domain model of clinical medicine to encode (a) a clinician's information need expressed as a question on the one hand and (b) the meaning of scientific publications on the other to yield a common representation. It is hypothesized that this approach will work well for (1) finding documents that contain answers to clinical questions and (2) extracting these answers from the

documents.

This work focuses on:

1. the use of medical domain knowledge and of the existing framework for formulation of clinical questions,
2. semantic domain-model motivated search for relevant documents,
3. extraction of important facts,
4. a hybrid approach to question answering that combines the classic AI frame-based approach with statistical NLP methods, and
5. synthesis and presentation of answers from multiple documents.

1.2 Contributions

Contributions of this thesis are both theoretical and practical and fall into two categories: Natural Language processing (NLP) and Medical Informatics.

NLP contributions:

- Design and development of a principled way to use the semantic domain model as a foundation for an end-to-end clinical question answering system.
- Identification of the semantic domain model components that allow for development of an end-to-end clinical question answering system.
- Demonstration of applicability of the system architecture for complex question answering in the clinical domain.
- Methodology for combining information extraction based on statistical methods and knowledge-based methods derived from the semantic domain model.

- Adaptation of question answering evaluation methods for the clinical domain.
- Development of annotation guidelines and test collections for information extraction and question answering evaluation.

Medical Informatics contributions:

- Development of CQA-1.0, a prototype Clinical Question Answering system.
- Development of information extraction modules that through ranking and clustering of medical documents facilitate clinical decision support, and can be used for extraction of information tailored to specific user needs, for example, a database of patient oriented outcomes.
- Evaluation of the impact of the semantic domain model components on the search for relevant documents.

1.3 Outline of Thesis

Chapter 2 introduces the semantic domain model of clinical medicine; reviews information needs of clinicians, and presents the framework for asking clinical questions, the nature of clinical questions and the nature of the document collection used in this study. This chapter also discusses related work on previous use of the semantic domain model in information retrieval and natural language processing. In addition, it provides a brief overview of information retrieval and application of natural language processing methods in medical domain. Finally, it reviews recent

advances in question answering.

Chapter 3 describes the domain resources used in implementing the CQA-1.0 system. Chapter 4 presents the architecture of the semantic-domain-model-based question answering system CQA-1.0, and design and implementation of the system's components. This chapter presents the three keystone components of the semantic domain model based on the principles of Evidence-Based Medicine (EBM), and algorithms that use these components as building blocks for the document selection and answer generation modules of the question answering system. This chapter concludes with the description of single-document and multi-document answer generation. A detailed description of each module, code, and pointers to third-party tools can be found at `http://www.umiacs.umd.edu/~demner/`.

Chapter 5 focuses on evaluation of the system components and the system as a whole. It presents a task-based expert evaluation of the system, and an expert evaluation of the document selection module. This chapter also discusses automatic evaluation methods adapted from other NLP tasks as well as widely accepted information retrieval evaluation metrics.

Chapter 6 concludes the thesis with the discussion of limitations of this work, of its potential impact, and of many remaining issues to be addressed in the future.

Chapter 2

Review of the Literature

This thesis draws on knowledge accumulated in several disciplines over considerable time periods, which makes presenting a comprehensive overview of each of the disciplines impractical and beyond the scope of this work. Rather, the goal of this chapter is to provide information necessary to understand the domain model and the user information needs; provide motivation for question answering in restricted domain; and reflect upon issues in information retrieval, natural language processing, and question answering that directly influence this work.

2.1 Information needs of clinicians

Information needs of health professionals have been actively researched for over two decades. Early surveys (Strasser 1978; Stinson & Mueller 1980) established a need for information on new developments in specialties and government regulations relating to health care. The most likely sources of information were journal papers, colleagues, and books. Subsequent observation studies focused on the types of information needs clinicians have and the number of questions arising due to attending to patients, as well as on the types and proportion of needs that are being pursued and on the most likely sources of information. In a review of the research of information needs of health professionals Smith (1996) identifies six

categorizes of information needs, provides sources of information for each category, and comments on the nature and importance of the needed information. The six categories of information needs identified by Smith are:

- information on particular patients,
- data on health and sickness within the local population,
- medical knowledge,
- local information on doctors available for referral,
- information on local social influences and expectations, and
- information on scientific, political, legal, social, management, and ethical changes affecting both how medicine is practiced and how doctors interact with individual patients.

Of these, medical knowledge is the focus of this thesis. Richardson and Wilson (1997) make a distinction between the background and the foreground medical knowledge. The background knowledge is a general knowledge of the basic facts of a disease. Information needs with respect to the background knowledge will generate information requests in the form of wh-type questions (who, what, where, when, why, and how) such as *What is this disorder?*, *What causes it?*, *What are the signs and symptoms?*, *What treatment options exist?* These questions are best answered from textbooks and regularly updated systematic reviews such as Cochrane reviews, a highly regarded source of evidence about the effects of healthcare interventions. These resources compiled by specialists are often called secondary sources, or secondary literature, as opposed to the primary sources that report original results of the clinical research. More often, practicing clinicians lack the foreground knowledge: information about choice of therapeutic interventions, the best diagnostic

test for a disease, or the best treatment strategy for a particular patient. These questions might be answered by systematic reviews, if the question was common enough to warrant a review, and if the review was updated recently. When no secondary sources are available, the online databases that provide access to results of clinical trials and observations are a potential source for answers. This assumption motivated studies striving to answer three questions:

- Is the need for answers substantial enough to warrant research of the best ways to answer the foreground questions?
- Do the existing resources contain information capable of answering these questions?
- To what extent is the available information accessible and used by clinicians?

The first question was answered positively. Interviewing and observing primary care physicians, researchers found information needs ranging from 0.33 per patient (Dee & Blazek 1993) to 2 per patient (Timpka, Ekström, & Bjurulf 1989). Other studies found 0.66 needs per patient (Covell, Uman, & Manning 1985), 0.57 needs (Gorman & Helfand 1995), and 0.42 needs (Cogdill & Moore 1997). Ely et al. (1999) found 0.32 needs when only foreground questions were considered.

The second question was studied in several controlled experiments in which physicians or medical librarians were given a set of foreground clinical questions and were asked to find answers using medical literature and online databases. In a study conducted by Gorman et al. (1994) experienced medical librarians deemed 88% of the clinical questions appropriate for MEDLINE, and found information judged relevant by physicians for 56% of the questions. Clear answers were found for 46% of those questions. Clinicians evaluated 40% of the answers as having impact on their

patients, and 51% of the answers as having impact on themselves and their practice. Similarly Giuse et al. (1994) answered 87% of clinical questions generated from patients' charts using online bibliographic databases. Chambliss and Conley (1996) found answers to 54% of the questions, 71% of these in MEDLINE. In a study that limited searches of online databases to 10 minutes, two experienced physicians found answers to 75% of the questions (Alper *et al.* 2001). Koonce et al. (2004) found secondary sources capable of answering 20% of the foreground questions and 47.5% of the background questions, which demonstrates the importance of the primary medical literature.

Knowing that online databases can potentially answer at least half of the clinical questions, researchers measured to what extent clinicians use these resources. De Groote and Dorsch (2003) in a survey of medical students, residents and faculty found 53% of the users search MEDLINE at least once a week, 72% of the survey responders used the online resources for patient care. The online databases, however, lag behind such sources as colleagues, books, journals, and newsletters (Ely *et al.* 2005). The key findings of the national audit of Clinical Question Answering Services conducted by Doctors.net.uk (Bryant 2005) indicate that 78% of the doctors consult colleagues as their first source in answering clinical questions. Gorman and Helfand (1995) found that doctors had pursued less than a third of the questions and found answers for less than a quarter. Obstacles in answering clinical questions were studied by Ely et al. (2005) who found that physicians pursued answers only to 55% of the questions due to failing to recognize an information need, uncertainty regarding existence of an answer, preference for convenient rather than appropriate

resources, and lack of skills formulating questions and search strategies. The latter issue is addressed in the practice of Evidence Based Medicine.

2.2 The Evidence-Based Medicine domain model

Evidence-based medicine (EBM) is a widely accepted paradigm that formalizes approaches to bridging the gap between the care that a patient will get and the best possible care in a given situation as determined by systematic research. Practicing evidence-based medicine involves integrating individual clinical expertise with the best available external clinical evidence when making a clinical decision. Addressing some reservations with respect to EBM used blindly in the individual patient care (Cohen, Stavri, & Hersh 2004), Sackett (2000) points out that neither component alone is sufficient: evidence cannot be used appropriately without expertise, and expertise without evidence becomes dated and potentially detrimental to a patient's health. Incorporating the best available research evidence in decision making involves: defining the question; finding the best information; appraising the information for validity and relevance; and summarizing the information (Rosenberg & Donald 1995). Booth and O'Rourke (1997) suggest that these guidelines for doctors seeking evidence roughly correspond to a reference interview and can be used in automatic question answering. Detailed studies of means to achieve success in seeking best evidence permit operationalizing the above guidelines, which are implemented in the proposed CQA-1.0 system based on the semantic domain model as follows: given a user's question, the CQA-1.0 system finds relevant documents,

estimates their validity and relevance to the question, and presents information in a multi-tiered answer.

The steps automatically performed by the CQA-1.0 system are based on three fundamental components identified in the EBM-based semantic domain model: clinical task, framework for question formulation and document appraisal, and strength of evidence. EBM identifies four major clinical tasks involved in decision making: etiology, diagnosis, therapy, and prognosis. An information need may occur while a clinician performs any of the tasks. Should the need arise, it should be clearly defined and well articulated. EBM provides guidelines for formulating clinical questions and translating these questions into successful information requests. For each of these clinical tasks, there are four elements of a well-formed question:

P Who/What is the **P**atient/**P**roblem being addressed?

I What is the intended **I**ntervention?

C What is the intervention **C**ompared to?

O What are the **O**utcomes?

The clinical question structure known by the first letters of the elements of the question frame as PICO plays a role not only in the question and search formulation, but also in appraisal of the information. The third important component of the EBM-based semantic domain model is the strength of the found clinical evidence. The strength of evidence in a clinical study is determined based on soundness of its design, number of patients participating in the study, and methods used to evaluate the results of the study.

Detailed descriptions of the three components of the EBM-based semantic domain model follow.

2.2.1 Clinical tasks

Physicians' everyday activities require familiarity with common health problems and adequate diagnostic, therapeutic and preventive services. Sackett et al. (2000) identified eight task categories that describe these activities: clinical findings, diagnostic tests, etiology, differential diagnosis, prognosis, treatment, prevention, and self-improvement. The Task Force rating scheme (AHRQ 2002) for assessing the quality of clinical evidence combines these tasks into the following four:

Etiology/Harm – identifying causes for diseases or conditions

Diagnosis – encompasses clinical findings, diagnostic tests, and differential diagnosis

- **Differential diagnosis** – identifying and estimating likelihood of potential causes for patient's condition
- **Diagnostic test** – selecting and interpreting diagnostic tests, considering their precision, accuracy, acceptability, cost, and safety

Therapy – select treatments that are worth the efforts and costs of using them (includes **Prevention** - actions to reduce the chance of disease by identifying and modifying risk factors)

Prognosis – estimate the patient's likely course with time and anticipate likely complications

It is worth mentioning that the distinction between the differential diagnosis and etiology questions is fine and not always obvious. For example, in a scenario-based assessment of physicians' information needs, *What is anemia?* was classified

as a question about diagnosis because the patient's test results showed abnormal findings indicating anemia, but *What is the cause of gastritis?* was classified as an etiology question because gastritis was not present in the test results (Seol *et al.* 2004). The task of self-improvement identified by Sackett et al. (2000) is one of the goals of the evidence based medicine practitioners. It is omitted in the AHRQ rating scheme, possibly because any questions of improvement arising during the practice of medicine fall into one of the above four categories. This task pertains to improvements in clinical practice through improving clinicians' learning skills, as illustrated in the following question: *To improve my understanding of the pathophysiology of ascites would I gain more from spending an hour in the library reading a textbook or spending 15 minutes on the ward computer looking at the CD ROM version of the same textbook?* (Straus & Sackett 1998) . An answer to this question depends on many factors outside of a clinical scenario. The goal of this work is to facilitate the learning process rather than answer the above question.

2.2.2 A framework for synopsis of a clinical scenario (PICO)

To standardize and improve the process of formalizing information needs of clinicians, Richardson et al. (1995) proposed PICO, a framework for constructing well-formulated questions. These questions identify the patient and/or problem, a planned intervention (e.g., a treatment or a diagnostic test), a desirable outcome of the intervention, and, if applicable, a comparison intervention as follows:

Patient (population)/**P**roblem: identify information about an individual patient or a group of patients (population) and the problem that needs clinicians' care.

Such information is routinely obtained during collection of preliminary case history and complaints of a patient and the subsequent diagnostic work-up.

Intervention: is the procedure, agent or other clinician's act of interfering with a condition to modify it or with a process to change its course that is being administered to either a single patient or a group of patients.

Exposure: is an alternative to the intervention slot of the PICO frame developed to accommodate the etiology (harm) questions. The exposure also reflects the act of interfering with a patient's condition, but in this case, the actions are that of a harmful agent, for example, prenatal exposure to cigarette smoking.

Comparison: provides a frame of reference for an intervention, for example, an alternative intervention, a different method of administration or pattern of dosage or a different timescale. Patient outcomes of a planned intervention might be measured against a comparison.

Outcome(s): summarize the effect of an intervention or an exposure on a patient or population, focusing on patient oriented outcomes such as few side effects, increased survival rates, restoration of functions, etc.

Questions containing these components are thought to be more "answerable."

Although the question construction process was initially developed to answer questions arising with respect to therapy, it was later adapted for all clinical tasks in a series of articles published in the Journal of the American Medical Association (JAMA) (Guyatt, Sackett, & Cook 1994; Jaeschke, Guyatt, & Sackett 1994; Levine *et al.* 1994; Laupacis *et al.* 1994). Each article focuses on constructing PICO representations of clinical questions for a major clinical task. The examples are given below in the order in which they are presented in the JAMA series.

2.2.2.1 PICO frame for Therapy or Prevention

Scenario: A 65 year-old man with controlled hypertension and a history of non-valvular atrial fibrillation resistant to cardioversion wants to know whether the benefits of long-term anticoagulants (to reduce the risk of embolic stroke) outweigh their risks (of hemorrhage from anticoagulant therapy). (Guyatt, Sackett, & Cook 1994)

The authors selected *non-valvular atrial fibrillation* from the detailed description of the patient and his problem to populate the *Problem* slot of the frame. Based on the background knowledge, the *Intervention* slot was populated with *warfarin*. Trustworthy evaluations of drug effectiveness are based on comparisons with other drugs known to be effective, or no treatment that often amounts to placebo treatment. In this case authors recommend populating the *Comparison* slot with *placebo*. Finally, there are several *Outcomes* of interest: risk of emboli (including embolic stroke) and the risk of the complications of anticoagulation.

2.2.2.2 PICO frame for Diagnosis

Although the JAMA article (Jaeschke, Guyatt, & Sackett 1994) discusses a PubMed search strategy and identifies important concepts that need to be included, there are no direct recommendations for assigning concepts to the PICO frame slots. A recent article (Zakowski, Seibert, & VanEyck 2004) offers the following example:

> *Question: accuracy of an increased respiratory rate to detect pneumonia in children presenting to a clinic with respiratory symptoms*

P: In children with upper respiratory symptoms

I: is measuring the respiratory rate

C: as effective as a chest x-ray

O: in detecting pneumonia?

However in an EBM training course offered by the British Medical Academy (BMA)[1] the suspected problem fills the problem slot and the outcome slot remains empty:

[1] `http://www.bma.org/ap.nsf/Content/LIBSeeKEvidenceMedline`

Q: What diagnostic tools are available for the screening for prostate cancer in young males, and how effective are they?

Population /**P**atient /**P**roblem: Prostate cancer

Intervention: Screening

Other limits: Male Adolescent Adult

This frame instantiation not only illustrates the difficulties in finding an appropriate slot for a hypothesized disease, it also indicates the need in separating the Patient and the Problem descriptions by placing patient's gender and age into a separate field.

2.2.2.3 PICO frames for Etiology

The JAMA article on etiology questions (Levine *et al.* 1994) describes an asthmatic patient asking his doctor about an increased risk of death associated with beta-adrenergic agonists in the treatment of asthma, but does not provide a filled PICO frame. The BMA tutorial offers the following example:

Question: What is the risk of psychiatric illness on taking the antimalarial drug Mefloquine?

Population /**P**atient /**P**roblem: Psychiatric illness

Intervention: Mefloquine

According to this example, in the original JAMA etiology question, asthma should be assigned to the population/problem slot frame; beta-adrenergic agonists to the intervention; and adverse effects and death to the outcome slot of the frame. There is however a difference in assigning an existing disease - asthma to the problem slot, and a potential result of exposure to the drug to the same slot, as in Psychiatric

illness. In the asthma example, beta-adrenergic agonists are an intervention with respect to asthma, and an exposure with respect to the patient outcome, in this case, risk of death. What if the patient outcome is an onset of a new pathologic condition, for example psychiatric illness? Such EBM sources as The Pediatric Residency Curriculum Handbook, University of Illinois at Chicago, recommend placing a potential disease into the outcome slot:

Patient/**P**roblem: Controlling for confounding factors,
do otherwise healthy children

Intervention/Exposure: exposed in utero to cocaine

Comparison: compared to children not exposed,

Outcome: have an increased incidence of
learning disabilities at age six years?

2.2.2.4 PICO frame for Prognosis

Prognosis refers to the possible outcomes of a disease and the frequency with which they can be expected to occur. In the JAMA example (Laupacis *et al.* 1994), the question for prognosis is a son's inquiry about an Alzheimer's patient prognosis, and whether she is likely to die soon. In this case, Alzheimer's disease fills the problem slot, and death - the outcome slot. There are controversial recommendations in filling out PICO frames for prognosis questions as well. The Chicago handbook provides the following example:

Patient/**P**roblem: In children with Downs Syndrome,

Intervention: is IQ an important prognostic factor

Outcome: in predicting Alzheimer's later in life?

The BMA tutorial suggests placing the motor neuron disease into the problem slot for the question: *What is the short / long term prognosis for a young adult recently diagnosed with motor neuron disease?*

Once the question is formulated, the JAMA series of articles recommends searching PubMed using all terms in the PICO frame. The articles recommend augmenting the search with a description of the clinical task, for example, prognosis for prognosis questions, and cause for etiology questions. In addition to the question and search formulation strategies, JAMA articles provide a framework for assessment of relevance and quality of the articles.

2.2.3 Quality of research / Strength of evidence in medical articles

The key points in evaluating articles are: whether the results are valid, and whether the information is relevant to the patient's condition that led to the question. The second point is verified using the PICO structure. The closer the problem, patients, interventions, and outcomes described in the article are to the clinical scenario, the more might be its significance to the patient.

Flaherty (2004) proposes a *PP-ICONS* method for physician's evaluation of the clinical literature. The *P(P)ICO* structure is augmented with N - the number of subjects in a research study, and S - statistics used to analyze the study. The method is illustrated using the clinical question: *Is duct tape an effective treatment for warts in children?* that arises after seeing a nine-year-old patient with common warts on her hands. Her mother had heard about treating warts with duct tape and asks

The efficacy of duct tape vs cryotherapy in the treatment of verruca vulgaris (the common wart).
OBJECTIVE: To determine if application of duct tape is as effective as cryotherapy in the treatment of common warts. DESIGN: A prospective, randomized controlled trial with 2 treatment arms for warts in children. SETTING: The general pediatric and adolescent clinics at a military medical center. PATIENTS: A total of 61 patients (age range, 3-22 years) were enrolled in the study from October 31, 2000, to July 25, 2001; 51 patients completed the study and were available for analysis. INTERVENTION: Patients were randomized using computer-generated codes to receive either cryotherapy (liquid nitrogen applied to each wart for 10 seconds every 2-3 weeks) for a maximum of 6 treatments or duct tape occlusion (applied directly to the wart) for a maximum of 2 months. Patients had their warts measured at baseline and with return visits. MAIN OUTCOME MEASURE: Complete resolution of the wart being studied. RESULTS: Of the 51 patients completing the study, 26 (51%) were treated with duct tape, and 25 (49%) were treated with cryotherapy. Twenty-two patients (85%) in the duct tape arm vs 15 patients (60%) enrolled in the cryotherapy arm had complete resolution of their warts (P =.05 by chi(2) analysis). The majority of warts that responded to either therapy did so within the first month of treatment. CONCLUSION: Duct tape occlusion therapy was significantly more effective than cryotherapy for treatment of the common wart.

Figure 2.1: MEDLINE abstract for PP-ICONS analysis.

about this treatment as an alternative to the usual cryotherapy. The PP-ICONS analysis illustrates physician's reasoning applied to the abstract in Figure 2.1 (Focht, Spicer, & Fairchok 2002).

The analysis starts with the problem. The abstract indicates that the researchers studied the same problem (the same type of warts.) If the investigated problems were not sufficiently similar to the girl's clinical problem, the results would not be relevant, and there would be no need in further analysis of the article. Next, the patient or population is considered. Is the study group similar to the patient in terms of age and gender, and is the clinical setting similar to the doctor's practice? If the patients in the study are not similar to the patient, the results might not be

relevant. The difference between the absolute constraint of the problem matching and the preference constraint of the patient matching is illustrated in this example: the age of the study participants ranges from 3 to 22 years, which includes, but is not limited to the girl's age, and no gender distribution is given. The clinical setting is also similar, but not identical to the doctor's practice. Next, the doctor needs to verify that the interventions in the article and in the question are the same. The question was about effectiveness of the duct tape for warts compared to cryotherapy, so this is a relevant study. In the outcome analysis physicians are particularly interested in the outcomes that their patients care about: symptoms, morbidity, quality of life, mortality, etc. Reports of such *patient-oriented evidence that matters (POEMs)* (Slawson & Shaughnessy 2000) indicate whether an intervention offers a true clinical benefit, as opposed to the surrogate such as tumor shrinkage or changes in cholesterol level, blood pressure, or other laboratory measures. The latter *disease-oriented evidence (DOE)* reflects changes in physiologic parameters. Until recently, physicians assumed that improving the physiologic parameters results in a better disease outcome, but this is not always the case. Therefore outcome measures such as *complete resolution of the wart* in the above article are considered more relevant than the ones based on DOEs. This outcome is something the patient is interested in, and therefore it satisfies the doctor. The number of subjects indicates whether the sample size was large enough to detect a clinically meaningful difference between the intervention and comparison groups. Studies with less than 100 subjects are usually considered inadequate to provide reliable statistics, which makes the wart study completed by fifty-one patient too small to generate good statistics. The

statistics analysis is twofold. First, the physician analyzes whether the statistics reported in the study are trustworthy. The effect of intervention is usually measured by comparing the probabilities of events in the control and treatment groups. For example, the absolute risk reduction is the difference in the probabilities of an event in the two groups of patients. The PP-ICONS method recommends the Number needed to treat (NNT) as the most reliable statistic. This statistic measures the number of patients that must be treated to prevent one adverse outcome or for one patient to benefit (Cook & Sackett 1995). The second aspect of the analysis of the intervention effect is needed to quantify the strength of evidence. This involves tests of significance or confidence intervals. In the final step of the PP-ICONS analysis the doctor concludes that the statistics, particularly the NNT, are reasonable, and accepts the approach as fair and worthy of a discussion with the patient's mother.

Clinicians are advised to select articles based on the potential strength of evidence, and then apply the above evaluation to promising articles. In 2002, the AHRQ identified seven systems that fully address grading the strength of a body of evidence in terms of quality, quantity, and consistency. The Strength of Recommendations Taxonomy (Ebell *et al.* 2004) builds upon the key issues identified in the AHRQ report, and explicitly addresses the issue of patient-oriented versus disease-oriented evidence. The taxonomy makes a distinction between the strength of a body of evidence and the quality of individual studies. A body of evidence could be of one of three grades:

grade A evidence is consistent and good-quality patient-oriented evidence. For example, evidence from meta-analysis of multiple, well-designed, controlled studies or from high quality Randomized Clinical Trials (double-blinded, randomized clinical trials.)

grade B evidence is based on inconsistent or limited-quality patient-oriented evidence. For example, non-randomized, controlled or cohort studies, matched case-controlled studies or cross-sectional studies.

grade C is disease oriented evidence, or evidence based on consensus, usual practice, case series, comparative, descriptive and case studies, case reports, clinical examples, and/or opinion.

Level of evidence in individual studies is determined similarly.

Level-1 evidence can be found in systematic reviews, meta-analysis and high quality RCTs (Randomized Clinical Trials) and cohort studies.

Level-2 evidence can be found in clinical trials and cohort studies without appropriate blinding and allocation, inadequate size, and insufficient follow-up.

Level-3 evidence comes from the studies that form group C in the body of evidence.

The recommendations described above have been developed to help clinicians apply the results of medical research into clinical practice by effectively finding and analyzing clinical articles. However the domain knowledge captured by the practitioners of EBM attracted attention of the researchers who employed various elements of the EBM-based domain model in development of information retrieval and expert question answering systems.

2.3 Prior use of the EBM-based domain model in Information Retrieval and Natural Language Processing

Each component of the EBM-based model has been explored in information retrieval separately and to a different extent. One of the first successful applications of one of the three basic components is a clinical-task specific query expansion. This research is actively pursued by the Hedges Project (Wilczynski, McKibbon, & Haynes 2001). The accuracy of the manually constructed search strategies for therapy, diagnosis, review, prognosis, causation (etiology), economics, cost, and clinical prediction guides was analyzed on 49,028 articles indexed for MEDLINE. The best performing terms were selected from 4,862 unique terms. These terms are available for query expansion in the form of Clinical Queries in PubMed, a service

of the U.S. National Library of Medicine that provides access to over 16 million citations from MEDLINE Database. These search strategies were used by Mendonça and Cimino (2001) to retrieve an initial set of 4,000 MEDLINE citations, from which additional terms associated with the four main clinical tasks were automatically extracted using hierarchical and semantic links in the Medical Entities Dictionary (Cimino *et al.* 1994). In the subsequent manual evaluation, 60% of the additional terms were found relevant to the corresponding task, with the best results for therapy and the worst for prognosis. The search strategies developed by the Hedges Project also serve as a foundation for complex search strategies developed to retrieve initial sets of documents for the Family Practitioner Inquiry Network (FPIN) database of answers to clinical questions (Ward, Meadows, & Nashelsky 2005). Pratt and Wasserman (2000) achieved 0.68 precision in categorizing queries into clinical tasks based on lexical and semantic analysis. Their ten query types extend beyond the conventional four tasks, and even the eight identified by Sackett to include prevention, risk factors/etiology, diagnostic tests, diagnosis, symptoms, treatment, side effects, prognosis, overview, and other.

The effectiveness of the PICO framework in its intended use (applied to query formulation) was evaluated by Booth et al. (2000), who compared data from 185 PICO-structured forms with 195 minimally structured forms on the axes of information elicitation, precision of search results, and acceptance by librarians participating in the study. PICO-structured forms yielded more detailed searches and statistically significant improvement in precision, but were rated lower by librarians. Similarly, in a survey of searches with handheld devices users evaluated positively the results

of the searches, but were divided evenly into those who found the PICO form easy to use and useful, and those who did not find it useful (Fontelo *et al.* 2005). The acceptance and effectiveness of the standard PICO query can be improved by specific instructions (Villanueva *et al.* 2001).

An interesting validation of the PICO framework comes from a study of informal consultations between 60 primary care physicians and 30 specialty physicians. In this study, e-mailed questions were less likely to remain unanswered if they identified a proposed intervention and a desired outcome. The presence of a comparison had no effect on the answer (Bergus *et al.* 2000).

Booth and O'Rourke (2000) pioneered application of the PICO framework to retrieval of documents and found that structuring abstracts according to PICO improved the precision of search for clinical questions when compared with unstructured single paragraph abstracts. Niu and Hirst (2004) applied PICO framework to search a database of reviews that summarize and appraise clinical evidence, and reported preliminary results of outcome identification in these reviews. However it is unlikely that findings from a study of peer-reviewed compilation for 200 medical conditions created by a limited number of specialists will scale to MEDLINE abstracts. As an example, a very specialized source permits using terms like *comparison* and *dependency* as indicators of patient outcomes. However, the term *comparison* can be found in 417,589 MEDLINE abstracts, often only in the title, e.g. *Comparison of preoperative anxiety in reconstructive and cosmetic surgery patients.* In a more general database, these terms lose their predictive power suggesting that simple cue words are only the beginnings of a solution.

The third component of the EBM-based model, the strength of evidence, has been available to MEDLINE users since 1991 in the form of manually assigned Publication Type controlled vocabulary terms, for example, search can be restricted to the strongest evidence in the form of meta-analysis and randomized clinical trials. Several studies exploit this component of the EBM-based model in document ranking. This component is implicitly taken into consideration in summarization (Fiszman, Rindflesch, & Kilicoglu 2004) that retrieves only abstracts with the most reliable publication types as the first step of the process. Similarly, McKeown et al. (2003) personalize search results to a patient profile taking into account whether a document is a "clinical study" as judged by a categorizer trained on publication types and other features of 7000 MEDLINE citations.

2.4 Information Retrieval and Natural Language Processing in the medical domain

Information retrieval research in biomedicine parallels the research on information retrieval in general. Some of the first studies were conducted at the National Library of Medicine in anticipation of computer-based retrieval systems. For example, Winifred Sewell (1964) reviewed the library's controlled vocabulary used for indexing in anticipation of the Medical Literature Analysis and Retrieval System (MEDLARS). Online access to a subset of references in the MEDLARS database became available in 1971 in the form of MEDLINE (MEDLARS Online). Sewell regarded the controlled vocabulary terms, called Medical Subject Headings, "as di-

rectional signals or vectors which, with other headings, serve to locate the essence of a particular paper or book in the universe of medical information." She pointed out that through greater coverage and deeper indexing a computerized system would increase the need for specificity in descriptors and delineation of hierarchical relationships useful for search purposes. Sewell identified four broad groups that encompass the majority of the vocabulary terms: *Anatomical Terms*; *Organisms*; *Diseases*; and *Chemicals and Drugs*. The desired level of specificity for adding concepts to the top level hierarchies was determined by frequency of appearance of concepts in journal articles and by the ability of a specific term to retrieve about a 100 citations from the 1961 collection of articles. According to a theory of sublanguages proposed by Harris, such computation oriented corpus-based definition of the controlled vocabulary is possible due to the structure and regularity of technical languages (Friedman, Kra, & Rzhetsky 2002).

This notion of the special biomedical sublanguage determined several directions in domain-specific information retrieval explored in addition to the mainstream research techniques applied to specialized literature. Domain specific research is largely influenced by existing resources such as curated databases of genomics information, for example Gene Ontology, and MEDLINE.

2.4.1 Document /Discourse structure utilization

Biomedical journal articles differ from documents comprising the widely used Information Retrieval (IR) collections (such as TREC collections that primarily in-

clude news stories) not only in the specifics of the biomedical sublanguage but also in their structure. Medical articles combine the generic structure of scientific articles (Bishop 1999) with specific domain elements. Purcell et al. (1997) captured the complex structure of medical articles in three hierarchical context models (clinical research articles, case reports, and reviews) for medical document representation, and identified a number of elements, which the authors call contexts, that characterize each of the structures. Purcell et al. proposed annotating each sentence in an article with a context, for example, experimental findings in the results section of a clinical research article, for the purposes of context-based information retrieval, and reported significant improvement in precision of full-text searching at fixed levels of recall.

Explicitly labeling structure not only in the full text of the articles, but also in the abstracts was suggested by the Ad Hoc Working Group for Critical Appraisal of the Medical Literature (1987) to support the assessment of reliability and content of a clinical report by readers and reviewers, and to aid accurate indexing and retrieval. Using this structure for indexing and giving more weight to some of the structural elements, for example, to titles or the *Purpose* and the *Conclusion* sections of an abstract has been shown to improve ad hoc retrieval results (Aronson *et al.* 2004a), and finding related citations (Tbahriti *et al.* 2004).

2.4.2 Domain knowledge utilization

Domain knowledge resources are widely used in the production information retrieval systems (PubMed, Ovid), in natural language processing (NLP), and in information retrieval research. The following sections present some of the key domain-knowledge-based techniques.

Query refinement

In addition to query refinement for clinical search developed by the Hedges project, many researchers studied query refinement techniques known to be useful in ad hoc retrieval. Srinivasan (1996) compared three sources for blind relevance feedback (query expansion using additional terms from the initial set of retrieved documents (Harman 1992).) In Srinivasan's study, initial search was expanded using only controlled vocabulary terms, only free text terms, and a combination of the two sources. Improvements in average precision at 11 standard recall points in the range of 9% to 119% were achieved for individual queries, with an overall improvement of 16%, primarily due to the controlled vocabulary feedback. Aronson and Rindflesch (1997) achieved 14% improvement in average precision through query expansion using automatically identified controlled vocabulary terms that were then expanded using inflectional variants from the Specialist lexicon (Browne *et al.* 2003) and synonyms encoded in UMLS (Lindberg, Humphreys, & McCray 1993). Hersh and Hickam (1994) have shown improved average precision using MeSH terms in searches of MEDLINE citations over the searches of the titles and abstracts of

the citations. Shatkay and Wilbur (2000) developed a probabilistic method for determining a theme – a set of terms from documents discussing a common topic that can be used either for query expansion (Aronson *et al.* 2004a) or for retrieval of documents by example as implemented in the Related Articles feature of PubMed.

Concept (entity) identification

The importance of MeSH terms and the slowness and cost of the current manual indexing process led to a fair number of studies of automatic extraction of the UMLS concepts from medical text. MetaMap (Aronson 2001), the most accurate and comprehensive method, is described in Section 4.4.1. Other methods include mapping of query terms into MeSH terms through a common semantic representation based on 3400 simple atomic concepts such as "heart" (Zieman & Bleich 1997), restricting UMLS concept matching to noun phrases (Denny *et al.* 2002), or first generating all possible UMLS concepts for each of the text tokens and then applying syntactic and semantic filters to eliminate irrelevant candidates (Zou *et al.* 2003). A related effort – protein and gene name identification – received much attention recently, following the growing interest in the "omics". "Omics" include genomics – study of a living organism in terms of the sequence of its genome; proteomics – study that focuses on identification of physiological roles of proteins and their structure; and a relatively new field metabolomics – study of metabolites that represent the end product of gene expression. Entity identification methods for "omics" include dictionary look-up, rule-based term recognition, machine learning, and hybrid ap-

proaches. A comprehensive review of gene and protein name recognition techniques is provided in (Krauthammer & Nenadic 2004).

Semantic indexing

One of the first specialized retrieval systems that implemented automatic concept-based indexing and extraction of the UMLS concepts from users' requests was SAPHIRE (Hersh & Greenes 1990). SAPHIRE utilized the UMLS Metathesaurus by breaking free text into individual tokens and constructing a list of Metathesaurus terms for each token. The terms were than weighted based on their length, overlap with the original text, and the proximity of the original tokens to each other. When compared with regular MEDLINE searches performed by physicians and experienced librarians, SAPHIRE performed equally well for physicians, but was outperformed by librarians using MEDLINE (Hersh *et al.* 1994). Chen et al. (2003) augment noun phrase indexing with automatic thesaurus generation in the HelpfulMed system, a Web portal that provides information retrieved from reliable medical domain sources with minimal manual effort. HelpfulMed also provides several presentation modes of retrieval results: in a traditional ranked list, in a self-organizing map, and in a list of automatically derived concepts, MeSH terms, and authors, which gives a user an opportunity to search phrases extracted from the text, related medical subjects headings, authors, or any combination of the three, thus accommodating users with different information needs and tasks.

Organization of retrieval results

One of the problems with searching medical literature is that in many cases the searchers are interested in all relevant publications, which sometimes amounts to a large number of documents. For example, in a recent evaluation of top 40 citations retrieved to answer 5 questions of the type *What is the best treatment for disease X?* conducted by a primary care physician, on average, 87% of the citations were at least topically relevant and 61.8% were judged as good quality research containing parts of an answer (Sneiderman *et al.* 2005). The combination of the amount of relevant information and the need for comprehensive coverage of the available data motivated research into organization of retrieval results. The basic principle for browsing and searching postulated as "Overview first, zoom and filter, then details-on-demand" (Shneiderman 1997) has been implemented in several specialized systems. Pratt et al. (1999) developed DynaCat, a knowledge-based system for dynamic organization of retrieval results into a hierarchy based on UMLS relations. Selection of the category labels and of the UMLS subset is based on the user's query. The Vivisimo service founded by Carnegie Mellon University researchers did not publish the details of the clustering algorithm and the label generation implemented to organize PubMed search results. Labels generated by the Vivisimo online demonstration appear to be extracted from the documents rather than looked up in the UMLS. Organization of MEDLINE retrieval results in the MEDLINE Database On Tap application is knowledge-based. The results are organized either into subject areas used for indexing journals according to discipline, or into strength

of evidence groups based on publication types (Demner-Fushman *et al.* 2004).

Another approach to reduction of the user's cognitive load is multi-document summarization. The most comprehensive study to date of multi-document summarization of medical articles was conducted in the PERSIVAL project. In this study, the summarization module used patients' records to provide personalized summaries in response to physicians' questions (McKeown, Elhadad, & Hatzivassiloglou 2003). The patient's record and the article are represented using vectors of UMLS concepts that fill out templates with the following slots: parameters, relation, dependence, and finding. Merging of identical results occurs on the template level. Another successful example of multi-document summarization is a graphical representation of conceptual condensates of retrieved documents (Fiszman, Rindflesch, & Kilicoglu 2004). Condensate generation utilizes relations identified by SemRep (Rindflesch & Fiszman 2003). SemRep is described below in section 3.6.

2.5 Question Answering

Question answering encompasses psychology, philosophy, linguistics, education, computer and library science. As a consequence, studies of the artificial intelligence, in particular natural language processing, and information retrieval aspects of question answering benefit from knowledge acquired in other disciplines. Philosophy and psychology provide insights into modeling of the question answering process. According to Singer's (2003) review of the theories contributing to understanding of the process, its first stage is the encoding of the question meaning. Singer fol-

lows Kintsch's tradition in encoding questions as propositions. He also points out that successful question comprehension and answering depend on understanding of which parts of the question contain information known to the person asking the question, and which part is the request for new information, i.e., the focal idea of the question. Identification of the question focus is based on the listener's knowledge. Lehnert (1977) illustrated the importance of finding focus in an implementation of a prototype question answering system SAM. SAM, which attempted approximating human cognitive process, answered questions about stories depicting eating in a restaurant. It used a sentence analyzer, a script application mechanism to create a memory for the story, and procedures for locating answers to questions in its memory. This prototype represents the first generation AI question answering programs. One of the first and well known representatives of such systems is LUNAR (Russell & Norvig 2003), a natural language interface to a database of chemical analysis data of lunar samples brought back by the Apollo mission, which in one test, answered 78% of the questions asked by geologists. A fair number of the first generation AI programs were developed for clinical decision support. The clinical decision support systems were primarily rule-based expert systems mimicking consultants in very restricted areas. One of the first medical expert systems is MYCIN (Shortliffe 1981), a system capable of diagnosing and recommending treatment for infections, and explaining its reasoning. The rules for the system were derived from observation and interviewing of experts, and took into consideration incompleteness and inexactness of information. The first expert systems were followed by many clinical decision support systems, such as PIP and INTERNIST-1 (Perry 1990). In a recent review

of the existing decision support systems, Garg et al. (2005) identify four types of systems:

- systems for diagnosis
- reminder systems for prevention
- systems for disease management
- systems for drug dosing and drug prescribing.

Even if these systems were capable of answering questions, a family doctor would need access to systems that are constantly updated in response to the changing body of knowledge, and also to a different system for each question type, since the support systems are designed to solve only one problem and often pertaining to one disease. These drawbacks by no means undermine the contribution of the first generation systems to the current state-of-the-art in question answering. Many techniques widely applied today were tested in these systems. For example, in his 1964 paper Cooper (1964) argues that question answering systems, which he calls "Fact Retrieval systems", should store information and accept queries in a natural language. He also suggests document retrieval could be used to find portions of stored information most relevant to the question. He then presents an internal language used by the system for inference, and syntactically motivated selection of controlled vocabulary for the internal representation of the stored information.

The generic architecture for question answering explored and tested in the above systems was reviewed by Hirschman and Gaizauskas (2001). It accommodates most of the current state-of-the art systems that retrieve answers in three basic steps:

- Query formulation based on the processing of the incoming question.

- Relevant document selection and answer extraction.
- Filtering of the retrieved results and ranking of the answers.

Although this system architecture is rooted in closed domain question answering, the most recent advances lie in the open domain question answering that focuses on returning brief answers, such as names, dates, locations, etc., to natural language questions on any topic.

2.5.1 Open Domain Question Answering

Moldovan et al. (2003) provide a broad taxonomy of existing open domain QA systems based on the following criteria:

(a) linguistic and knowledge resources

(b) natural language processing involved

(c) document processing

(d) reasoning methods

(e) assumptions about answers being explicitly stated in documents

(f) necessity to generate answers

The five classes of systems identified in this taxonomy are strongly associated with the types of questions and the available test collections.

Class 1 systems are capable of processing factual questions and typically extract answers using keyword matching.

Class 2 systems use semantic alternations, world knowledge axioms and simple reasoning to relate snippets of text containing answers with the questions.

Class 3 systems generate answers to list, script, or template-like questions from parts found in several documents.

Class 4 interactive systems answer questions in the context of previous interactions, which involves complex reference resolution.

Class 5 systems answer speculative questions, which involves knowledge extraction from relevant documents and case-based, temporal, spatial and evidential reasoning.

Since for the most part large-scale evaluations at the Text Retrieval Conferences (TREC) (Voorhees & Tice 1999; Voorhees 2003), NTCIR (Fukumoto, Kato, & Masui 2004) and CLEF (Magnini *et al.* 2004) focused on the fact-based questions, many Class 1 systems that successfully use surface text pattern matching have been developed (Soubbotin 2001; Brill *et al.* 2001). The major advantage of this approach is simplicity: the use of surface patterns requires minimal processing, resources, and knowledge engineering compared to the other types of systems. The patterns in these systems are either hand crafted (Hildebrandt, Katz, & Lin 2004), or induced automatically similarly to Riloff's (1996) techniques for information extraction. Ravichandran and Hovy (2002) use bootstrapping to build a large tagged corpus starting with a few examples of QA pairs. Lita and Carbonell (2004) rather than extracting patterns, cluster training questions and learn models that explain the cluster, i.e. successfully answer questions from a given cluster. By learning the distribution of the expected answer type, this instance-based approach to question answering postpones decisions about an expected answer type until the answer extraction step. The use of surface patterns was extended to definition questions (Class 3) by Cui et al. (2005) who introduce two models for probabilistic lexico-syntactic pattern matching (soft pattern matching). Both models: the bigram model (a simplified first-order Markov model with one state for each token), and PHMM that aggregates token sequence probability into state transition probabilities regard def-

inition patterns as sequences of tokens. The PHMM was shown to be more robust with respect to gaps in patterns caused by language variations. The above automatic approaches rely heavily on the availability of large amounts of training data. Blair-Goldensohn et al. (2004) approached answering definition questions as a combination of data-driven document summarization methods with knowledge-based techniques that utilize empirically identified key elements of definitions.

Scenario-based question answering has recently advanced to exploration and analysis of the answer space, which might require interactions with a user. Users of a prototype scenario-based system HITIQA (Strzalkowski *et al.* 2005) are expected to submit exploratory, analytical questions. Rather than seeking just an exact answer, a user might be interested in related information. The answer is shaped by the available document collection and is not presented immediately. In a position paper, Harabagiu and Lacatusu (2004) suggest that complex questions should be decomposed into a series of simple questions, for which concept-based or pattern-based resolving techniques exist or may be developed, and that answers should be fused based on user's background and interactions with a system. This paper presents an appealing approach to answer fusion using seven fusion operators: contradiction, addition, refinement, agreement, generalization, trend, and no information.

2.5.2 Closed Domain Question Answering

Closed domain question answering has recently regained the interest of researchers. The definition of the closed domain ranges from working in a specific

domain to using closed document collection restricted in size and subject. The term restricted domain is often used interchangeably with the closed-domain, but Benamara (2004) defines it to be broader in terms of subject coverage, for example, tourism covers several subject areas, e.g. accommodation, transportation, etc., as opposed to Unix manuals. Interestingly, there seem to be no reported attempts to apply systems developed for open domain to the closed domain even by researchers with relatively successful open-domain systems (Diekema, Yilmazel, & Liddy 2004). However several closed domain systems are starting explorations in other domains, for example, the ExtrAns system developed to answer questions using Unix and Aircraft Maintenance manuals is being re-targeted to answer genomics questions (Rinaldi *et al.* 2004). The AQUA project plans to answer open domain questions by using different ontologies in addition to the currently used research domain ontology (Vargas-Vera & Motta 2004). Besides relying upon domain ontologies, the emerging closed domain systems share such features as:

- Concept recognition and matching in questions and documents (Personalphone in the Bell Canada QA system (Doan-Nguyen & Kosseim 2004).)
- Utilization of the domain-specific document structure and features such as tense, voice and style (Gabbay & Sutcliffe 2004).
- Logical representation of documents (Benamara 2004).
- Mapping questions to restricted set of semantic frames (Chung *et al.* 2004).
- Modeling questions as syntactico-semantic patterns (Jacquemart & Zweigen-

baum 2003).

An interesting experiment attempts to overcome the size limitations of the closed domain collections by searching the Web for biographic information that is missing in the closed collection, and at the same time maintain the quality of biographic data using automatic text classification (Tsur, de Rijke, & Sima‘an 2004).

The main difference between the CQA-1.0 system developed in this work and the systems described above is in the level of abstraction provided by consistent application of the semantic domain model to all steps of the question answering process.

Chapter 3

Resources, Tools and Test Collections

The CQA-1.0 system relies upon many freely available resources to generate answers to clinical questions. The resources include the domain knowledge encoded in the Unified Medical Language System (UMLS) described in Section 3.1, the database of citations into biomedical literature – MEDLINE (Section 3.2), search engines to retrieve MEDLINE citations (Sections 3.3 and 3.4), and tools that identify biomedical entities and relations in a given text (Sections 3.5 and 3.6.) Real-life clinical questions and answers from several high quality online collections described in section 3.7 were used to build the test collections for CQA-1.0 development and evaluation (see Section 3.8.)

3.1 Unified Medical Language System (UMLS)

The development of knowledge-intensive methods and tools in medical domain is made possible by the Unified Medical Language System[1]. The UMLS, maintained at the National Library of Medicine, consists of three knowledge sources: Metathesaurus, Semantic Network, and SPECIALIST Lexicon. The UMLS Metathesaurus contains information about biomedical concepts, their various names, and the relationships among them. It represents many source vocabularies (thesauri, classifica-

[1] `http://www.nlm.nih.gov/research/umls/`

tions, code sets, and lists of controlled terms) in a single database format with the purpose of linking alternative names and views of the same concept together. The Metathesaurus preserves the names, meanings, hierarchical contexts, attributes, and inter-term relationships present in its source vocabularies. It also establishes new relationships between terms from different source vocabularies. The concept *Lou Gehrig's Disease* illustrates linking of a concept name to its synonyms. Through its Metathesaurus unique concept identifier (CUI = C0002736) this string is recognized as synonymous with the following terms:

- Amyotrophic Lateral Sclerosis
- ALS
- ALS (Amyotrophic Lateral Sclerosis)
- ALS - Amyotroph lat sclerosis
- Amyotrophic Lateral Sclerosis/Progressive Muscular Atrophy
- Amyotrophic lateral sclerosis (disorder)
- amyotrophy; lateral sclerosis
- Bulbar motor neuron disease
- Gehrig's Disease
- Lou Gehrig Disease
- Motor Neuron Disease, Amyotrophic Lateral Sclerosis
- palsy; creeping
- spinal; sclerosis, lateral (amyotrophic)

The 14 synonyms were found in 19 source vocabularies (Alcohol and Other Drug Thesaurus, Clinical Problem Statements, COSTAR, CRISP Thesaurus, DXplain, ICD-9-CM, ICPC2-ICD10 Thesaurus, Library of Congress Subject Headings, MedDRA, MedlinePlus, MeSH, UMLS ICD-9-CM Terms, NCI Thesaurus,

National Drug File - Reference Terminology, Quick Medical Reference, Read Codes, SNOMED 1982, SNOMED Intl 1998, and SNOMED Clinical Terms.)

The 2005 version of the UMLS Metathesaurus contains information about over 1 million biomedical concepts and 5 million concept names from more than 100 controlled vocabularies. The Metathesaurus is used in the CQA-1.0 system both directly, in hierarchical clustering of answers(see Section 4.6.1); and indirectly, in named entity recognition using MetaMap (see Section 4.4.1.)

The UMLS Semantic Network categorizes each Metathesaurus concept into at least one of the basic semantic types. It also defines the set of relationships that may hold between the semantic types. The current release of the Semantic Network contains 135 basic semantic types and 54 relationships. There are major groupings of semantic types for organisms, anatomical structures, biologic function, chemicals, events, physical objects, and concepts or ideas. The Semantic Network provides textual descriptions of semantic types and defines important relations in the biomedical domain, in addition to 54 relationships allowed between the semantic types. The primary relation between the semantic types is "IS-A". It establishes the hierarchy of types within the Network and is used for deciding on the most specific semantic type available for assignment to a Metathesaurus concept. There is also a set of non-hierarchical relationships, which are grouped into five major categories:

(a) physically related to
(b) spatially related to
(c) temporally related to
(d) functionally related to
(e) conceptually related to

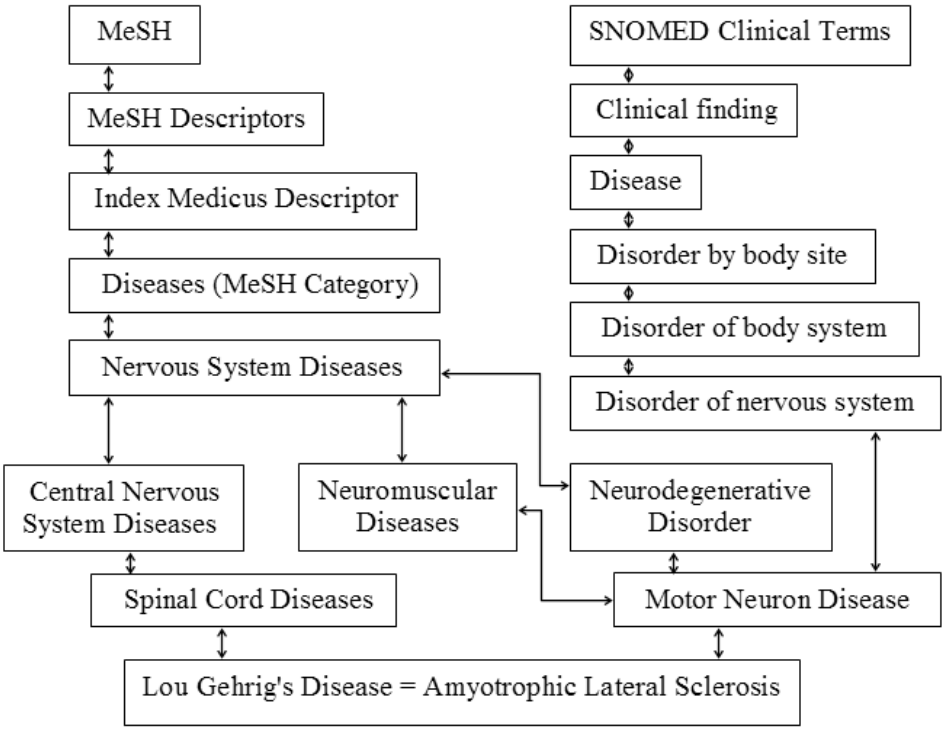

Figure 3.1: The concept *Lou Gehrig's Disease* and some of its hypernyms.

Applied to the *Lou Gehrig's Disease* example, the Semantic Network categorizes it as *Disease or Syndrome*, provides several hierarchies (see Figure 3.1) that contain this disease and related concepts, for example, ALS3 gene. In the CQA-1.0 system, semantic types are used to assign concepts identified in the documents using MetaMap to the slots of the PICO frame (see Section 4.4.2.1.)

The SPECIALIST lexicon is a general English lexicon that includes biomedical terms. The lexicon entry for each word or term contains the syntactic, morphological, and orthographic information. Lexical entries may be single or multi-word terms. Each lexical record has a base form, a part of speech, a unique identifier and optionally a set of spelling variants. The base form is the uninflected form of the lexical item; the singular form in the case of a noun, the infinitive form in the case

of a verb, and the positive form in the case of an adjective or adverb. Lexical information includes syntactic category, inflectional variation (e.g., singular and plural for nouns, the conjugations of verbs, the positive, comparative, and superlative for adjectives and adverbs), and allowable complementation patterns. The lexicon and lexical tools are not used in the proposed system directly; however both MetaMap and SemRep rely upon the lexicon and tools.

3.2 MEDLINE database

PMID- 12361440
...
TI - The efficacy of duct tape vs cryotherapy in the treatment of verruca vulgaris (the common wart).
...
LA - eng
PT - Clinical Trial
PT - Journal Article PT - Randomized Controlled Trial
PL - United States
JT - Archives of pediatrics and adolescent medicine.
JID - 9422751
RN - 0 (Adhesives)
SB - AIM
SB - IM
MH - Adhesives
MH - Adolescent
MH - *Bandages
MH - Child
MH - Humans
MH - *Cryotherapy
MH - Prospective Studies
MH - Treatment Outcome
MH - Warts/*therapy
EDAT- 2002/10/04 04:00
SO - Arch Pediatr Adolesc Med. 2002 Oct;156(10):971-4.

Figure 3.2: MEDLINE citation in MEDLINE format.

MEDLINE (Medical Literature Analysis and Retrieval System Online)[2], a large bibliographic database maintained by NLM, is known as an authoritative and comprehensive source of peer reviewed clinical evidence. In a 2001 BMJ editorial, it was called one of America's two greatest gifts to the world, and the best free starting point for finding high quality medical information (Smith & Chalmers 2001). MEDLINE contains over 16 million references to articles from approximately 4,800 biomedical journals in 30 languages, dating back to the 1950's. With the exception of few weeks for a yearly update of the controlled vocabulary, 1,500 to 3,500 references are added to the database every day, which amounted to over 571,000 new citations added to the database in 2004. The scope of database is biomedicine and health, broadly defined to encompass those areas of the life sciences, behavioral sciences, chemical sciences, and bioengineering needed by health professionals and others engaged in basic research and clinical care, public health, health policy development, or related educational activities. MEDLINE also covers life sciences important to biomedical practitioners, researchers, and educators, including aspects of biology, environmental science, marine biology, plant and animal science, as well as biophysics and chemistry.

Each MEDLINE citation includes basic information such as the title of the article, authors, journal and publication type, date of publication, language, etc. The majority of publications covered in MEDLINE are scholarly journals. For citations added during 1995-2003: about 48% are for cited articles published in the U.S., about 88% are published in English, and about 76% have English abstracts

[2]http://www.nlm.nih.gov/bsd/licensee/2006_baseline_doc.htmlz

written by authors of the articles. For about 4,500 journals, MEDLINE provides a link to the publisher's Web site to request or view the full article, depending on the publisher's access requirements. Many publishers provide free full text of the article.

Additional metadata are associated with each MEDLINE citation. Of these, the controlled vocabulary terms assigned by human indexers and information about the journal are used in this work. The NLM controlled vocabulary thesaurus, Medical Subject Headings (MeSH), contains approximately 23,000 descriptors arranged in a hierarchical structure and more than 151,000 Supplementary Concept Records (additional chemical substance names) within a separate thesaurus. Indexing is performed by approximately 100 indexers with at least a bachelor's degree in life sciences and formal training in indexing provided by NLM. Since mid-2002, the Library has been employing software that automatically suggests MeSH headings based on content (Aronson *et al.* 2004b). Metadata is provided in several formats, of which MEDLINE format and XML are most widely used. Figure 3.2 provides metadata in MEDLINE f format for the abstract shown in Figure 2.1. To represent different aspects of the topic described by a particular MeSH heading (descriptor), up to three subheadings(qualifiers) may be assigned, as indicated by the slash notation. An asterisk placed next to a MeSH term indicates that the human indexer interprets the term to be the main focus of the article. More than one MeSH term can be identified as representative of the focus of the article. In the above example, *Warts/*therapy* indicates that the article focuses on treatments for warts. The other two starred MeSH headings describe the treatment options: bandages and

cryotherapy. The publication type data indicate that the study was a randomized clinical trial. The Treatment Outcome heading indicates that the goal of the study was to determine effectiveness of the treatments.

3.3 PubMed

MEDLINE is publicly accessible on the Web through PubMed, the National Library of Medicine's gateway, or through third-party organizations that license MEDLINE from NLM. PubMed[3] is a Boolean search engine that indexes titles, abstracts and metadata separately. These indices allow users to specify which fields or indices should be searched. For example, tagging the search term as follows: *warts[mh]* indicates that only metadata should be searched for the term. It is worth mentioning that the wide variety of advanced search options makes PubMed a highly competitive search engine when used by an experienced searcher (Hersh *et al.* 1994). PubMed automatically recognizes and translates controlled vocabulary terms, and expands identified MeSH headings. The automatic term mapping process matches untagged query terms against the entries in the following tables/indexes:

- MeSH Translation Table (contains MeSH terms, entry terms for MeSH terms, MeSH Subheadings, Publication Types, Pharmacologic action terms, Terms derived from the UMLS that have equivalent synonyms or lexical variants in English, Supplementary concept (substance) names and their synonyms,)
- Journals Translation Table (contains Full journal title, MEDLINE abbreviation, ISSN)
- Author Index

[3]http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html

When a match is found for a term or phrase in a translation table, the mapping process is complete and does not continue on to the next translation table. If a match is found in the MeSH Translation Table, the term will be searched as MeSH (that includes the MeSH term and any specific terms indented under that term in the MeSH hierarchy), and as a Text Word. For example, if there were no restriction on MeSH expansion searching for osteoporosis, this term would have been expanded internally with its child Osteoporosis, Postmenopausal. Or when searching for GERD, this entry term is expanded using the MeSH term as follows: *"(gastroesophageal reflux"[TIAB] NOT Medline[SB]) OR "gastroesophageal reflux" [MeSH Terms] OR gerd[Text Word].* PubMed appends search field tags (in square brackets) to each search term. The search field tag indicates which indices will be searched, for example, [Text Word] indicates searching all textual fields of PubMed records. The above search snippet requests a union of the searches for *gastroesophageal reflux* in the titles and abstracts of citations ([TIAB]) not indexed for MEDLINE (NOT Medline[SB]), *gastroesophageal reflux* as MeSH, and *gerd* as Text Word.

Clinical Queries filter

One of the advanced PubMed search options, Clinical Queries, is a set of filters designed to find clinically relevant and scientifically sound studies. These filters automatically expand queries using predefined sets of terms designed to skew search results in favor of one of the four clinical tasks (etiology, diagnosis, therapy, and

prognosis.) For each task, Clinical Queries provide two search choices: specific (narrow) or sensitive (broad.) For example, a "narrow therapy" Clinical Query ANDs a user's query with the following string: *randomized controlled trial [Publication Type] OR (randomized [Title/Abstract] AND controlled [Title/Abstract] AND trial [Title/Abstract])*. The terms for the Clinical Queries were originally derived based upon a manual review of articles from 10 internal medicine and general medicine journals in 1986 and 1991 by the Hedges Project researchers. The filters were recently updated by the same group (Wilczynski, McKibbon, & Haynes 2001).

In addition to PubMed, NLM provides utilities for batch retrieval of MEDLINE citations with all capabilities of PubMed. These Entrez Programming Utilities[4] were used to retrieve initial sets of citations for all experiments in the thesis.

3.4 Essie

Essie, a probabilistic search engine developed at NLM for the ClinicalTrials.gov database, provides access to many databases maintained at NLM. It incorporates a number of strategies aimed at alleviating the need for sophisticated user queries. These strategies include a fine-grained tokenization algorithm that preserves punctuation information, concept searching using UMLS-derived synonymy, and phrase searching based on the user's query. Essie was the best performing search engine in the 2003 TREC Genomics track and achieved results comparable to those of the high ranking systems on the 2005 TREC Genomics track data. Essie has been

[4]http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

used as the search engine for ClinicalTrials.gov since 2001. It is also used to serve consumer information about genetic conditions through the Genetics Home Reference site. The NLM Gateway uses Essie directly to search some of the NLM data sets as well as indirectly in some of the systems it queries. Another Essie user is MEDLINE Database On Tap, which provides MEDLINE access through wireless handhelds at the point of care. The CQA-1.0 document ranking was compared with Essie retrieval results in the pilot experiment (See Section 5.3.1.)

3.5 MetaMap and MMTx

MetaMap, designed to find Metathesaurus concepts in biomedical text, was developed and is maintained at the NLM (Aronson 2001). An original MetaMap annotation of a text is obtained by submitting batch jobs to NLM. Alternatively, a JAVA implementation of the MetaMap, MetaMap Transfer (MMTx), could be downloaded and used locally through its Application Programming Interface (API). Although several other tools have been developed for the purpose of mapping text to concepts in the UMLS Metathesaurus in the past (Zieman & Bleich 1997; Denny *et al.* 2002), and more recently (Zou *et al.* 2003), only MetaMap attempts identifying all UMLS concept in any document type. Many MetaMap parameters are configurable. For example, a user can select one of the three data models that differ from each other by the level of filtering they do on the UMLS Knowledge Sources. The strict model is considered to be most appropriate for semantic-processing applications, and is used in the present work. MetaMap first uses a minimal commit-

ment parser to break the text into phrases. For each phrase, MetaMap generates acronyms, abbreviations, synonyms, derivational, inflection and spelling variants using the SPECIALIST lexicon and a list of synonyms. The variants are used to retrieve a set of Metathesaurus strings called candidates. Each candidate is evaluated for the strength of mapping to the original text. The strength score is an average of four metrics: centrality, variation, coverage and cohesiveness. Centrality is a binary value, which equals 1 if a Metathesaurus string matches the head of the noun phrase, and 0 otherwise. Variation is computed as inverse edit distance between a Meta string and the noun phrase, where edit distance for spelling variants is 0, for inflectional variants is 1, for synonym or acronym/abbreviation variants is 2, and for derivational variants is 3. Coverage value indicates how much of the Metathesaurus string and the phrase are involved in the match. The cohesiveness value is similar to the coverage value but, unlike coverage, it does not ignore gaps in phrases and strings. Candidates with the best scores are selected as final mappings.

MetaMap performance in terms of its ability to find concepts was evaluated in a small scale study using 133 unique reference concepts identified by six people in 60 titles of medical articles (Pratt & Yetisgen-Yildiz 2003). Of the 133 concepts only 73 were in the UMLS Metathesaurus. Under lenient conditions, where a MetaMap identified concept was considered a match if at least one subject also identified the concept, MetaMap achieved 93.3% recall and 84.5% precision.

There are several sources of MetaMap errors. UMLS coverage and ambiguity are external to MetaMap processing, but influence its results significantly. The results of an ongoing effort to disambiguate mappings using the context (Humphrey

et al. 2006) are not yet incorporated in the MetaMap processing. Accepting a candidate depending on its context might be helpful even if only one mapping is available, as in the following example:

```
Processing phrase: "with a concise description"
Meta Mapping (888):
   694 Concise [Biomedical or Dental Material,Organic Chemical]
   861 description [Research Activity]
```

The dental material Concise is the only candidate for the phrase, however of 5532 PubMed hits for the word *concise* only 2433 are about *bisphenol a-glycidyl methacrylate*, the chemical named Concise.

Another source of errors is overmatching, for example, phrase: "aided" was mapped with a high score to:

```
AIDS (Acquired Immunodeficiency Syndrome) [Disease or Syndrome]
Aid <2> (Manufactured aid) [Manufactured Object]
```

utterance('00000000.tx.1',"Lou Gehrig's Disease").

phrase('Lou Gehrig' 's Disease', [head ([lexmatch (['Lou Gehrig' 's disease']), inputmatch (['Lou','Gehrig',"",s,'Disease']), tag(noun), tokens ([lou,gehrig,disease])])]).

candidates ([ev (-1000, 'C0002736', 'Lou Gehrig Disease', 'Amyotrophic Lateral Sclerosis', [lou,gehrig,disease], [dsyn], [[[1,1],[1,1],0],[[2,2],[2,2],0],[[3,3],[3,3],0]], yes,no), ev(-901, 'C0002736', 'Gehrig"s Disease', 'Amyotrophic Lateral Sclerosis', [gehrig,disease], [dsyn], [[[2,2],[1,1],0],[[3,3],[2,2],0]], yes,no), ev(-827, 'C0012634', 'Disease', 'Disease', [disease], [dsyn], [[[3,3],[1,1],0]], yes,no)]).

mappings([map(-1000, [ev(-1000, 'C0002736', 'Lou Gehrig Disease', 'Amyotrophic Lateral Sclerosis', [lou,gehrig,disease], [dsyn], [[[1,1],[1,1],0],[[2,2],[2,2],0],[[3,3],[3,3],0]], yes,no)])]).

'EOU'.

Figure 3.3: MetaMap machine output example.

In some cases a correct concept is generated as a candidate phrase, but is ranked lower than other candidates, and therefore not available in final mappings. For example, phrase: "prognostic value" had a candidate score 623 as

```
Prognosis <1> (Forecast of outcome) [Health Care Activity]
```

but 694 as prognostic [Intellectual Product], which was retained as a final mapping.

The MetaMap output shown above is "human" readable, and loses some information. The original MetaMap machine output (see Figure 3.3) was used in the pilot experiments (see Section 5.3.1). The machine output consists of the original text broken up into utterances (approximately equivalent to sentences). Each utterance starts with the tag *utterance*, ends with the tag *EOU*, and contains a set of phrases with their candidates and mappings. Each candidate and mapping contains in that order: a concept score; its CUI; Preferred UMLS name(s) for the CUI; a comma separated list of textual tokens that mapped to the concept (except for omitted stopwords); a comma separated list of the concept semantic types; information about how the phrase words mapped to the concept; whether the head of the phrase is involved with the concept (yes in the above sample machine output); and whether the phrase was overmatched (no above). The Prolog-friendly machine output requires special processing to be used by the implemented knowledge extraction modules described in Section 4.4.2.1. MMTx permits using the MetaMap capabilities described above within the CQA-1.0 system without additional processing, and is used for all but the pilot experiments in this work.

3.6 SemRep

SemRep uses the UMLS and MetaMap processing to identify semantic propositions, for example, "cryotherapy treats verruca vulgaris" in biomedical text. SemRep identifies semantic relationships between the UMLS concepts using underspecified parsing and MetaMap processing for concept identification. Semantic relationships are identified through syntactic and structural phenomena called indicators. Constraints on allowed relationships are encoded in 227 manually created indicator rules that link indicators to relations encoded in the UMLS Semantic Network. Dependency grammar rules that enforce syntactic constraints are used to identify arguments of a semantic relationship. For example, the proposition *cryotherapy treats verruca vulgaris* is derived from the phrase: *cryotherapy in the treatment of verruca vulgaris* because the indicator **in** points to the Semantic Network relationship *Therapeutic or Preventive Procedure—treats—Disease or Syndrome*, and the arguments *cryotherapy* and *verruca vulgaris* are mapped to concepts having the allowed semantic types. SemRep was recently enhanced to handle semantic interpretation of comparative structures. At the moment, identification of such structures is restricted to the semantic group Chemicals and Drugs. A set of rules identifies two types of comparison relations: the first asserts only that two drugs are compared; the second provides additional information about the scale on which the drugs are compared, for example, effectiveness, and the relative position of the drugs on the scale, for example, lower_than. This thesis explores the use of comparative structures identified by SemRep for extraction of an answer distributed in multiple documents,

if possible, and falls back to less sophisticated methods if no comparative structures are available.

3.7 Online databases of clinical questions and answers

A fair number of high quality examples of clinical question-answer pairs could potentially be obtained online. Three freely available sources: Parkhurst Exchange[5], Family Practitioner Inquiry Network[6], and Clinical Evidence[7] are used for development and evaluation of the CQA-1.0 system.

3.7.1 Parkhurst Exchange Forum

The first source, Parkhurst Exchange, represents the traditional, and still most popular among clinicians, method of finding answers to clinical questions, namely asking an authoritative colleague. Questions are submitted to the database by family practitioners, and answered by a member of a panel of specialists. Questions usually provide a relatively full description of a clinical situation, for example,

> Q: ALAN RUSSELL, MD, asks, "I understand that a high-sensitivity C-reactive protein (hs-CRP) test is now part of Ontario's arsenal of laboratory tests. Could you give us some guidelines on its use?" He adds, "What red flags does an elevated CRP raise in a middle-aged patient with a family history of ischemic heart disease? Would this indicate the use of long-term ASA or other antiplatelet agents?"

The 1-3 paragraph answers sometimes provide background knowledge and evaluate the problem presented in the question, and then provide informal advice, rarely

[5] `http://www.parkhurstexchange.com/qa/index.php`

[6] `http://www.primeanswers.org/primeanswers/`

[7] `http://www.clinicalevidence.com/ceweb/conditions/index.jsp`

referencing literature sources or clinical research, as follows:

> A: CRP has a number of promising features. Because atherosclerosis is an inflammatory process and CRP is a marker for low-level inflammation, a series of epidemiologic studies suggests that it can also predict vascular events, in addition to other conventional risk factors. My current clinical practice is to measure CRP when I'm unclear on how to manage a patient. For instance, in overweight individuals with multiple cardiac risk factors, in high-risk diabetics and in people with established vascular disease, the routine use of CRP makes little sense since I'm already aggressively treating them. In patients such as a middle-aged man or woman with one or two cardiac risk factors and an intermediate risk of vascular disease, particularly if he or she doesn't want therapy, I find the measurement of this newer risk factor helpful.

3.7.2 American Family Physicians Inquiries Network

Similarly to the Parkhurst Exchange Forum, Clinical Inquiries provide answers to questions submitted by practicing family physicians to the American Family Physicians Inquiries Network (FPIN). Members of the network select a question based on its relevance to family medicine. However, rather than providing an expert opinion, answers are drawn from an approved set of evidence-based resources and undergo a peer review. The strength of recommendations for individual studies contributing to an answer is rated using criteria developed by the Evidence-Based Medicine Working Group (Ebell *et al.* 2004). For example, a similar question: *How useful is high-sensitivity CRP as a risk factor for coronary artery disease?* is answered as follows:

> Little evidence supports the use of the high-sensitivity C-reactive protein assay (hs-CRP) as a screening test for cardiovascular disease (CVD) in the healthy adult population. There is significant debate about its use in populations at moderate risk for cardiovascular disease, with some evidence suggesting its use if the results of the test will alter treatment recommendations (strength of recommendation [SOR]: C, based on extrapolation of consistent level 2 studies). Research to date is inadequate

to determine the role of hs-CRP in risk-stratification of patients when considered in light of other standard risk factors.

In addition to a short answer, FPIN provides more details in an evidence-based summary, for example, the summary for the above answer elaborates:

> The updated National Cholesterol Education Panel Adult Treatment Panel III guidelines list elevated hs-CRP ($> 3mg/L$) as an influencing factor in deciding whether to use an LDL-lowering drug for moderately high-risk patients with LDL-cholesterol values $< 130mg/dL$.

In addition, FPIN provides references to publications that contributed to the answer and the summary.

3.7.3 Clinical Evidence

Similarly to the FPIN, Clinical Evidence contributors adopt the multi-tiered answer model. Clinical Evidence (CE) from the BMJ Publishing Group summarizes currently available information about treatments for more than 200 medical conditions. As opposed to the two previous sources, the questions are selected based on the UK data on morbidity, mortality, and health care priorities. The questions focus on the benefits and harms of preventative and therapeutic interventions, with emphasis on outcomes that matter to patients. For example, the CE topic related to the risk factors for coronary disease questions above is *Primary prevention of cardiovascular disorders*. The top tier of an answer presents known interventions organized into patient-outcome oriented categories:

Likely to be beneficial: Eating more fruit and vegetables; Physical activity; Smoking cessation

Trade-off between benefits and harms: Anticoagulant treatment (warfarin); Aspirin in low risk people

Unknown effectiveness: Antioxidants (other than β carotene and vitamin E)

Likely to be ineffective or harmful: β Carotene; Vitamin E

The second tier elaborates each of the bullets in a 2-3 paragraph summary, for example,

> We found five systematic reviews, which included five large RCTs comparing regular aspirin versus control among individuals with no prior history of vascular disease, with or without vascular risk factors. The average control group risk of a serious vascular event (myocardial infarction, stroke, or death from a vascular cause) in each of these trials was low (about 1% a year).

This source also provides references to the original articles used in compilation of the answer. Motivated by these widely used secondary sources, the CQA-1.0 system implements the multi-tier form of answer presentation when an overview of the information landscape is needed.

3.8 Test Collections

Three test collections were created for the development and evaluation of the CQA-1.0 system.

1. The PICO-annotated collection was instrumental in the development and evaluation of the knowledge extraction modules.
2. The FPIN collection was created to tune the system, evaluate system components and best answer generation.
3. The CE collection were created to evaluate the multi-tiered full answers.

The FPIN and the CE collections consist of questions and answers obtained from the above-described online databases. The questions were used to retrieve

MEDLINE citations using PubMed. The retrieved citations were judged by several judges (which led to creation of several subcollections within each collection). The relevance judgments determined whether a citation contains an answer or could lead to the original expert answer contained in the online database. Each collection was subdivied into training/validation and evaluation sets. The evaluation sets were used in several evaluations.

PICO-annotated collection

Table 3.1: **Number of retrieved citations, annotators, and inter-annotator agreement for the PICO-annotated collection**

Search	Annotators	Count	Annotation	Annotator agreement		
				All	Clinicians	Best Pair
1	RN1	275	Outcome			
2	RN1, student	123	Outcome	0.42	0.42	0.42
3	PhD, author	135	Outcome	0.75		0.75
4	RN1, RN2, PhD, author	50	PICO	0.65	0.63	0.75
4	RN1, RN2, PhD, author		Outcome	0.81	0.80	0.98
5	RN1, RN2, PhD, author	50	PICO	0.63	0.77	0.84
5	RN1, RN2, PhD, author		Outcome	0.78	0.94	0.97

The PICO-annotated collection consists of 633 MEDLINE citations retrieved using PubMed search strategies presented in Appendix A. The collection was created by a group of five NLM employees and visiting researchers (including the author.) Agreement between annotators was measured using Cohen's kappa (Siegel & Castellan 1988) on a sentence-by-sentence basis and ranged from moderate to good (see Table 3.1.) If possible, the differences were reconciled before using the collection for training and evaluation. If one of the annotators could not participate in the rec-

onciliation, PICO elements annotated by the majority of the assessors were marked as true positives. In case of two judgments, the intersection of two annotations constitutes true positives.

The initial goal of the annotation effort was to identify succinct patient health outcome statements in abstract text. The definition of outcome was taken from the MeSH scope notes that define the outcome as "...the results or consequences of management and procedures used in combating disease." The annotators were instructed to identify outcomes as a component of the PICO framework. Passages containing outcome statements were identified and annotated in 592 citations. One hundred abstracts were in addition annotated with population, problems, and interventions. With the exception of 50 citations retrieved to answer a question about childhood immunization (Search 4 in Table 3.1), the rest of the results were retrieved by querying on diseases or treatment outcomes.

Feasibility of outcome annotation was established using 275 articles retrieved in an expert search conducted by a registered nurse (RN1 and Search 1 in Table 3.1 and Appendix A) with more than 20 years of experience. RN1 annotated 2.25 sentences per abstract (on average) as outcome statements. The expert search strategy was then used to obtain additional 123 citations (Search 2 in Table 3.1.) The RN1 annotated on average 1.9 sentences per abstract as outcome statements in these citations. The second annotator, a medical student, on average annotated 4.3 sentences per abstract as outcome statements; 83% of the statements identified as outcomes by RN1 were also marked as such by the medical student. Because of the difference in the size of the annotated passages, agreement between the annotators was only fair.

Analysis of the disagreements showed that the medical student tended to include disease-oriented outcomes and statistical information in support of the outcome in addition to the patient outcome statements. These observations led to revision of the annotation scheme, in which the outcome statements were separated from the supporting data, and annotation of the population and interventions/comparisons became explicit. During the revision, the annotators strongly disagreed identifying the PICO elements at the phrase level. For example, given the following sentence:

> This double-blind, placebo-controlled, randomized, 3-period, complete block, 6-week crossover study examined the efficacy of simvastatin in adult men and women (N = 151) with stable type 2 DM, low density lipoprotein-cholesterol 100 mg/dL, HDL-C < 40 mg/dL, and fasting triglyceride level > 150 and < 700 mg/dL.

all annotators agreed that the sentence contained the problem, population, and intervention. However, it was hard to determine the exact phrasal boundaries of each element, and more importantly, general guidelines for ensuring consistent annotations. That is, should the whole clause starting with adult men and women be marked as population, or should type 2 Diabetes Mellitus (type 2 DM) be marked-up only as the problem? Should every instance of diabetes and simvastatin be annotated, or is it enough to mark-up each PICO element once? Should other characteristics of the study population, such as their cholesterol levels, be annotated explicitly, or included within the population annotation? The decision was made to keep the scheme as simple as possible, annotate each element once, keeping problem and population together, and extend boundaries of identified PICO elements to the boundaries of sentences that contain the elements.

The extended annotation scheme (see Table 3.2) was applied to the abstracts

Table 3.2: **Extended scheme for annotation of clinically relevant elements in MEDLINE citations. Original elements shown in bold**

Tag	Definition
Background	Material that informs and may place the current study in perspective, e.g., work that preceded the current; information about disease prevalence, etc.
Population	The group of individual persons, objects, or items comprising the study's sample, or from which the sample was taken for statistical measurement
Intervention	The act of interfering with a condition to modify it or with a process to change its course (includes prevention)
Statistics	Data collected about the results of the intervention demonstrating its effect
Outcome	The sentence(s) that best summarizes the consequences of an intervention
Supposition	An assumption or conclusion that goes beyond the evidence presented in an abstract
Other	Any sentence not falling into one of the other categories and presumed to provide little help with clinical decision making.

retrieved using the last two search strategies in Table 3.1. The 50 most recent abstracts were selected for annotation from each of the search results, yielding 100 citations annotated with all elements in Table 3.2. Using the new scheme and annotating sentences rather than phrases resulted in good inter-annotator agreement (see Table 3.1.) The intra-annotator consistency was measured comparing the original judgments made by an assessor with the consensus annotation. The intra-annotator consistency in annotating outcomes was excellent for RN1, RN2, and the author (kappa ranging from 0.91 to 0.99) and good for the PhD (kappa = 0.81). For RN1, RN2, and the author, the intra-annotator consistency on all PICO elements (kappa

= 0.92, 0.8, and 0.9 respectively) was also better than that for the PhD (kappa = 0.73.)

The annotated citations were used as follows: 275 citations annotated by RN1 (Search 1 in Table 3.1) were used for training and rule derivation for knowledge extractors. Three hundred and forty eight of the remaining 358 citations (including one hundred fully annotated citations) were used to evaluate the outcome extractor. The fully annotated citations were used to evaluate the population and intervention extractors.

FPIN collection

Table 3.3: **Questions distribution by task in the FPIN test collection.**

	Therapy	**Diagnosis**	**Prognosis**	**Etiology**	**Total**
Training	10	6	3	5	24
Evaluation	12	6	3	5	26

Two sources of questions asked and answered by doctors were used to create this test collection (the Family Practitioner Inquiry Network and Parkhurst Exchange.) To avoid inadvertently biasing selection in favor of the system, the questions were gathered from FPIN and Parkhurst Exchange by Xiaoli Huang, an information studies graduate student unfamiliar with the CQA-1.0 system (Huang, Lin, & Demner-Fushman 2006). Selection of 59 questions was guided by typical instance sampling (Lindlof & Taylor 2002), thus capturing a realistic sampling of the scenarios that a clinical question answering system would be confronted with. These questions were minimally modified from their original form as downloaded

from the Web. In a few cases, a single question actually consisted of several smaller questions; such clusters were simplified by preserving a single question about the central clinical problem. All questions were manually classified into one of the four clinical tasks, yielding 25 therapy, 15 diagnosis, 12 etiology, and 7 prognosis questions. The distribution of the questions follows the prevalence of each task type as observed in natural-settings, noted by Ely et al. (1999). Nine of 59 questions were discarded because they retrieved no citations, or not enough to warrant further processing. For example, the question *Can finasteride cause or contribute to osteoporosis in men?* could not be answered using MEDLINE at the date of search, because only three citations were retrieved using terms finasteride and osteoporosis without any advanced search strategies that limit the size of the retrieved set. In fact, the answer to this question in Parkhurst Exchange is *That isn't known...* The remaining questions are divided into 24 questions for training and 26 for evaluation. Table 3.3 presents the distribution of questions by clinical task. Appendix B presents the questions. Each question was manually translated into a PubMed query as described in Section 4.2. The top fifty results retrieved for each query were used in the evaluation. In total, 2309 citations were gathered because some queries returned fewer than fifty hits. All abstracts were evaluated by the author for relevance on a four-point scale:

Contains answer: the citation directly contains information that answers the question.

Relevant: the citation does not directly answer the question, but provides topically-relevant information (containing information on the topic of request.)

Partially relevant: the citation provides information that is marginally relevant.

Not relevant: the citation does not provide any topically-relevant information.

In total, the relevance assessment (which resulted in creation of the FPIN-train and FPIN-eval-1 subsets) took approximately 100 hours, or about an average of 2 hours per question.

Additional relevance judgments by two MDs (Dr. CS and Dr. KWF) were obtained for answers extracted from 221 training set citations (up to 10 citations for each question.), resulting in the FPIN-eval-2 subset.

Clinical Evidence collection

The second collection, the *CE* collection, consists of 30 questions of the type *What is the current opinion on the best pharmacotherapy for disease X?* randomly selected from 55 questions of this type identified by an MD unfamiliar with the CQA-1.0 system (Dr. MF) in the June 2004 issue of Clinical Evidence. The randomly selected diseases are presented in Appendix C. The question type was chosen as the most frequently occurring (15% of all clinical questions), based on Ely et al. (1999). The reference answers to these questions primarily discuss drug therapy. The structure of the four-tiered reference answers in this collection is relatively uniform.

1. The top tier consists of interventions categorized into:
    - beneficial
    - likely beneficial
    - trade-off between benefits and harms
    - unknown effectiveness
    - unlikely beneficial

- likely to be ineffective or harmful

2. The second tier presents short two to three sentence key messages providing more context for the top-tier answers.
3. The third tier provides variable length summaries for each answer.
4. The fourth tier provides references to articles used in compiling the review, with links to PubMed, if available.

The top-tier answers are of three types:

- broad classes of interventions, for example, topical anti-infective agents for otitis externa
- specific interventions, such as Permethrin, Crotamiton, Oral ivermectin, and Lindane for scabies
- mixed, such as Selegiline and Dopamine agonists for Parkinson's disease.

This distinction matters in evaluation: there are three possibilities: 1) the system finds an exact match, 2) the system finds a hypernym of the reference answer, 3) the system finds a hyponym (a UMLS concept whose semantic range is included within that of another concept, its hypernym) of the reference answer.

In two cases the system answer is considered to be correct: for the exact match and if the hyponym is found, because if the reference answer names anti-infective agents, then any specific representative of this class could be used. On average, 11.3 interventions are listed as the top-tier answer for a disease in this collection. Of those, 2.3 on average are marked as beneficial and 1.9 as likely beneficial. These top-tier answers were used in the manual evaluation of the key points generated by the CQA-1.0 system (see Section 5.4.2). The CQA-1.0 answers were generated using PubMed search results obtained submitting each disease name to PubMed in the template query described in Section 4.2. The answer quality and the strength

of evidence in support of the system-generated key points were evaluated by two assessors, an MD enrolled in the NLM medical informatics rotation program (Dr. CA) and the author, yielding two overlapping sets of relevance judgments. The 244 citations evaluated by Dr. CA formed the CE-eval-2 subset and 273 citations evaluated by the author formed the CE-eval-3 subset.

The respectable number of references associated with each disease in the CE collection (48.4 on average) provides an opportunity to explore an automatic evaluation described in Section 5.4.3. Many of the references (34.7 citations on average) appeared in MEDLINE. These citations contributed to answer generation only if they were actually retrieved by the system from MEDLINE using question frames. There were 189 (7.6%) citations in the intersection of references and retrieval results (6.3 citations per topic on average.) The distribution of these citations in the evaluation sets varies (see Section 5.4.3.) To compare ROUGE results to human judgments, 267 of citations retrieved using PubMed were evaluated by the author on the four-point scale described above (subset CE-eval-4). The above overview of the test collections and their use in experiments is summarized in Table 3.4.

Table 3.4: **Subsets of Three Test Collections for System Development and Evaluation.**

Collection	Size	Use	Judged by:
PICO-train	275 citations	knowledge extractors development	RN1
PICO-eval-1	358 citations	outcome extractor evaluation	RN1, RN2, PhD,[author]
PICO-eval-2	100 citations	population and intervention extractor evaluation	RN1, RN2, PhD, [author]
FPIN-train	24 questions	CQA-1.0 system training	author
FPIN-eval-1	26 questions	CQA-1.0 system components evaluation	author
FPIN-eval-2	221 citations	CQA-1.0 best answers evaluation	Dr. CS, Dr. KWF
FPIN-eval-3	30 questions	CQA-1.0 multi-tiered answers evaluation	author
CE-train	5 questions	CQA-1.0 System tuning	author
CE-eval-1	25 questions	CQA-1.0 multi-tiered answers evaluation	Dr. CA, author
CE-eval-2	244 citations	answers support evaluation (manual)	Dr. CA
CE-eval-3	273 citations	answers support evaluation (manual)	author
CE-eval-4	267 citations	answers support evaluation (automatic)	author

Chapter 4

EBM-based Question Answering System

The crux of the implemented system is the matching of the semantic representation of the user's information needs and the semantic representation of documents automatically derived from MEDLINE citations. Building upon the idea of a frame proposed in the 1970's (Minsky 1975), semantic matching uses the frame-based representations of a document and a question. Figure 4.1 illustrates all steps in the flow of the question answering process. The process starts with a manually constructed focused clinical question encoded in a PICO frame and results in two types of answers: a multi-tiered full answer, or the best answer. The multi-tiered answer presents an overview of available information in a concise bulleted form with a possibility to drill down to each individual context that contributed to the answer key points. The authoritative human-compiled answers widely used by clinicians often use this form of presentation. Clinicians need an alternative, the best answer (or a short ranked list of best answers), to verify a fact or get an update for a known item.

The semantic domain knowledge is applied consistently at every step of the process: providing guidance and vocabulary in query formulation; enabling document retrieval and ranking; enabling entity identification and extraction of document frames; and enabling answer generation. The CQA-1.0 system uses many available

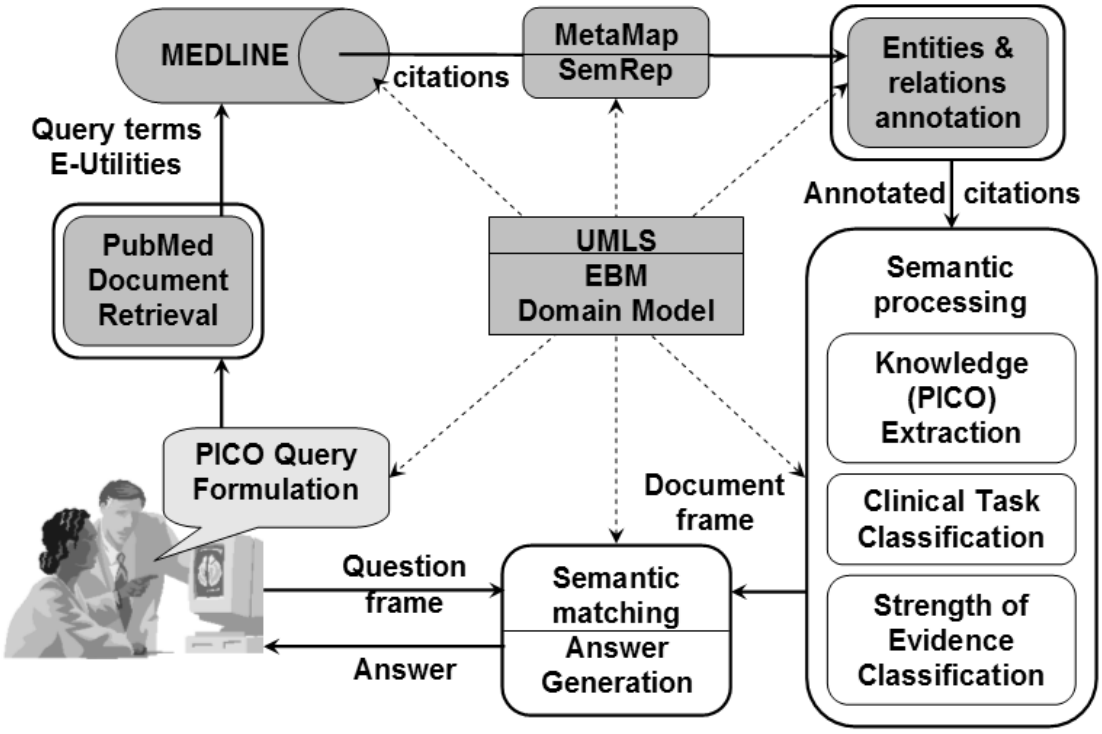

Figure 4.1: EBM-based Question Answering

resources and tools described in Chapter 3. The available tools are shown in gray in Figure 4.1. The white rims around the tools indicate the wrappers implemented by the author to use the tools within the system.

This chapter starts with an overview of the system architecture followed by a detailed description of the manual coding of PICO frames and retrieval of documents for answer generation. Then this chapter presents the details of the originally developed algorithms and components of the CQA-1.0 system (shown in white in Figure 4.1.) The presentation order follows the process flow.

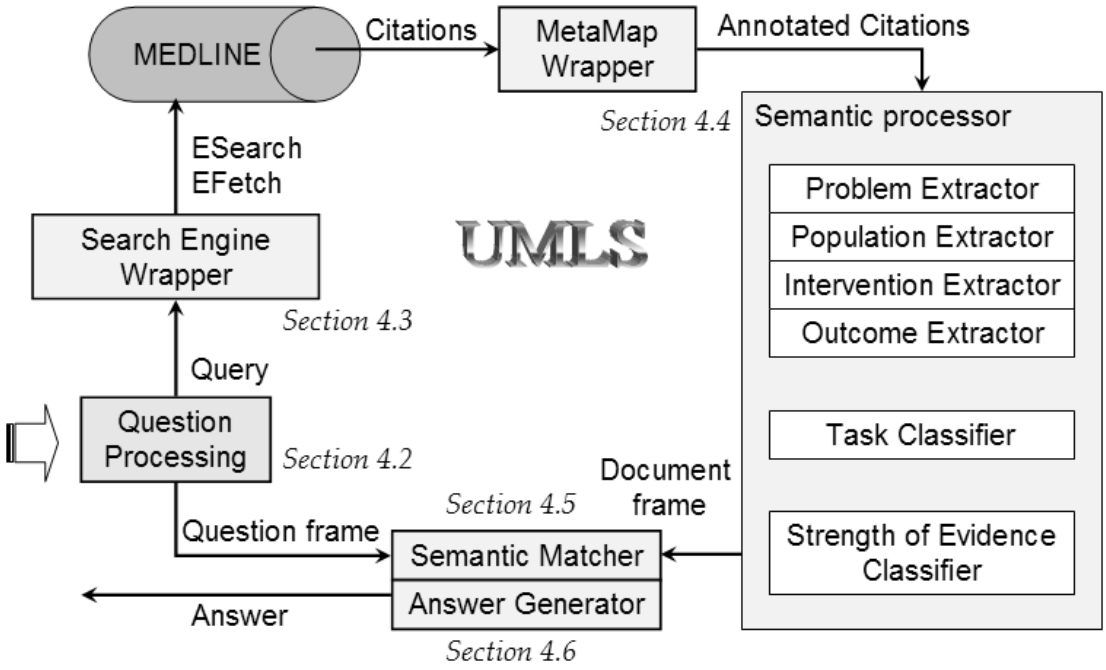

Figure 4.2: System architecture

4.1 System architecture overview

The CQA-1.0 system is organized into several modules that provide the capabilities necessary to approach question answering as a semantic matching process (determining whether a question and a document share a common meaning.) These capabilities include semantic processing of documents and knowledge extraction described in Section 4.4, document ranking (Section 4.5), and answer generation(Section 4.6.) Modules of the system implemented in Java (see Figure 4.2) maintain the process flow shown in Figure 4.1.

A manually formed question PICO frame and a clinical task of interest serve as inputs to the system. To ensure feasibility of the knowledge-intensive processing needed for answer generation, the implemented system follows the two-step de-

sign adopted by many of the open- and closed-domain QA systems: CQA-1.0 uses PubMed to reduce the number of documents from which the answers are generated.

Each task and PICO frame entered into CQA-1.0 is used to generate a PubMed query in the first step of the process. A manual and an automatic strategy for translating PICO frames into PubMed queries (described in Section 4.2) were used in the system evaluation. The query is submitted to PubMed using E-Utilities, an http-based mechanism provided by NLM. PubMed returns a list of MEDLINE citations in XML format.

A standoff annotation is created for each of the returned XML documents. Metadata, such as MeSH headings, publication type, journal descriptor, year of publication, and other, are externalized using the Uniform Retrieval Architecture (URA) toolkit developed by Jimmy Lin.

The title and the abstract of a citation are then submitted to MetaMap. UMLS concepts found in the title and the abstract text are added to the standoff annotation along with their semantic types and unique identifiers.

The annotated documents enter a semantic processor – a chain of knowledge extractors and classifiers. The semantic processing adds the extracted elements of the semantic domain model to the standoff annotation. The semantic processing results in a complete document frame (see Section 4.4.)

A fully annotated document frame is compared with the submitted question PICO frame and task in the semantic matcher. The semantic matcher scores each document based on the match between the document and the question frame. The strength of evidence and matching on the clinical task also contribute to the docu-

ment score (see Section 4.5.)

Finally, the best or multi-tiered answer is generated (see Section 4.6.)

4.2 User Query: PICO frames

The PICO frames, which serve as input to the CQA-1.0 system, were coded by the author for all questions in the FPIN and CE collections (FPIN-train, FPIN-eval-1, CE-train and CE-eval-1 in Table 3.4.) The translation of a clinical question into a frame for the CQA-1.0 system and a PubMed query was a three step process.

1. A conceptual PICO frame was formed by identifying PICO elements in a real life clinical question.
2. The conceptual PICO frame created in the first step was used to fill the slots in a PICO frame to be submitted to the CQA-1.0 system. Manual coding ensured that the frame contained the most widely used terms (not necessarily encoded in the UMLS) and their synonyms contained in the UMLS.
3. The instantiated PICO frame was used to develop a PubMed query focusing on the terms likely to retrieve relevant documents. Since document retrieval is an intermediate step, recall was more important than precision in this step.

4.2.1 Identifying PICO elements in a real life clinical question

The coding of the clinical questions in the FPIN and CE collections only approximates a real clinical situation. The important entities were identified in the questions submitted by doctors, whereas EBM practitioners are taught to first identify the important PICO elements in a clinical situation, and then formulate a question. It is unknown whether questions submitted to the FPIN and Parkhurst Exchange databases were formulated according to these recommendations. Some of the questions, therapy questions in particular, seem to be created with PICO

in mind, and are therefore easily coded. For example, *Do acetaminophen and an NSAID combined relieve osteoarthritis pain better than either alone?*(B.2.1.5) translates to:

Problem: *osteoarthritis*

Intervention: *acetaminophen and non-steroidal anti-inflammatory drugs combination*

Comparison: *acetaminophen*

Comparison: *non-steroidal anti-inflammatory drugs*

Outcome: *pain relief*

The above question serves as an example of one of the few almost fully instantiated frames in this set. The population/problem slot(s) were present in all but three therapy questions, and for seven questions it was the only populated slot. These seven questions fall into the type: *What is the best current treatment for X?* for example, *What is the most effective treatment for ADHD?* Three therapy questions that do not name a problem follow two patterns: *What is the most effective representative of intervention Y?* for example, *What is the most effective nicotine replacement therapy?* and *Is intervention Y effective?* for example, *What's the current success rate of electroconvulsive therapy (ECT)?* In both cases world and professional knowledge help instantiate the remaining slots. For example, when preparing the frame for the nicotine replacement therapy question, world knowledge leads to the following inferences:

Problem: *smoking*

Population: *smokers*

Intervention: *nicotine replacement therapy*

Outcome: *smoking cessation*

A good answer to this question would provide an overview of options for different population groups and tobacco use behavior. Note that due to the insights gained in studying the EBM literature, the problem and the population slots are separated in the conceptual and subsequently in the input PICO frames. To get the best answer for a specific clinical situation, the problem and the population in the above question need to be further clarified.

A high-level abstraction of diagnostic questions amounts to:

- *What is the most likely condition to cause the observed manifestations?* and
- *Which tests could confirm/reject the manifestations are caused by condition X?*

The first question is the question of differential diagnosis. Due to the word sense ambiguity (for example, *anemia* might refer to the disease and its clinical manifestations, or to the reduction in the number of red blood cells), it is hard to distinguish the differential diagnosis questions from the etiology questions without a direct indication from a user, or from additional data (Seol *et al.* 2004). Because of the similarity of etiology and differential diagnosis questions, identification of PICO elements for differential diagnosis does not differ from that for etiology questions, and will be described together later. The second question addresses diagnostic tests. This type lends itself to PICO representation almost as naturally as therapy questions. Six of 15 diagnostic questions in the collection belong to this type. In all six, following the BMA recommendations, the presenting patient's problems fill the

Patient slot, the hypothesized problems fill the **P**roblem slot, and the tests fill the interventions and comparison slots. For example, question B.1.2.2 (*For knee pain, how predictive is physical examination for meniscal injury?*) is represented as:

Patient: *knee pain*

Problem: *meniscal injury*

Intervention: *physical examination*

Preparing the conceptual etiology frames was relatively straightforward: the existing problems were assigned to the problem slot, causes to the exposure/intervention slot, and potential problems to the outcome slot. For example, question B.1.4.2 (*Does maternal smoking cause ADHD?*) is represented as:

Intervention/Exposure: *maternal smoking*

Outcome: *ADHD*

The same principle was applied in preparing the prognosis frames: existing patient's problems were assigned to the problem slot, and potential problems to the outcome slot. For example, question B.2.3.3 (*Does a polyp in the gallbladder pose any risk of becoming malignant?*) was coded as follows:

Problem: *Gallbladder polyp*

Outcome: *malignancy*

Preparing conceptual frames for the second set of questions was extremely simple. Because all questions are about best known treatments for a disease, only the problem slot was filled with the name for each disease.

Table 4.1: Mapping of a conceptual PICO frame into a system input frame.

Conceptual PICO frame		Query PICO frame	
Problem:	*chronic fatigue syndrome*	Problem: Problem CUI:	*chronic fatigue syndrome* *C0015674*
Intervention:	*stimulants*	Intervention: Intervention CUI:	*stimulants* *C0304402*
Task:	*Therapy*	Task:	*Therapy*

4.2.2 Instantiating input PICO frames

Conceptual PICO frames were used to instantiate the CQA-1.0 question frames. Two types of representation were used for the concepts identified in a clinical question:

- the widely used surface representation(s) of the concept, and
- the UMLS unique concept identifier (if the concept is represented in the UMLS.)

Although the mappings for the CQA-1.0 frames were performed manually in this work, the process could be automated using the UMLS API to find out if a concept is in the UMLS, and using document frequency of the term and concept surface representations to determine their "popularity" in MEDLINE. If a term is mapped to the UMLS concept, the concept unique identifier (CUI) is added to the Query Pico frame. The CUI is used in the process of matching the document and the question frames described in Section 4.5.1. The process of PICO frame instantiation applied to most of the conceptual frames amounts to one-to-one mapping and is illustrated using question B.2.1.11 (*Could stimulants be useful for chronic fatigue syndrome?*) in Table 4.1.

In a few cases the CQA-1.0 question frame is much more detailed then the conceptual frame. For example, the conceptual frame for question B.1.3.2 (*What's the prognosis of lupoid sclerosis?*) is:

Problem: *lupoid sclerosis*

Task: *Prognosis*

Lupoid sclerosis is a UMLS concept contributed by the COSTAR (Computer-Stored Ambulatory Records) vocabulary, but it occurs in MEDLINE only 13 times, probably because it is an old term used to describe features observed in two diseases: *systemic lupus erythematosus* and *multiple sclerosis*. The question does not provide any more details, so it is not clear which of the diseases the doctor had in mind. Only one of 13 abstracts containing the term presents an observation of a long-term outcome. To better answer the prognosis question, the frame submitted to the search engine lists all three diseases and the clinical task *Prognosis*. Clearly, such extreme cases cannot be easily automated.

Because the users are unlikely to thoroughly prepare the conceptual frames, and it is unlikely that anyone but a trained medical librarian will use controlled vocabulary terms, conceptual frames created for the CE collection questions were used to populate the CQA-1.0 frames without any additional processing. For example, both frames for question C.2.13 *What is the treatment for genital warts?* consist of one slot: Problem: *genital warts*, and identify clinical task as *Therapy*.

Unique concept identifiers for both sets of question frames were looked up in the UMLS and added to the CQA-1.0 frames manually. However, judging by results of Pratt's experiments (2003) it is reasonable to assume that up to 93% of the unique

concept identifiers can be obtained automatically.

4.2.3 Frames to PubMed queries

The retrieval of MEDLINE citations using PubMed is necessitated by the two-step architecture of the CQA-1.0 system. The query formulation process is PubMed-specific. In general, PubMed searches using all terms from fully instantiated frames as recommended in the EBM tutorials proved to be too restrictive. For example, searching for *osteoarthritis* AND *acetaminophen NSAID combination* AND *pain relief* (question B.2.1.5) retrieves 11 citations that do not include any studies of NSAID-acetaminophen combinations. Relaxing the search to *osteoarthritis* AND *acetaminophen NSAID combination* retrieves 32 citations, one of which provides partial information that a *naproxen/paracetamol combination* is more effective than treatment with *naproxen* alone. Such observations shaped translation of PICO frames into PubMed queries.

The goal of PubMed queries was to retrieve a set of documents covering all aspects of the question, and at the same time containing high quality relevant documents among the first 10-20 retrieved, since this set was used as the baseline in the CQA-1.0 evaluation. These requirements resulted in a thorough inspection of documents retrieved using PubMed, and several iterations of query formulation for each question. The initial queries were based on techniques recommended in the EBM sources and widely-accepted for narrowing search results: restricting search to citations 1) with abstracts, 2) published in English, 3) containing MeSH term

Humans, and 4) applying Clinical Query filters (see Section 3.3), which was too restrictive in some cases. For example, for question B.1.1.1 (*What is the best treatment for analgesic rebound headaches?*), the initial search contained the term *analgesic rebound headache* and a narrow therapy filter. PubMed translated this query to:

> (("headache disorders"[TIAB] NOT Medline[SB]) OR "headache disorders"[MeSH Terms] OR analgesic rebound headache [Text Word]) AND (randomized controlled trial [Publication Type] OR (randomized[Title/Abstract] AND controlled[Title/Abstract] AND trial[Title/Abstract])) AND hasabstract[text] AND English[Lang] AND "humans"[MeSH Terms]

PubMed automatically identifies concepts and the terms for these concepts to be matched against MeSH headings assigned to citations by indexers (see Section 3.3.) In this case, because none of the top 20 results were relevant, the query was manually expanded with terms *side effects* and *analgesics* to emphasize the aspect of the problem requiring an intervention. During further manual modifications, the therapy filter that amounts to retrieving only results of clinical trials was removed from the query. The final query for the above question was:

> ((("analgesics"[TIAB] NOT Medline[SB]) OR "analgesics"[MeSH Terms] OR "analgesics"[Pharmacological Action] OR analgesic[TextWord]) AND (("headache"[TIAB] NOT Medline[SB]) OR "headache"[MeSH Terms] OR headaches[TextWord]) AND ("adverse effects"[Subheading] OR side effects[Text Word])) AND hasabstract[text] AND English[Lang] AND "humans"[MeSH Terms]

This is an example of an exceptionally hard question about headaches caused by medications taken to alleviate headaches. An incomplete question about a high-impact disease might also be hard to answer. For example, a search for *diabetes* with the narrow therapy filter retrieves 7,268 hits. This question cannot be narrowed down using the given incomplete PICO frame. In the experiments, the clinical

scenario provided in the original databases and the top 10-20 retrieved citations were reviewed to refine the frame and the search. On average, query generation required about forty minutes per question. This process was useful not only in the creation of the test collection, but also in deriving rules for translating PICO frames to queries. The following rules were derived:

- ▷ The initial search should use all populated frame slots and a narrow Clinical Query filter for a given task.
- ▷ If 50 to 500 abstracts are retrieved, then there is enough high quality evidence (such as clinical trials for therapy questions) for this topic.
- ▷ If more than 500 abstracts are retrieved, the question is not specific enough and needs to be refined.
- ▷ If less than 50 abstracts are retrieved, the frame is too specific, and important information might be missed. The search needs to be relaxed:
  - ◦ removing some of the search terms from the query, and/or
  - ◦ looking for any evidence, not only high quality studies (removing the Clinical Query filter.)

Table 4.2: **Conceptual and input frames for the question** *Do TCAs or SSRIs have any effect on decreasing tinnitus?*

Conceptual PICO frame		Query PICO frame	
Problem:	*tinnitus*	Problem: Problem:	*tinnitus* *ringing*
Intervention:	*TCA*	Intervention: Intervention:	*TCA* *tricyclic antidepressant*
Intervention:	*SSRI*	Intervention: Intervention:	*SSRI* *selective serotonine re-uptake inhibitors*
Outcome:	*decrease tinnitus*	Outcome:	*decrease tinnitus*
Task:	*Therapy*	Task:	*Therapy*

The following rules were derived to relax a query by removing filters and terms from the PICO slots that serve as soft constraints in the semantic matching process:

1. Remove terms from the outcome slot for therapy and diagnostic questions. This rule is based on observations that the intended patient-oriented outcome for these tasks is most often implied but not stated explicitly in the abstracts. For example, a sufficient number of hits were retrieved by removing *pain relief* from the query generated for the *NSAID/acetaminophen* question discussed above.

2. If the result set is still empty after the first step remove or relax the population terms. For example, removing *young athletes* from the search for question B.1.1.3(*Does quinine reduce leg cramps for young athletes?*) results in an appropriate retrieved set.

3. In the next step, the task-specific Clinical Query filter is removed. This step retrieves enough results to answer question B.2.1.6 (*What regimens eradicate Helicobacter pylori?*), but is still not sufficient for the question about *tinnitus*.

4. If the result set is still empty or too small after the previous relaxation steps, revise all terms. For the question in Table 4.2 ANDing *TCA*, *SSRI*, and *tinnitus* is problematic. The solution is to replace the drug classes with a hypernym, *antidepressive agents*. The final query, which preserves the requirements for having an abstract and being restricted to human studies, is successful:

   (("tinnitus"[MeSH Terms] OR tinnitus[Text Word]) AND (("antidepressive agents"[TIAB] NOT Medline[SB]) OR "antidepressive agents"[MeSH Terms] OR "antidepressive agents"[Pharmacological Action] OR antidepressants[Text Word])) AND hasabstract[text] AND "humans"[MeSH Terms]

The elimination of the outcome slot contradicts findings of Bergus et al. (2000), who observed that when asking a specialist, doctors had a better chance of receiving recommendations or an answer if the intended intervention and desirable outcomes were stated in the question. Following these recommendations would require a deeper automatic understanding of the outcome statements then is currently available. An intermediate solution is to rely on the generic patient outcome indicators as recommended in EBM. For example, an EBM recommendation for *diagnostic tests* is to instantiate the outcome slot with the term *sensitivity and specificity*, an indexing term for results of diagnostic and screening tests. Because clinicians

need to find reliable tests to confirm suspected diagnosis, they are interested in the predictive power and quality of the tests. The MeSH heading *Sensitivity and Specificity* or those words themselves are indicative of the evaluations of diagnostic tests (Wilczynski, McKibbon, & Haynes 2001). Because the retrieved citations are not the end product, but serve as an input to the CQA-1.0 system, use of these indicators is deferred to the later matching stages, and there is no need to add the outcome slot terms to the search as recommended in the EBM tutorials.

Approximate translation of the input PICO frames into PubMed searches based on the above rules might permit automating the process. Although testing this hypothesis is beyond the scope of this dissertation, an automatic approximate translation strategy was used in retrieving documents for the CE collection questions. The following manually constructed search template was used for every disease:

> (DISEASE NAME[mh:noexp]) AND drug therapy[sh] AND hasabstract[text] AND Clinical Trial[pt] AND English[Lang] AND humans[mh]

In this template, the[mh:noexp] tag next to the disease indicates that only the MeSH index should be searched, but not the abstract text (the *mh* part of the tag). The *noexp* part of the tag requires an exact match without expansion (no narrower terms are included in the search). In addition, the search is restricted to *human* clinical trials (*Clinical Trial[pt]*), having an abstract, published in English, and indexed with the subheading *[sh] drug therapy*. The template, with *DISEASE NAME* replaced by each of the diseases in Appendix C, was automatically submitted to PubMed using E-Utilities. An in-depth study of the ways to generate reliable

query templates, and map questions to templates is needed in the future.

4.3 Document Retrieval using Entrez Utilities

The CQA-1.0 system submits queries to PubMed using the ESearch utility which retrieves primary IDs (PMIDs) of the documents. The term translation described in Section 4.2.3 needs to be done manually (or use a manually constructed template) only for queries with the desirable advanced features, such as having a term as a MeSH heading but not as a text word, or restricting a MeSH heading to just the term without expansion (see Section 3.3.) If none of the above is required, terms in the frame may be submitted through ESearch as is.

The list of PMIDs returned by ESearch is submitted to the EFetch utility that returns documents in requested format (which is XML for the CQA-1.0 system.)

4.4 Document frame generation

Each MEDLINE citation retrieved using PubMed is processed to generate its internal CQA-1.0 representation – a document frame. Each document frame generated by the CQA-1.0 system contains elements of the clinical scenario (Problem(s), Population, Intervention(s), and Outcome), Clinical Task scoring derived from MeSH headings, and the Strength of Evidence scoring based on metadata (MeSH headings, Publication Types, MEDLINE subset, ISSN of the journal, and Publication Date). Metadata for scoring is extracted from the citation in the PubMed XML format. The elements of the clinical scenario are extracted from

the title and the abstract text of a citation using Knowledge Extractors (see Section 4.4.2.1) developed specifically for this task.

4.4.1 Citation text preprocessing

The extraction of the elements of a clinical scenario relies upon 1) the knowledge of the discourse structure of the abstract of a MEDLINE citation, and 2) identification of the UMLS concepts in the abstract text. This information is obtained in the preprocessing step. The preprocessing starts with identification of the discourse structure of an abstract. The discourse structure is determined in structured abstracts using a simple finite-state machine (FSM) developed by the author.

Next, the acronyms and abbreviations are expanded in the titles and abstracts of retrieved citations using an abbreviation-expansion module developed by the author.

The abstract text is then submitted to MetaMap for entity identification. MetaMap processing concludes the preprocessing step.

Discourse structure annotation

Recommendations of the Ad Hoc Working Group for Critical Appraisal of the Medical Literature (1987) for structured abstracts are enforced by many leading clinical journals. The abstracts have to be structured, however the structure and the section headings are merely suggested. The suggested structure follows that of

an article, for example, JAMA proposes these sections: *Context, Objective, Design, Setting, Patients or Other Participants, Intervention(s), Main Outcome Measure(s), Results, Conclusions*. The headings vary widely from journal to journal, and even within one journal, as the headings are only recommended. A total of 2688 section heading variations were collected by the author from a 10-year (1993-2004) subset of MEDLINE citations. To facilitate further processing, all structural headings are substituted with one of the four traditional headings:*introduction, methods, results, and conclusions*. For section headings provided by publishers, instructions for authors were used for the substitution. For example, JAMA provides the following description of the context section: *The abstract should begin with a sentence or 2 explaining the clinical (or other) importance of the study question*, from which it is clear that *context* maps to *introduction*. If no instructions were available, the substitution rules were generated by the author based on inspection of random samples of the abstracts. The decision of whether an abstract is structured (or partially structured) is based on finding at least one structure heading in the abstract (some citations have only the *conclusions* section.) Structure headings are recognized using a simple FSM built on the following observations:

- in a section heading either all letters or the first letter of every word is capitalized;
- section headings may contain commas, slashes, and whitespaces in addition to alphabetical characters;
- section headings end in a colon or a dash followed by a space and a capital letter;
- section headings could be found at the beginning of a sentence;
- except for heading *Aim*, section headings are longer than four characters.

The identified headings are assigned to each sentence between the recognized heading and the next.

There are several reasons for taking the discourse structure of a document into consideration, but not relying completely upon it. For example, expecting to find patient outcomes (and only patient outcomes) in the *conclusions* section. These reasons are: 1) the headings only loosely follow the recommendations; 2) only a few recommendations are concerned with the PICO representation; and 3) the abstracts of many high-quality research articles are not structured.

Abbreviation expansion

Before submitting the title and the abstract text of a citation to MetaMap, abbreviations are identified and expanded (if possible) using the local context. Although MetaMap recognizes many abbreviations, the context-based expansion is preferable because many short (3-4 character) strings in the UMLS are ambiguous. For example, *cold* has the following meanings (including one abbreviation):

- cold temperature
- Common Cold
- Cold Therapy
- Chronic Obstructive Airway Disease
- Cold Sensation
- Cold brand of chlorpheniramine-phenylpropanolamine.

In addition, many abbreviations are either not present or incomplete in the UMLS. For example, *TCA* is expanded only to *Turks and Caicos Islands*. The

tricyclic antidepressants meaning intended in question B.2.1.12 is not in the UMLS. To alleviate expansion errors, a local rule-based expansion algorithm is implemented. The rules are based on finding a group of upper-case characters preceded or followed by a parenthetical expression, or an n-character long upper-case string in parenthesis that is preceded by n words starting with characters forming the upper-case string and in that order. Skipping over stopwords such as it in, of, for, and is allowed. After this step, the title and abstract text are submitted to MetaMap/MMTx.

Entity identification (MetaMap/MMTx processing)

MetaMap processing is done by manually submitting batch files through a Web interface, and then downloading the results. UMLS concepts for each phrase, their unique identifiers, semantic types and the position in the original text are recovered from the machine output (shown in Figure 3.3) using regular expressions. There are several inconveniences in this process – a need for a person in the loop, large-size files produced by MetaMap, differences in sentence tokenization in MetaMap and in the URA framework, and the fact that when MetaMap skips over stopwords, or matches a concept approximately, the position of the original string in the text is hard to recover. Many of these inconveniences are resolved in MMTx. MMTx is integrated with CQA-1.0, which allows: (1) sentence-by-sentence processing using URA tokenization; (2) direct communication (no human interaction, no file reading/writing); and (3) recovering the exact positions of the original strings that were mapped to concepts. The disadvantages of using MMTx are: (1) slower process-

ing speed, and (2) insignificantly (for the purposes of this work) lower accuracy of annotations. Using either tool results in a document annotated with all identified occurrences of the UMLS concepts. MMTx was used in all but the pilot experiments described in Section 5.3.1. After the MetaMap/MMTx processing, a document is annotated with everything needed for knowledge extraction and generation of the document frame.

4.4.2 Semantic Processing

The next step in the CQA-1.0 document frame generation is the identification and extraction of the basic elements of the EBM-based semantic domain model: 1) extraction of the elements of a clinical scenario (Knowledge (PICO) extraction); 2) clinical task classification; and 3) strength of evidence classification.

4.4.2.1 Knowledge (PICO) extraction

Each knowledge extractor developed by the author focuses on one element of the clinical scenario (PICO) and processes the title and text of the abstract of a MEDLINE citation. The corresponding MeSH terms, which might or might not be assigned manually by indexers, are not used in this extraction. The following elements are identified: *Population*, *Problem(s)*, *Intervention(s)*, and *Outcome*. All but the Outcome extractor process the text independently. The Outcome extractor uses elements annotated in the other three extractors. The Problem and Intervention extractors select one or more of the concepts identified using MetaMap. There

is no need to implement a comparison extractor separately, because the intervention extractor will capture interventions that are compared. The Population extractor determines the number of people participating in the study and the population group to which they belong. The Outcome extractor selects up to three sentences having high probability of being an outcome. As described below, the Population, Problem, and the Intervention extractors are based largely on recognition of semantic types and a few manually constructed rules; the Outcome extractor, in contrast, is implemented as an ensemble of classifiers trained using supervised machine learning techniques. These two very different approaches can be attributed to differences in the nature of the frame elements: whereas problems and interventions can be directly mapped to UMLS concepts, and population easily maps to patterns that include UMLS concepts, outcomes are complex descriptions of the results of a clinical process. Six base classifiers capture features determined to be important in the outcome recognition by annotators of the PICO-annotated collection (see Section 3.8.)

Population extractor

The PICO framework for question formulation makes no distinction between the population and the problem. A clinician has to identify the presenting problem and select the details of a patient's examination, history, and lab results that characterize the patient and are essential in question focusing. Asking a colleague or using a search engine requires presenting the details, but does not require separating the

problem and the patient's description: a clinician has no difficulties identifying the elements, whereas even the most sophisticated search engines that use the UMLS concepts are unaware of semantic types and PICO elements. However, separating the problems and population in the document analysis phase is important for three reasons: 1) many clinical questions ask about a particular problem without specifying a population; 2) the elements are not always described together in the abstracts; 3) some of the problems found in abstracts could be co-morbid conditions, but not the focus of the study. For example, the sentence *We examined 23 type 2 diabetic patients in a rural and resource-poor area of South Africa.* does not indicate that the study focused on oral health of this population group, specifically on periodontal disease.

Population elements (the number of study participants or observations, and a population group recognized using MetaMap, typically a noun phrase) are identified using manual rules based on the following assumptions:

- The concept involved in the description of population belongs to the UMLS semantic type *Group* or any of its children. For example, *Population Group*, *Patient or Disabled Group*, or *Age Group*.
- Certain nouns are often used to describe study participants in medical texts. For example, an often observed pattern is *subjects* or *cases* followed by a concept from the semantic group *Disorder*.
- The number of subjects that participated in the study often precedes or follows the concept identified as a *Group*. In the latter case, the number is sometimes given in parenthesis using a common pattern *n=number*, where *n=* is a shorthand for the number of subjects, and *number* provides the actual number of study participants.

Given the above assumptions, the population extractor searches for the following patterns:

- Group ([N|n]=[0-9]$^+$) (for example, *in 5-6-year-old French children (n=234)*, *Subjects (n = 54)*)
- number* Group (for example, *forty-nine infants*)
- number* Disorder* Group? (for example, *44 HIV-infected children*)

The population extractor examines each sentence, but does not cross the sentence boundaries. If one of the above patterns is found, it is assigned a confidence score based on the following assumptions:

- The confidence that a clause with an identified number and *Group* contains information about the population is inversely proportional to the distance between the two entities.
- The confidence that a clause contains the population is influenced by the position of the clause, with respect to headings in the case of structured abstracts and with respect to the beginning of the abstract in the case of unstructured abstracts.

The confidence score assigned to a particular pattern match is a function of both its position in the abstract and its position in the clause from which it was extracted. If a number is followed by a measure, for example, year or percent, the number is discarded, and pattern matching continues. After the entire abstract is processed in this manner, the match with the highest confidence value is retained as the population description.

Problems extractor

The problem extractor relies on MetaMap recognition of concepts primarily belonging to the UMLS semantic types assigned to semantic group *Disorder*. The 136,389 concepts in the *Disorder* group have the following sematic types: *Congenital*

Abnormality, *Acquired Abnormality*, *Injury or Poisoning*, *Finding*, *Pathologic Function*, *Disease or Syndrome*, *Anatomical Abnormality*, *Neoplastic Process*, *Mental or Behavioral Dysfunction*, and *Sign or Symptom*. Two semantic types assigned to this group, *Experimental Model of Disease* and *Cell or Molecular Dysfunction*, are not used in the CQA-1.0 processing because they are of limited interest to clinicians. The semantic type *Hazardous or Poisonous Substance* was added to the group to capture problems caused by substances such as most drugs of abuse, and agents that require special handling because of their toxicity, for example, *carcinogens*, *crack cocaine*, or *pesticides*. Although *addiction* and its subordinate concepts are UMLS concepts of the type *Mental or Behavioral Dysfunction*, some articles use terms that are mapped only to *Hazardous or Poisonous Substance*. For example, the phrase *of chronic crack-cocaine use* is mapped as follows:

> Chronic [Temporal Concept], Crack (Crack Cocaine) [Hazardous or Poisonous Substance], COCAINE USE (Cocaine Users) [Population Group]
>
> or as
>
> Chronic [Temporal Concept], Crack Cocaine [Hazardous or Poisonous Substance], use (utilization) [Quantitative Concept].

In both mappings, *crack-cocaine use* that is a problem of interest will be found through its semantic type.

The string representations of several dozens of concepts annotated by MetaMap are suppressed in the CQA-1.0 processing either because the **disorder** sense of the term is no longer used in the literature, for example *consumption* for *tuberculosis*, or because the alternative sense and part of speech is prevalent in the literature but not present in the mappings, for example, *block* is recognized as a verb in *GAL1*

antagonist did not block GAL effect, but still mapped to the *pathologic function Obstruction*. In addition, CQA-1.0 distinguishes general terms that are useful in identifying that an abstract discusses disorders, but not particularly useful being tagged as a problem discussed in the article. These terms are: *disease(s)*, *syndrome(s)*, *sign(s)*, *symptom(s)*, *inflammation*, *pain*, *disorder(s)*, and *finding(s)*. The general terms and the suppressed concepts are assigned zero scores. Each remaining concept is assigned a confidence score, which is a function of its frequency and position in the abstract. Concepts in the title, in the introduction section of structured abstracts, or in the first two sentences in unstructured abstracts are given higher confidence values. The highest-scoring problem (or problems in a tie) is/are designated as the primary problem(s). The co-occurring conditions identified in an abstract are retained in annotation and used in semantic matching with patient description, but not in problem matching.

Interventions/Comparison extractor

The interventions extractor identifies both the intervention and comparison elements. In many abstracts, it is not clearly stated which intervention is the primary one and which are the comparisons, but a ranked list of interventions will reflect the salience of each intervention under study. In defining the intervention and comparison slots, the PICO framework names procedures, agents or other clinician's acts that influence patient's condition. Specific interventions are UMLS concepts belonging to various semantic types. The UMLS does not define a semantic group

for Interventions, but the UMLS Semantic Network defines relations associated with each clinical task. These relations include: *treats*, *prevents*, and *carries out* for *therapy*; *diagnoses* for *diagnosis*; *causes* and *result of* for *etiology*; and *prevents* for *prognosis*. Restrictions on the semantic types allowed in these relations determine the set of possible clinical interventions.

The intervention extraction starts with identification of the following semantic types: *Therapeutic or Preventive Procedure*, *Laboratory Procedure*, *Diagnostic Procedure*, *Health Care Activity*, *Educational Activity*, *Medical Device*, *Clinical Drug*, *Drug Delivery Device*, *Antibiotic*, *Pharmacologic Substance*, *Biomedical or Dental Material*, *Neuroreactive Substance or Biogenic Amine*, and *Steroid*. As with the problem extractor, some of the mappings are suppressed. For example, several MeSH entry terms for *Bisphenol A-Glycidyl Methacrylate (Biomedical or Dental Material)* are suppressed because the brand names of this material, *Conclude* and *Concise*, are prevalently used in their common word sense in the medical literature. Another example is *pace* which is a UMLS-derived synonym of the *cisplatin/cyclophosphamide/doxorubicin/etoposide* chemotherapy protocol. An ongoing word sense disambiguation initiative (Humphrey *et al.* 2006) is devoted to disambiguation of such terms in context during MetaMap processing, which might eliminate the need for the suppressed concepts lists in the future. There are also general intervention terms, such as *medicament*, *treatment*, *intervention*, *regimen*, etc. The candidate scores are assigned as a function of concept's frequency, position in the abstract, and contextual cues. In structured abstracts, concepts of the relevant semantic types are given additional weight if they appear in the title, aims,

or methods sections. In unstructured abstracts, concepts towards the beginning of the abstract text are favored. The score increases if the sentence containing an intervention concept also contains certain cue phrases that describe the aim and/or methods of the study, such as *This ⋆ study examines* or *This paper describes*.

Outcome extractor

Outcome extraction is approached as a classification problem at the sentence level (for each sentence in an abstract, the outcome extractor attempts to estimate the likelihood or probability that it belongs to an outcome statement.) This approach differs from the rule-based extraction of the other frame elements. Preliminary explorations demonstrated that neither a rule-based approach, nor supervised machine learning alone are sufficient for outcome identification. These findings lead to a strategy based on an ensemble of classifiers, which include:

1. a rule-based (cue-terms) classifier,
2. a Naive Bayes (unigram bag-of-words) classifier,
3. an n-gram classifier,
4. a position classifier,
5. a document length classifier, and
6. a semantic (heuristic) classifier.

With the exception of the rule-based classifier, all classifiers were trained on the 275 citations from the annotated collection of abstracts (PICO-train in Table 3.4) , leaving 317 citations with outcome statements and 41 without (358 total, PICO-eval-1 in Table 3.4) for testing. The choice was made after experimenting with the

sizes of the sets following recommendations that the test set size should be 5-10% of the collection size (Manning & Schütze 1999). The preliminary experiments using the WEKA toolkit[1] were conducted as follows: ten iterations of randomly selecting from 633 citations and setting aside 60 as the test set and another 60 citations as the verification set and using the rest for training. In these experiments, the relatively large training set did not improve the classification results over the results obtained using only 275 citations for training. Furthermore, the small size of the test set prevented testing the system performance for each of the four clinical tasks. The subsequent experiments use the 275 citations as the training set, which is sufficient to maintain the performance achieved in the preliminary experiments. In these experiments, a Naive Bayes classifier outperformed both a linear SVM and a decision tree classifier in identifying outcome statements, and was chosen as the baseline classifier for further experiments. Additional preliminary experiments used the state-of-art Naive Bayes classifier provided with the MALLET toolkit (McCallum 2002). This Naive Bayes classifier achieved 100% recall and 27% precision, which lead to creation of a coordinated ensemble of classifiers (training complementary classifiers, and then classifying sentences in each citation using (1) linear interpolation with ad-hoc weights assigned based on intuition and (2) a weighted sum of the classifiers combined in an optimum way using stacking (Ting & Witten 1999).) The base classifiers that contribute to the final score for each sentence operate either locally (on a sentence) or on the whole abstract: the rule-based, Naive Bayes, and n-gram based classifiers treat each sentence disregarding the context of the abstract. The

[1] `http://www.cs.waikato.ac.nz/ml/weka/`

position classifier and the semantic classifier use the abstract structure and context, and the document length classifier operates solely on the number of sentences in the abstract.

The **rule-based (cue-terms) classifier** estimates likelihood of the sentence to be an outcome based on cue phrases such as *significantly greater*, *well tolerated*, and *adverse events*. Knowledge for the rule-based classifier was hand-coded by RN1 (see Section 3.8) prior to the annotation effort. The likelihood of a sentence being an outcome (as indicated by cue terms) is measured by the ratio of the cumulative score for found phrases to maximal possible score. For example, the following sentence: *The dropout rate due to adverse events was 12.4% in the moxonidine and 9.8% in the nitrendipine group* is segmented into eight phrases during MetaMap processing, which sets the maximal possible score to 8, and the two phrases *dropout rate* and *adverse events* contribute one point each to the cumulative score, which results in likelihood estimate of 0.25 for the sentence.

The **unigram bag-of-words classifier** is a Naive Bayes classifier implemented with the API provided by the MALLET toolkit. This classifier outputs the probability of a class assignment. The Naive Bayes classifier treats each sentence as a bag of words and generates the probability of the sentence to be an outcome statement, rather than a binary decision with respect to the class of the sentence being an outcome or not.

The **n-gram classifier** generates the probability in a manner different from the Naive Bayes classifier: whereas the probability assigned by the Naive Bayes classifier is based on probabilities of all words encountered during training, the n-gram

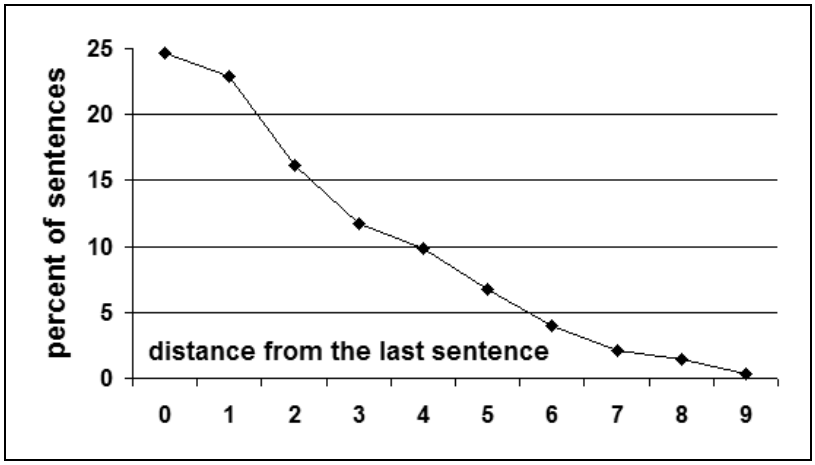

Figure 4.3: Positions of outcome statements in 275 training abstracts.

based classifier uses only features that are strong positive predictors of outcomes. These features were selected as uni- and bi-grams by first identifying the most informative features using information gain measure (Yang & Pedersen 1997), then selecting only positive outcome predictors using odds ratio (Mladenic & Grobelnik 1999), and finally by a manual revision by RN1. During manual revision, the topic-specific terms, such as rheumatoid arthritis, one of the three diseases used to retrieve the training documents, were removed from the feature set to ensure generality of the features. The differences in term selection based on information gain and odds ratio can be illustrated using terms *superior* and *placebo controlled*. Both have a high information gain value, but *superior* also has a high positive odds ratio value and is selected as a feature for the n-gram classifier, as opposed to *placebo controlled* that has a high negative odds ratio, and is therefore discarded.

The **position classifier** returns the maximum likelihood estimate that a sentence is an outcome based on its position in the abstract. The position classifier is based on the discourse structure of the abstract and the relative position of the

sentence in the abstract. As can be seen in Figure 4.3, the likelihood estimate that a sentence contains an outcome statement is very high for the last three sentences of an abstract. This is also true for the sentences in the results and the conclusions sections of structured abstracts. Of the 275 citations used for training, twenty-two (2.5%) were not structured. In the rest, the outcome statements were found in conclusions in 63.6% of the structured abstracts, in the results section of another 36%, and in the interventions section of one abstract.

The **document length classifier** returns a smoothed (add-one smoothing estimating likelihood of lengths unseen in training (Jurafsky & Martin 2000)) probability that a document of given length (in the number of sentences) contains an outcome statement. For example, the probability that a three-sentence long abstract contains an outcome statement is 0.2, and the probability of finding an outcome statement in an 11- to 14-sentence long abstract is 0.95. Implementation of this classifier is motivated by the observed difference between the lengths of abstracts in which the outcome statements were found, and were not found. The average length of the former is 11.7 sentences, whereas the length of the latter is 7.95 sentences on average. The contribution of the document length classifier is essential when selecting documents most likely to contain outcome statements.

The **semantic (heuristic) classifier** generates the maximum likelihood estimate of a given sentence being an outcome statement based on the presence of UMLS concepts belonging to semantic groups highly associated with outcomes, such as *therapeutic procedure*, or *pharmacological substance*. The semantic classifier is global, i.e., it takes into consideration the previously seen content of an abstract

stored during its sequential processing. For example, if the problem and interventions identified in a sentence correspond to the primary problem(s) and top-ranking interventions identified by the problem and intervention extractors, the likelihood that the sentence is an outcome statement increases.

The probabilities and likelihood estimates of being an outcome statement (assigned to a sentence by the base classifiers and used as probabilities) are then combined by the meta-classifier using either a simple weighted linear interpolation scheme with an ad hoc weight selection based on intuitions about accuracy of the base classifiers, or in a more principled way using stacking – a version of least squares linear regression adapted for classification (Ting & Witten 1999). This multiple linear regression (MLR) meta-classifier, which has been shown by Ting and Witten (1999) to outperform other methods of combining classifiers, is described by the following equation:

$$LR(x) = \sum_{k=1}^{N} \alpha_k P_k(x) \tag{4.1}$$

$P_k(x)$ is the probability that sentence x belongs to an outcome statement, as determined by classifier k. To predict the class of a sentence, the probabilities generated by K classifiers are combined using the coefficients (α_0,..., α_k). The coefficients values are determined in the training stage as follows: probabilities predicted by base classifiers for each sentence are represented as a $K \times N$ matrix A, where N is the number of sentences in the training set, and K is the number of classifiers. The reference set class assignments for each sentence are stored in a vector

b, and the coefficients values are found by calculating the vector α that minimizes $||A\alpha - b||$. The coefficients were found using singular value decomposition (SVD), as provided in the JAMA basic linear algebra package released by NIST [2].

The knowledge extraction process results in filling the Problem, Population, Intervention and Outcome slots of the document frame. Each slot contains the surface representation of the element(s), the UMLS identifiers (if applicable), and the score that reflects the CQA-1.0 confidence in the identified element.

4.4.2.2 Clinical Task classification

The identification of the elements of a clinical scenario (PICO knowledge extraction) is followed in the semantic processing by the identification of the clinical task under study. To determine the task-specific orientation of a MEDLINE citation, it is processed using six binary rule-based task classifiers: therapy, prevention, diagnostic methods, differential diagnosis, etiology, and prognosis. Each classifier returns a confidence score that a study described in the article focused on a given clinical tasks. The score for each clinical task is based on the terms that are positive and negative indicators. Positive indicators for each task were derived from: 1) the PubMed Clinical Query filters (Haynes *et al.* 1994; Wilczynski, McKibbon, & Haynes 2001); 2) the JAMA EBM tutorial series on critical appraisal of medical literature 3) MeSH scope notes; and 4) observations. A set of positive indicators for the non-clinical orientation is used as an additional set of negative indicators common to all tasks. These terms were extracted by Susanne Humphrey (Aron-

[2] `http://math.nist.gov/javanumerics/jama/`

son *et al.* 2004a) from the articles provided for the secondary task in the TREC 2004 genomics track evaluation (Hersh, Bhupatiraju, & Corley 2004). The terms, for example, *genetics* and *cell physiology* were originally developed as positive indicators for genomics and other basic scientific research articles. The positive and negative weights assigned to each term heuristically encode the relative importance of different MeSH headings. The task score, S_{task}, is given by:

$$S_{\text{task}} = \sum_{t \in \text{MeSH}} \alpha(t) \tag{4.2}$$

The function $\alpha(t)$ maps a MeSH term to a positive score if the term is a positive indicator for that particular task type, or a negative score if the term is a negative indicator for the clinical task[3]. The highest S_{task} score determines the primary orientation of the study described in the article. Since the classifiers rely on MeSH headings assigned by indexers based on the full text of an article, it is appropriate to assume the task classifiers determine the orientation of the whole article, and not that just of the abstract. Although at present the classifiers use only MeSH headings assigned manually by indexers, should a need arise, the system could rely entirely upon automatic Medical Text Indexing that is currently suggesting terms for indexers' review (Aronson *et al.* 2004b). This might lead to approximately 20% degradation in performance.

[3]Each indicator score in equation 4.2 implicitly includes its coefficient. Appendix E contains a complete list of equations for scoring of the CQA-1.0 components

Indicators and score for therapy task

The examples of strong positive therapy indicators derived from the Clinical Query filters, MeSH scope notes, and JAMA EBM tutorials are: *treatment outcome*, *drug combinations*, *drug therapy*, *therapeutic use*, *surgery*, and *radiotherapy*. A score of 1 is given if the above MeSH descriptor or qualifier is marked as the main theme of the article (indicated via the star notation by indexers), and a score of 0.5 otherwise. The starred non-clinical indicators decrease the score by 1, and by 0.5 otherwise. In addition, two MeSH sub-trees were observed to be weak negative indicators of the task (with a score decrement of 0.1), and one sub-tree as a weak positive indicator (with a score increment = 0.2). The weak positive indicators are *drug administration routes* and any of its children in the MeSH hierarchy. The weak negative indicators are *Health Care Economics and Organizations* and *Health Services Administration*.

Indicators and score for prevention task

In addition to the therapy score determined by the previous classifier, the following MeSH terms are considered positive indicators and each add 0.8 to the score:

MeSH Qualifiers: preventive medicine, primary prevention, life style, risk, risk factors, health behavior, infection control, epidemiologic methods.

MeSH Descriptors: prevention & control, epidemiology, prevention, prophylaxis, preventive therapy, preventive measures, control, preventive therapy.

Indicators and score for differential diagnosis

A single strong positive indicator of differential diagnosis is the term *differential diagnosis*. The remaining diagnosis MeSH terms such as *diagnosis*, *diagnostic use*, *findings*, *examination*, *diagnostic tests*, *predictive value of tests*, *sensitivity and specificity*, etc. are weak positive indicators. The non-clinical and therapy indicators are weak negative indicators for this task (with score decrements 0.4 and 0.1)

Indicators and score for diagnostic methods

Positive indicators for therapy are also used as negative indicators for diagnosis because the relevant studies are usually disjoint. It is highly unlikely that the same clinical trial will study both diagnostic methods and treatment methods. The MeSH term *diagnosis* and any of its children are considered positive indicators. As with therapy questions, MeSH terms marked as the major theme get a score of ± 1.0, and ± 0.5 otherwise. To distinguish clinically oriented *diagnostic methods* from the research oriented *Investigative Techniques*, this term and all its children, for example, *Animal Experimentation*, are used as weak negative indicators, decreasing the score by 0.5. This rule might be changed or, resources permitting, learned in the future when clinical implications of methods in this sub-tree, for example, *Cytogenetic Analysis* will be of practical interest.

Indicators and score for prognosis task

Positive indicators for prognosis include the following MeSH terms: *survival analysis*, *disease-free survival*, *treatment outcome*, *health status*, *prevalence*, *risk factors*, *disability evaluation*, *quality of life*, and *recovery of function*. For terms marked as the major theme, a score of +2 is given; +1 otherwise. There are no negative indicators (other than those common to all tasks).

Indicators and score for etiology task

Negative indicators for etiology include strong therapy-oriented MeSH terms; these terms are given a score of −0.3. Positive indicators for diagnostic methods and differential diagnosis are weak positive indicators for etiology, and receive a positive score of 0.1. The following MeSH terms are considered highly indicative of citations relevant to etiology: *population at risk*, *risk factors*, *etiology*, *causality*, and *physiopathology*. If one of these terms is marked as the major theme, a score of +2 is given; otherwise, a score of +1 is given.

4.4.2.3 Strength of Evidence classification

The semantic processing of a MEDLINE citation concludes with the identification and scoring of the third basic element of the EBM-based semantic domain model – the Strength of Evidence. The Strength of Evidence indicates how influential a given MEDLINE citation should be in contributing to a clinical decision. Several factors determine the Strength of Evidence: 1) the type of the clinical study,

2) the authority and orientation of the journal in which the article was published, and 3) the recency of the publication. Given these factors, the Strength of Evidence score of a citation is determined as a sum of the scores for each factor:

$$S_{\text{SoE}} = S_{\text{study}} + S_{\text{journal}} + S_{\text{date}} \tag{4.3}$$

Strength of the study

Metadata associated with most MEDLINE citations are extensively used in determining strength of evidence, and scoring of its three components. The first component is the type of a clinical study. The potential highest level of the strength of evidence for a given clinical study type can be identified using the Publication Type and MeSH terms pertaining to the type of the clinical study assigned by indexers.

Table 4.3 shows the publication type and MeSH terms mapped to evidence grades according to the principles defined in the Strength of Recommendations Taxonomy (Ebell *et al.* 2004).

Level A publication types and MeSH terms increase the overall study type score by 0.5; Level B, by 0.3; Level C by 0.2. The highest evidence level is used to score citations with several publication types and MeSH terms pertaining to different evidence levels. All non-clinical publications decrease the score by 2. Otherwise, a zero score is assigned to S_{study}.

Table 4.3: **Publication Type and MeSH-based strength of evidence categories.**

Strength of Evidence	**Publication Type/MeSH**
Level A(1)	Meta-Analysis, Controlled Clinical Trials, Randomized Controlled Trials, Multicenter Studies, Double-Blind Method, Cohort Studies, Follow-up Studies
Level B(2)	Studies: Case-Control, Cross-Sectional, Cross-Over, Evaluation, Longitudinal, Retrospective, Case Series
Level C(3)	Case Report, In Vitro, Animal and Animal Testing, Alternatives studies

Journal contribution to the score

Citations published in core and high-impact journals such as the Journal of the American Medical Association (JAMA) get a score of 0.6 for $S_{journal}$. The score increases by 0.3 for citations published in one of the approximately 100 journals most likely to contain patient oriented outcomes (identified by the group of clinicians that developed the Strength of Recommendations Taxonomy), for example, in the American Family Physician journal. The remaining journals get a zero score.

Recency of the study

Finally, recency contributes to the strength of evidence score according to Equation 4.4.

$$S_{\text{date}} = (year_{\text{publication}} - year_{\text{current}})/100 \tag{4.4}$$

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.

The assignment of the Strength of Evidence score concludes the semantic processing.

4.5 Document scoring and ranking

The semantic processing described above results in a set of document frames fully annotated with the semantic domain model elements. The document frames contain: 1) the elements of a clinical scenario (PICO), 2) a set of confidence scores for each clinical task, and 3) a score for the strength of evidence of the study. The document set is now ready to be ranked with respect to its relevance to the question. The ranking takes place in the CQA-1.0 Semantic Matcher module.

Formally, the relevance of a citation with respect to a clinical question includes contributions from matching the PICO elements, the strength of evidence of the citation, and matching of the clinical task that generated the question and the task orientation of the citation:

$$S_{\text{EBM}} = \lambda_p S_{\text{PICO}} + \lambda_s S_{\text{SoE}} + \lambda_t S_{\text{task}} \tag{4.5}$$

With few exceptions, the score components were derived heuristically based on recommendations for critical appraisal of medical literature, intuition and observations on a training set (Appendix E presents all coefficients and indication of how they were set: based on heuristics, or automatically learned using a training set.) The simplest linear combination of scores was used primarily because it is not clear if, and how the three basic components of the semantic domain model interact when applied to document ranking. For the same reason, the λ coefficients are initially set to 1. It is safe to assume that these top-level scores are generated by three different scoring systems, and resort to fusion of the scores. Fox and Shaw (1994) explored different methods for combining scores, and showed the "sum" method to be the best fusion approach. Many fusion methods have been explored since (Zhang *et al.* 2001), however adding the scores is still a viable approach to exploration of fusion (Aronson *et al.* 2005). The successful ranking of citations with respect to their relevance to the question described in Chapter 5 provides empirical evidence that the above approximations capture some important document characteristics. Ideally, the ad hoc scores will be replaced with probabilities derived from MEDLINE data, when enough annotated data becomes available. This, in turn, will allow to explore more sophisticated methods for combining scores of the three components contributing to document ranking.

Only the S_{PICO} score needs to be adjusted with respect to the question. The S_{SoE} and the S_{task} scores derived in semantic processing contribute to the overall score without further adjustments.

4.5.1 Question–Document frame matching. (PICO score)

Matching of the PICO elements in a question and in a document frame is the primary responsibility of the Semantic Matcher. This process results in the assignment of the PICO score to the document. Each extracted PICO element contributes to the score proportionally to its role as a hard or a soft constraint. To reduce the system's susceptibility to automatic mapping errors, the most widely used surface representation(s) of a concept are used in the matching process independent of the concept's presence in the UMLS. The rules for score assignment are built into individual components scoring. There are two types of rules: global (task-independent) and local (task-specific) rules. The individual components' scores are combined linearly according to equation 4.6.

$$S_{\text{PICO}} = S_{\text{problem}} + S_{\text{population}} + S_{\text{intervention}} + S_{\text{outcome}} \tag{4.6}$$

Problem matching and scoring

The first component in the above equation, S_{problem}, depends on a match between the primary problem in the question frame and the primary problem in the abstract (the highest-scoring problem identified by the problem extractor). A score of 1 is given if the problems match exactly based on their unique UMLS identifier as provided by MetaMap. Matching based on concept identifiers, provides for a conceptual match disregarding surface representation of the term used in a document. Failing an exact match of concept identifiers, a partial string match is given a score of 0.5. The string match accounts for cases in which one of the frames contains

a more specific term. For example, if $CUI = C0029456$ is used in the question frame for *osteoporosis* question (C.1.5), but *postmenopausal osteoporosis* found in a document maps to $CUI = C0029458$, the Problem slots are partially matched on *osteoporosis*. If the primary problem in the query has no overlap with the primary problem from the abstract, a score of -1 is given. The initial intent was to remove such citations from the set, however given the accuracy of current tools and the incompleteness of the ontology, the hard binary constraint was replaced with demoting the documents, which resulted in better performance in the exploratory experiments. Finally, if the problem extractor could not identify a problem (but the query frame does contain a problem), a score of -0.5 is given. The primary problem matching rules are global (applied universally independent of the clinical task.)

Co-occurring problems must be taken into consideration in the *differential diagnosis* and *etiology* tasks because knowledge of the problems is typically incomplete in these scenarios. Therefore, physicians might be interested in any problems mentioned in the abstracts in addition to the primary problem specified in the query frame. For example, answering the question: *What is the differential diagnosis of chronic diarrhea in immunocompetent patients?* (B.1.2.4) Although *chronic diarrhea* is the problem that conceivably prompted the patient's visit to the doctor, citations that discuss additional related disorders are instrumental in answering this question. According to local rules for *differential diagnosis* and *etiology*, disorders mentioned in the title receive three points, and disorders mentioned anywhere else receive one point for these (in addition to the match score based on the primary problem, as discussed above).

Population matching and scoring

The population score is global. It is based on the premise that a question frame can contain only one description of a patient. If the patient description matches population identified in a document, the document score is incremented by one. For example, finding the population group *children* from a question frame in the document population slot increments the match score by one. There is no penalty for not matching the patient slot.

Intervention matching and scoring

According to the global scoring rules, for each intervention in the question frame that matches an intervention in a document intervention slot, the intervention score is incremented by one. The intervention score is then normalized (divided by the number of interventions in the question frame), and added to the document score. If no intervention in the question frame matches interventions in the document frame, a score of -0.5 is given. For therapy and diagnosis questions with empty intervention slots, a score of one is given to documents in which interventions with an appropriate semantic type were identified.

Outcome scoring

The outcome score is not based on frame matching even in such rare cases as question B.2.1.5, where the desired outcome of the intended intervention is specified (*pain relief* for *osteoarthritis*). This decision is based on the analysis of the real life

questions that mostly do not specify a high level outcome. More importantly, it is highly unlikely a clinician will chose a surface representations of the desired outcomes that will match those in the article literally, and semantic matching on the outcomes level requires deeper understanding and reasoning about outcomes than is currently available. For example, a high-ranking outcome for the question B.2.1.5 states: *The main finding was that treatment with naproxen and paracetamol is more effective than treatment with higher naproxen doses alone* (Seideman, Samuelson, & Neander 1993). Understanding that *more effective* pertains to pain relief is possible through analysis of outcome measures listed elsewhere in the abstract: *clinical assessment of pain, joint movement, activity of daily life and side-effects were performed at the end of the 5 treatment periods.* A clinician reading this abstract would associate pain intensity ratings within the comfort goals during clinical assessment, as well as being comfortable while performing activities of daily life with *pain relief*, infer that *more effective* means the pain intensity was significantly lower in the group of patients on combination medication than in the naproxen group, and come to the conclusion that *combining NSAIDs and acetaminophen provides more relief of pain in osteoarthritis*, as stated in the reference answer to this question. Although some steps towards automating this reasoning process have been undertaken recently (Harabagiu & Hickl 2006), this level of language processing is beyond the state-of-the-art. Rather then forgo the outcome statements, the score of the highest-ranking outcome sentence generated by the outcome extractor is added to the document score. This decision is motivated by the assumption that outcome statements contain answers to questions, or at least present enough information to

predict whether an answer could be found in the text of the article. Given a match on the primary problem and other elements, all highly ranked patient outcomes are likely to be of interest to the physician.

4.5.2 Document scoring example

A randomized controlled trial of clonidine added to psychostimulant medication for hyperactive and aggressive children.
OBJECTIVE: To compare **clonidine**$_{intervention}$ with placebo added to ongoing **psychostimulant**$_{intervention}$ therapy for the treatment of **attention-deficit/hyperactivity disorder**$_{problem}$ with comorbid oppositional defiant disorder or conduct disorder. METHOD: **Children**$_{population}$ 6 to 14 years of age recruited through 2000 to 2001 were randomized to receive clonidine syrup 0.10 to 0.20 mg/day (n = 38) or placebo (n = 29) for 6 weeks. Primary outcome measures were the Conduct and Hyperactive Index subscales of the parent-report Conners Behavior Checklist. Side effects were monitored using physiological measures and the Barkley Side Effect Rating Scale. RESULTS: Evaluable patient analysis showed that significantly more clonidine-treated children than controls were responders on the Conduct scale (21 of 37 versus 6 of 29; chi2(1) = 8.75, p ¡.01) but not the Hyperactive Index (13 of 37 versus 5 of 29). Compared with placebo, clonidine was associated with a greater reduction in systolic blood pressure measured standing and with transient sedation and dizziness. Clonidine-treated individuals had a greater reduction in a number of unwanted effects associated with psychostimulant treatment compared with placebo. CONCLUSIONS: **The findings support the continued use of clonidine in combination with psychostimulant medication to reduce conduct symptoms associated with attention-deficit/hyperactivity disorder.**$_{outcome}$ Treatment is well tolerated and unwanted effects are transient.

Figure 4.4: PICO elements automatically annotated in the abstract of MEDLINE citation 12874489

Table 4.4 presents the scoring of a MEDLINE citation (Hazell & Stuart 2003) with respect to two question frames. This citation (see Figure 4.4) was retrieved by PubMed queries for two questions:

B.2.1.1 *What is the most effective treatment for ADHD in children?* (therapy)

B.1.2.3 *Does a Short Symptom Checklist accurately diagnose ADHD?* (diagnosis)

The same citation is used to illustrate the Task score assignment (see Table 4.5) and the Strength of Evidence scoring (see Table 4.6.)

Table 4.4: **PICO scores for abstract 12874489 (see Figure 4.4) with respect to two questions about *ADHD*.**

	Question B.1.2.3		Document	Question B.2.1.1	
Slot	Frame	Score	Frame	Score	Frame
Problem CUI	*C1318965*	1	*C1318965*	1	*C1318965*
Problem:	*ADHD*		*ADD*		*ADHD*
Population			*Children*	1	*children*
Intervention:	*checklist*	−0.5	*Clonidine*		
			Psychostimulant	1	
Outcome:		0.78	*The findings...*	0.78	
Task:	*Diagnostic methods*		*Therapy*	*Therapy*	
PICO score		1.28		3.78	

Table 4.5 illustrates the assignment of the clinical task score for *Diagnostic methods* and *Therapy* to a MEDLINE citation. The remaining tasks have a zero score. MeSH terms relevant to the task score assignment are shown in MEDLINE format. MH stands for MeSH heading, sh stands for subheading. Indexing terms are broken up into components with subheadings displayed beneath the corresponding headings, for example, *MH - Methylphenidate/*therapeutic use* fills two rows in the table. To improve readability of the table, some of the index terms are replaced with a shorter Entry Term, for example, *DSM-IV* is used instead of *Diagnostic and Statistical Manual of Mental Disorders*.

Table 4.6 illustrates the Strength of Evidence scoring for the citation presented in Figure 4.4 and Tables 4.4 and 4.5. The journal is not contributing to the score

Table 4.5: **Task scores for MEDLINE citation 12874489 (see Figure 4.4.)**

	MeSH term	Diagnostic Methods	Therapy
MH	Adolescent		
MH	*Aggression		
MH	Attention Deficit Disorder		
sh	complications		
sh	diagnosis	0.5	
sh	*drug therapy	−0.5	1
MH	Central Stimulants		
sh	*therapeutic use	−0.5	1
MH	Child		
MH	Conduct Disorder		
sh	complications		
MH	DSM-IV	0.5	
MH	Drug Therapy, Combination	−0.5	0.5
MH	Female		
MH	Follow-Up Studies		
MH	Humans		
MH	Male		
MH	Methylphenidate		
sh	*therapeutic use	−0.5	1
MH	Sympatholytics		
sh	*therapeutic use	−0.5	1
	Task score:	−0.07	0.2

of this citation because it is not one of the core clinical journals. As indicated by the SB (subset) tag, this journal is indexed as Index Medicus (IM) containing 4,401 journals, but not as Abridged Index Medicus (AIM), a list of about 120 core clinical English language journals. This journal is also not listed as likely to be regularly reviewed for patient oriented evidence by the developers of the Strength of Recommendations Taxonomy.

Table 4.7 presents the final score for the citation with respect to two *ADHD* questions. These scores promoted the answer (the extracted outcome) to rank four for the therapy question, and to rank 22 for the diagnosis question (from the original

Table 4.6: **Strength of Evidence(SoE) score for MEDLINE citation 12874489 (see Figure 4.4.)**

Metadata		Score
DP –	2003 Aug	$(2003 - 2006)/100 = -0.03$
SB –	IM	
MH –	Follow-Up Studies	
PT –	Clinical Trial	
PT –	Journal Article	
PT –	Randomized Controlled Trial	0.5
JT –	Journal of the American Academy of Child and Adolescent Psychiatry	
	SoE score:	0.47

Table 4.7: **Final score for citation 12874489.**

EBM model component	Diagnosis Score	Therapy Score
PICO score	1.28	3.78
Task score	−0.07	0.2
SoE score	0.47	0.47
Total score	1.68	4.45

rank 34 in both retrieved sets.) The corresponding key point in the reference answer for the therapy question states: *The combination of methylphenidate and clonidine (Catapres) improves symptoms in children with both ADHD and tics*, which is fairly close to the CQA-1.0-generated answer. The rank change for the diagnosis question is not influencing the answer.

The output of the Semantic Matcher is a list of documents fully annotated semantically and ranked with respect to the question frame. Answer generation based on this list is presented next.

4.6 Answer generation

Ely et al. (2005) through observing and interviewing clinicians found that the most desirable form of an answer is bottom-line clinical advice. Popularity and financial success of the secondary sources that present bulleted key messages confirms this observation. Manually generated answers integrate information from multiple clinical studies, pointing out both similarities and differences. An automatic system should follow this design; it should detect and eliminate redundancy and provide the best representation for multiply encountered findings. The system should also detect controversial evidence and present it to the clinician.

Some of these desirable features are implemented in the CQA-1.0 system. The CQA-1.0 system eliminates redundancy through semantic clustering on problems and interventions. Clustering of the outcome statements and finding controversial evidence in the outcome statements are hard problems and probably beyond the current state of the art in question answering and multi-document summarization. An exploration of deeper understanding of the document using SemRep that would potentially allow for controversy and comparison detection and enable outcome clustering is described in Appendix D.

The CQA-1.0 system generates answers of two types: 1) multi-tiered answers and 2) best answers. The goal of the multi-tiered answer generated by the CQA-1.0 system is to provide an overview of available information, which is appropriate for the majority of clinicians' information needs. However there are cases when an overview is not needed. This situation is similar to a known item search. In general,

when searching for a known item, the user knows of a particular document, but does not know where it is. In clinical practice that could actually be the case, but more often clinicians want to verify that their recollection of a fact is correct. For example, they might want to verify that there is a contraindication for a generally accepted treatment in a certain group of patients. In such situations generating a short list of answers or a single best answer is appropriate.

The semantic representation of each document and the EBM-based ranking described in Section 4.5 provide a means for extracting and presenting both answer types. A TREC QA model is used to generate answers for a known fact confirmation: outcome statements from top N documents are extracted as a short list of likely answers to the best of the system's knowledge. In this case, answer generation amounts to displaying the title and the annotated outcome statements from the top ranking documents, since all information is already available in the CQA-1.0 document frames.

An overview answer is generated using clustering – a method known to provide a good overview of data and often used to visualize and interactively explore large document collections and knowledge bases (Card, Mackinlay, & Shneiderman 1999). To generate this answer type, documents discussing the same intervention, or interventions belonging to the same drug class are identified. A list of interventions serves as the top-tier answer. Each intervention is supplemented with supporting evidence (the outcome statement extracted from the top ranking citation in the cluster), and with the ranked list of citations in the cluster. This approach provides a full answer that is hypothesized to be better suited for the domain. The implemented CQA-1.0

system provides a possibility to test this hypothesis in the future, conducting user experiments with both types of answers.

4.6.1 Semantic clustering

The retrieved MEDLINE citations are organized into semantic clusters based on the main interventions (interventions with top scores) identified in the abstract text by the intervention extractor (See Section 4.4.2.1), and using hierarchical agglomerative clustering based on the UMLS hierarchical relationships (Demner-Fushman & Lin 2006b; 2006a). The clustering process starts by placing each concept in its own group (with N identified interventions, each in its own cluster.) Iteratively, interventions that fall under a common parent (a UMLS hypernym), are grouped together, ascending the UMLS hierarchy in the process. For example, rofecoxib would be grouped with ibuprofen because they were both Anti-Inflammatory Agents according to UMLS. The process is applied until no new clusters can be formed. In order to preserve granularity at the level of practical clinical interest, the tops of the UMLS hierarchies were truncated; for example, the MeSH category *Chemical and Drugs* is too general to be useful. This truncation process was performed manually by the author prior to the evaluation described in Section 5.4.2. This crude stopping condition might be replaced in the future with a semantic distance metric along the lines of the metrics originated in (Resnik 1999). An abstract may appear in multiple clusters if more than one intervention was identified (for example, if the abstract compared the efficacy of two treatments that belong to different se-

mantic types.) The most general ancestor concept, for example, antibiotics, is then used as the cluster label. Zhao and Karypis (2002) demonstrated that the overall computational complexity of a naive agglomerative clustering approach, known to be $O(n^3)$, can be reduced to $O(n^{2/3} \log n)$, if the number of intermediate clusters is sufficiently large (on the order of $\sqrt{n}$) and the space over which agglomeration decisions are made is constrained, so that each document is only allowed to merge with other documents that are part of the same partitionally discovered cluster. To satisfy these conditions, the clustering algorithm is implemented as follows: each initial cluster is expanded with all its ancestors, looked up in a hash table. In the subsequent iterations the clusters are merged only if they share a common ancestor, thus agglomerating only within a partition.

Once the clusters have been formed, the citations need to be sorted within a cluster, and the presentation order of the clusters must be determined as well. These two questions would each be a complex study on their own. A preliminary decision was made to order clusters by size, assuming that the more important intervention types warranted more studies, and rank by document scores (or recency for ties) within the cluster. Independent of the order of presentation, a four-tier answer is then generated for each question. The first tier is a list of cluster labels, for example, for the question *What are the interventions for otitis externa?* (C.2.20) the following labels are presented as the first tier in this order:

+ anti-infective agents
+ antimicrobials
+ hormones

+ physically based treatment method
+ musculoskeletal medications
+ generic operative procedures

Expanding, for example, the anti-infective group, the answers are *ciclopiroxolamine, ofloxacin, and boric acid*. In the majority of the citations in this cluster ofloxacin is identified as the main intervention. Following the second-tier answer ofloxacin, the third tier presents the following answer (the top-scoring outcome sentence):

> Ofloxacin given twice daily is as safe and effective as Cortisporin given 4 times daily for otitis externa. [SOE-1: Based on Randomized Clinical Trial].

The outcome sentence (and indeed, the full abstract of the article) is an insufficient basis for a clinical decision, but it can serve as an entry point into the medical literature, which the physician can explore further in depth. Information about citations contributing to the answer is available in a list of unique document identifiers. Given these identifiers the abstracts, many of which provide access to full text, could be examined next.

The first-tier answers (cluster labels) represent 34 citations retrieved for this question, and the main intervention *ofloxacin* represents five documents. At present, the relationships between *ofloxacin* and the other two second-tier answers is not known. The second-tier answer would benefit from a quality ranking and/or information about relationships between the listed interventions (similar to the manually generated answer, the top tier of which is shown in Figure 4.5.) For example, if interventions A, B, and C are present in a cluster, a sophisticated system should be

Ear, nose, and throat disorders

Otitis externa

Search date July 2003

Daniel Hajioff

QUESTIONS

Effects of empirical treatment

INTERVENTIONS

Likely to be beneficial

Topical aluminium acetate drops (as effective as topical anti-infective agents)
Topical anti-infective agents (antibiotics or antifungals with or without steroids)
Topical steroids

Unknown effectiveness

Oral antibiotics
Specialist aural toilet

Unlikely to be beneficial

Oral antibiotics plus topical anti-infective agents (no better than topical anti-infective agents alone)

Footnote

See glossary

To be covered in future updates

Prophylaxis for otitis externa
Surgery for ear canal stenosis after otitis externa
Treatment for necrotising otitis externa

Figure 4.5: Otitis Externa key points in BMJ Clinical Evidence.

able to generate an answer saying A is as good as B, and better than C (or rank the interventions similarly to the CE collection). This task turned out to be fairly complex, primarily because of the sparseness of comparative studies. For example, the *ofloxacin-cortisporin* comparison above is the only comparison for ofloxacin. *ciclopiroxolamine* and *boric acid* were never compared to it directly. These terms do not even co-occur in a single citation, although there are 5528 citations containing *ofloxacin*, 997 containing *boric acid*, and 220 containing *ciclopiroxolamine*. The only MEDLINE citation that contains both *cortisporin* and *boric acid* is not a comparison; moreover *cortisporin* is present only in the metadata list of chemicals, but not in the abstract text. Despite these difficulties, a preliminary investigation of SemRep comparative processing indicates that upon completion, it will often provide information for ranking of the top-tier answers (for interventions that were compared in clinical studies.)

Summary

The question answering process implemented in the CQA-1.0 system follows these steps:

- ▷ Clinical questions are translated to question frames and PubMed queries.
- ▷ MEDLINE documents are retrieved and translated to CQA-1.0 document frames.
    - ▷ Knowledge extractors fill the population, problem(s), intervention(s), and outcome slots of the document frame.
    - ▷ Each clinical scenario element is scored, and preserves a pointer to its position in the citation.

- ▷ The result of the semantic processing is a set of document frames fully annotated with the elements of the three components of the semantic domain model: PICO, Clinical Task, and Strength of Evidence.

- ▷ The semantic matcher module (see Section 4.5.1)compares the document frame with the question frame described in Section 4.2. The scoring and ranking of the document frames in the semantic matcher concludes the processing of the documents retrieved to answer the question.

- ▷ The desired form of the answer determines the next processing step:

  - ▷ For the best answer, the title of the abstract and the top three outcome sentences in the order they appeared in the abstract are combined, then the answers from n-best citations are returned without any further processing.
  - ▷ Multi-tiered answers are generated using hierarchical agglomerative clustering of the annotated citations described in Section 4.6.1. The description and the evaluation in Section 5.4.2 focus on interventions, primarily because this will answer all therapy and diagnostic methods questions, (over 80% of the clinical questions overall (Ely *et al.* 1999)), however the algorithm can be applied to any UMLS semantic group without any changes.

Chapter 5

System evaluation

Ideally, a clinical question answering system should be judged by its impact on patient outcomes. A conceivable ethical way to conduct such evaluation is to measure patient outcomes before and after a system became available to a group of clinicians. An approximation of this approach is to ask experienced clinicians if the answers are correct and if they have a potential to influence clinicians' decisions. This approximation is used in practice in medical informatics for evaluation of manually generated secondary sources (Alper, White, & Ge 2005). It was approximated in the instructions for the document- and answer evaluations described below.

One of the difficulties in working in a fairly unexplored domain such as evaluation of clinical and biomedical question answering systems is the absence of established test collections and evaluation metrics. However, the availability of online resources that to various extent contain components of a test collection, permitted creation of several collections for the purposes of the summative evaluation of the system (see Table 3.4.) In addition to the system evaluation, it is also important to know how well individual modules perform their tasks, and how much every component contributes to the overall quality of an answer. This chapter first describes the metrics and statistical tests used in the evaluation and then presents the evaluation of system components. Using PubMed for document retrieval requires re-ranking of

retrieved documents as the first step in answer generation. semantic-model-based document re-ranking is evaluated and compared with different baselines. Two manual evaluations for two types of answer generation – extracting the best answer, and emulation of interactive multi-tiered answer examination based on semantic clustering are described next. Known automatic evaluation methods applied to clinical question answering conclude the description of evaluation design, results, and analysis.

5.1 Evaluation metrics

Several evaluation metrics (Baeza-Yates & Ribeiro-Neto 1999) accepted as de-facto standards in one of the fields that contribute to question answering are used in all evaluations. These are:

Precision: The fraction of the retrieved documents that is relevant.

Recall: The fraction of the relevant documents that is retrieved.

Mean Average Precision (MAP): For multiple topics, it is the mean of the average precision scores for each of the topics. The Average Precision score for a single topic is computed by averaging the precision after each relevant document is retrieved (Harman 1996). This metric has recall and precision components and is widely-accepted in information retrieval as reflecting the level of performance a user should expect for a new topic retrieved using a system that achieves a given MAP value.

Bpref: A preference-based measure that depends on the number of judged non-relevant documents retrieved before the relevant ones, as opposed to MAP that is determined by the ranks of the relevant documents in the result set and makes no distinction between documents explicitly judged as not relevant and documents that are not judged (Buckley & Voorhees 2004). This measure is reported to be more stable than MAP with incomplete judgments, which is probably the case for the pilot studies presented below.

R precision: measures precision after R documents have been retrieved, where R is the total number of relevant documents for a query.

Precision at five retrieved documents (P@5): measures the fraction of relevant documents in the top five results.

Precision at ten retrieved documents (P@10): measures the fraction of relevant documents in the top ten results.

Mean Reciprocal Rank (MRR): is the metric used in TREC QA evaluation. It quantifies the "expected search length". It is computed as the mean of the individual questions Reciprocal Ranks. Reciprocal Rank of top relevant document is the reciprocal of the rank at which the first relevant document was found.

Statistical tests

Statistical tests are needed to determine if the differences in evaluation scores reflect differences in the systems' performance, or occurred by chance. The parametric methods are often not applicable to results of IR experiments because of the small to moderate sample size, and because the evaluation metrics are discrete. In such cases, non-parametric tests are used to evaluate significance of the differences in the results. When the values in the two results being compared are naturally paired (for example, the same document is ranked by two systems), and the relative magnitude as well as the direction of the differences is considered, the Wilcoxon signed ranks test is used (Siegel & Castellan 1988).

To test whether two samples are from the same population when the underlying distributions are unknown, the Kolmogorov-Smirnov two-sample test is recommended, especially to determine if the samples differ in any respect (the most powerful test). This non-parametric alternative to the t-test uses distributions of the data rather than ranking. This test is less sensitive to the data ordering (Siegel & Castellan 1988). This metric is better suited to test if the best answers generated

by the system outperformed the baseline. In all other evaluations the Wilcoxon signed ranks test was used.

5.2 Evaluation of knowledge extractors

Knowledge extractors were evaluated on the held-out set of the manually annotated abstracts described in section 3.8. Because of the different nature of PICO elements, each of the evaluations is element-specific and will be described separately.

Evaluation of the population extractor

The PICO-eval-2 collection (see Section 3.8 and Table 3.4 was used to evaluate the population extractor. The output of the population extractor was judged to be correct if it occurred in a sentence that was annotated as containing the population in the reference standard. This evaluation is lenient, and represents an upper bound on the performance of a population extractor that outputs noun phrases. It was adopted because annotators were asked to annotate whole sentences, since it was very hard and time consuming to indicate exact boundaries of a population being studied. For comparison, the first three sentences of an abstract were considered as a baseline. The baseline was considered to be correct if any one of the sentences was annotated as containing the population in the reference standard (an even more lenient criterion). This baseline was motivated by the observation that the aim and methods sections of structured abstracts, which roughly correspond to the first three sentences of unstructured abstracts, are likely to contain the population information.

Table 5.1: **Accuracy of the population extractor**

	Correct		**Unknown**		**Wrong**	
	RS	RN1	RS	RN1	RS	RN1
Baseline	53.3%	51.1%			46.7%	48.9%
Extractor	80%	78.9%	10%	11.1%	10%	10%

The accuracy of the population extractor is shown in Table 5.1. Since the reference standard (RS) includes the author's judgments, an additional evaluation of accuracy with respect to RN1 judgments is provided. Little difference in accuracy with respect to the two judgments is observed due to the fact that RN1 had the best intra-annotator agreement ($kappa = 0.95$ for the population annotation.) A manual error analysis revealed three sources of error:

1. Not all population descriptions contain a number explicitly, e.g., *The medical charts of all patients who were treated with etanercept for back or neck pain at a single private medical clinic in 2003.*

2. Not all study populations are population groups, as for example in *All Primary Care Trusts in England.*

3. Part of speech tagging and chunking errors propagate to the semantic type assignment level and affect the quality of MetaMap output, as for example, in the following sentence:

   > We have compared the LD and recombination patterns defined by singlenucleotide polymorphisms in ENCODE region ENm010, chromosome 7p15 2, in Korean, Japanese, and Chinese samples.

Both *Korean* and *Japanese* were tagged as nouns, which led to the following erroneous chunking:

> [We] [have] [compared] [the LD] [and] [recombination patterns] [defined] [by single-nucleotide polymorphisms] [in] [ENCODE] [region ENm010,] [chromosome 7p15 2,] [in Korean,] [Japanese,] [and] [Chinese samples.]

which led to the tagging of *Japanese* as a population. Errors of this type affect other extractors too. For example, *lead* was mis-tagged as a noun in the phrase

Echocardiographic findings lead to the right diagnosis, which caused MetaMap to identify the word as a Pharmacological Substance (lead, the metal, is sometimes used as a homeopathic preparation).

Evaluation of the problem extractor

Although the problem extractor returns a list of clinical problems, only performance on identification of the primary problem was evaluated. Because all searches were conducted for a specific problem, assessors were not asked and therefore did not explicitly annotate problems in the abstracts. Assuming that because the search was conducted for a certain problem, it therefore is the main problem is incorrect: PubMed is a Boolean search engine: a mere presence of a term anywhere in the abstract text is sufficient for a citation to be retrieved. However, for some abstracts, MeSH headings can be used as ground truth, since one of the human indexers' tasks in assigning terms is to identify the main topic of the article (sometimes a disorder). MeSH terms are not used in problem identification; therefore the starred MeSH descriptors can be used as the ground truth for the problem extractor. The problem extractor evaluation is based on fifty randomly selected abstracts with disorders indexed as the main topic from the set retrieved using PubMed to answer five clinical questions described in (Sneiderman *et al.* 2005).

The problem extractor was applied on different segments of the abstract: the title only, the title and first two sentences, and the entire abstract. These results are shown in Table 5.2. Here, a problem was considered correctly identified only if

Table 5.2: **Accuracy of the problem extractor**

	Correct	**Unknown**	**Wrong**
Abstract title	85%	10%	5%
Title + 1st two sentences	90%	5%	5%
Entire abstract	86%	2%	12%

it shared the same concept id as the ground truth problem (from the MeSH heading). The performance of the best variant (abstract title and first two sentences) approaches the upper bound of 93% on MetaMap performance – which is limited by human agreement on the identification of semantic concepts in medical texts, as established in (Pratt & Yetisgen-Yildiz 2003). Although problem extraction largely depends on disease coverage in UMLS and MetaMap performance, the error rate could be further reduced by more sophisticated recognition of implicitly-stated problems. For example, with respect to a question about immunization in children, an abstract about the measles-mumps-rubella vaccination never mentioned the disease without the word vaccination; hence, no concept of the type Disease or Syndrome was identified.

Evaluation of the intervention extractor

The intervention extractor was evaluated in the same manner as the population extractor (PICO-eval-2 collection) and compared to the same baseline with respect to the reference standard (RS) and RN1 judgments.

The output of the intervention extractor was judged to be correct if it occurred in a sentence that was annotated as containing the same intervention in the refer-

Table 5.3: **Accuracy of the intervention extractor**

	Correct		**Unknown**		**Wrong**	
	RS	RN1	RS	RN1	RS	RN1
Baseline	60%	55.6%			40%	44.4%
Extractor	80%	76.7%			20%	23.3%

ence standard. As with the evaluation of the population extractor, this represents an upper bound on performance. Results are shown in Table 5.3. The difference in the accuracy with respect to the two judgments is somewhat greater than for the population extractor. RN1 intra-annotator agreement was lower for this annotation ($kappa = 0.92$.) Some of the errors were caused by ambiguity of terms. For example, in the clause *serum levels of anti-HBsAg and presence of autoantibodies (ANA, ENA) were evaluated*, *serum* is recognized as a *Tissue*, levels as *Intellectual Product*, and *autoantibodies* and *ANA* as *Immunologic Factors*. In this case, however, *autoantibodies* should be considered a *Laboratory or Test Result*. MetaMap does provide alternative candidate mappings, but the current extractor considers only the best candidate.

In other cases, extraction errors were caused by summary sentences that were very similar to intervention statements, e.g.,

> This study compared the effects of 52 weeks' treatment with pioglitazone, a thiazolidinedione that reduces insulin resistance, and glibenclamide, on insulin sensitivity, glycaemic control, and lipids in patients with Type 2 diabetes.

For this particular abstract, the correct intervention is contained in the sentence

> Patients with Type 2 diabetes were randomized to receive either pioglitazone (initially 30 mg QD, n = 91) or micronized glibenclamide (initially

1.75 mg QD, n = 109) as monotherapy.

Evaluation of the outcome extractor

The PICO-eval-1 test set was used to evaluate the outcome extractor. The set contains 358 citations, which were evaluated as a whole, and with respect to clinical tasks (153 for therapy; 37 for diagnosis; 111 for prognosis; and 57 for etiology.)

Table 5.4: **Accuracy of the outcome extractor with respect to reconciled judgments for PICO-eval-1 test set (B = baseline, returns *N* last sentences in abstract; AH = *ad hoc* weight assignment; LR = least squares linear regression. Numbers following the abbreviated extractor names correspond to the number of sentences in the outcome statement. Statistically significant improvement over the baseline ($p < 0.01$) is shown in bold)**

	B1	B2	B3	AH1	AH2	AH3	LR1	LR2	LR3
Etiology	34.5%	63.6%	78.2%	47.4%	68.4%	82.5%	52.6%	**73.7%**	**87.7%**
Diagnosis	44.4%	72.2%	75.0%	56.8%	70.3%	78.4%	67.6%	78.4%	**89.2%**
Therapy	38.6%	74.0%	75.0%	49.0%	75.0%	**95.0%**	51.0%	77.0%	**92.8%**
Prognosis	49.5%	73.0%	84.7%	63.1%	75.7%	87.4%	60.4%	**79.3%**	89.2%

Table 5.5: **Accuracy of the outcome extractor with respect to RN1 judgments for PICO-eval-1 test set.**

	B1	B2	B3	AH1	AH2	AH3	LR1	LR2	LR3
Etiology	35.1%	59.6%	71.9%	59.6%	71.9%	**84.2%**	56.1%	66.7%	82.5%
Diagnosis	35.1%	70.3%	78.4%	70.3%	72.9%	89.2%	70.3%	81.1%	**91.9%**
Therapy	39.9%	58.7%	74.1%	52.4%	77.6%	90.8%	51.0%	79.7%	**91.5%**
Prognosis	34.2%	64.9%	86.5%	50.5%	72.1%	89.2%	51.4%	72.9%	**90.1%**

The output of the outcome extractor is a ranked list of sentences sorted by confidence. Based on the observation that annotators typically mark two to three

sentences in each abstract as outcomes, the extractor performance was evaluated at cutoffs of one, two, and three sentences. The results of outcome identification are shown in Table 5.4, where numbers 1 through 3 indicate the sentence cutoffs in selecting sentences with top scores assigned by the outcome classifiers. The columns marked AH1, AH2, and AH3 show performance of the weighted linear interpolation approach with ad hoc weight assignment at one-, two-, and three-sentence cutoffs, respectively; the columns marked LR1, LR2, and LR3 show performance of the least squares linear regression model at the same cutoffs. In the evaluation, the extracted outcome was considered correct if it contained at least one sentence judged as belonging to the outcome statement by annotators. This lenient evaluation was adopted because of the importance of pointing the physician in the right direction, even if the results are only partially relevant. Because the author participated in the reconciliation of the reference standard, Table 5.5 presents the accuracy of outcome identification with respect to RN1 judgments. Since outcome statements are typically found in the conclusion of a structured abstract (or near the end of the abstract in the case of unstructured abstracts), the baseline of returning either the final sentence, or two or three final sentences in the abstract (B1, B2, and B3 in Table 5.4 served as a comparison for the outcome extractor. As can be seen, the outcome extractors perform better than the baseline in few cases at the two-sentence cutoff. Improvements can also be seen at the three-sentence cutoff level. The assignment of weights in the *ad hoc* model is primarily geared towards therapy questions. Better overall performance is obtained with the least squares linear regression model.

Table 5.6: **Precision of the Stacking Meta-Classifier for Each Major Clinical Task**

Task	Etiology	Diagnosis	Therapy	Prognosis
Precision	0.30	0.37	0.31	0.36

Table 5.6 presents precision of outcome identification (three best sentences generated using least squares linear regression.)

The majority of errors in the outcome extractor were related to inaccurate sentence boundary identification, chunking errors, and word sense ambiguity in the Metathesaurus. The top-ranking three sentences were used as extracted answers in the best answer evaluation (see Section 5.4.1.)

5.3 Re-ranking evaluation

As the system relies upon PubMed to retrieve a manageable set of citations for further processing, and then re-ranks citations according to the scoring algorithm described above, it is important to verify that re-scoring actually improves upon PubMed results, and performs comparably to, or better than existing state-of-the art search engines. The methodology for such evaluations is well developed in TREC and is applied in the re-ranking evaluation without any modifications. A pilot evaluation was performed at the early stages of system development. In these experiments, CQA-1.0 re-ranking was compared with PubMed and two other experimental systems that use domain knowledge for document ranking. The relevance judgments for these experiments were provided by Dr. CS unfamiliar with

implementation of the systems. in a larger scale evaluation CQA-1.0 re-ranking was compared with that of the two well-known search engines that consistently perform well in TREC.

5.3.1 Pilot document re-ranking experiments

The pilot re-ranking experiments compare four systems:

- PubMed as baseline (results sorted by recency.)
- Essie, a probabilistic search engine developed for NLM's Clinical Trials database.
- SemRep (ranking based on semantic propositions identified in citations.)
- CQA-1.0 (EBM-semantic-domain-model-based ranking.)

The pilot experiments use 15 *FPIN* questions and relevance assessments by Dr. CS. The questions are divided into three sets by the number of PICO elements present in the questions: all questions in the first set ask *what is the best treatment for Problem X?*, problems being *genital warts*, *acute low back pain*, *panic disorder*, *osteoporosis*, and *obesity in children*. The last question specifies *Population* as well. Questions in the second set contain at least two PICO elements. All questions contain a *Problem*; three contain *Population*; four ask about specific *Interventions*; in addition, two pertain to the prevention aspect of therapy. Questions in the last set contain at least three PICO elements:

- *Antiviral agents*$_{intervention}$ for *pregnant women*$_{population}$ with *genital herpes*$_{problem}$.
- *Intravenous fluids*$_{intervention}$ for *children*$_{population}$ with *gastroenteritis*$_{problem}$.
- What is the best treatment$_{intervention}$ for *gastroesophageal reflux and vomiting*$_{problem}$ in *infants*$_{population}$?

- Is $methylphenidate_{intervention}$ useful for treating $adolescents_{population}$ with $ADHD_{problem}$?
- What are the best therapies$_{intervention}$ for $acute\ migraine_{problem}$ in $pregnancy_{population}$?

Table 5.7: **Pilot evaluation of domain model based document re-ranking. (Best results shown in bold.)**

System	MAP	Bpref	R-prec	MRR	P@5	P@10
Question set 1: broad questions						
PubMed	0.141	0.515	0.119	0.622	0.240	0.180
Essie	**0.440**	**0.609**	0.378	**1.000**	**0.880**	0.840
SemRep	0.292	0.370	0.324	0.767	0.840	0.840
CQA-1.0	0.413	0.469	**0.422**	**1.000**	0.840	**0.880**
Question set 2: intermediate questions						
PubMed	0.364	0.400	0.339	0.840	0.640	0.560
Essie	0.299	0.371	0.326	0.800	0.680	0.640
SemRep	0.362	0.397	0.416	**1.000**	**0.800**	0.720
CQA-1.0	**0.440**	**0.547**	**0.437**	0.867	0.720	**0.760**
Question set 3: specific questions						
PubMed	0.413	0.294	0.453	0.867	0.640	0.580
Essie	0.536	0.636	0.539	**1.000**	0.880	0.860
SemRep	0.216	0.234	0.252	0.800	0.720	0.500
CQA-1.0	**0.629**	**0.720**	**0.601**	**1.000**	**1.000**	**0.980**

A total of 1305 documents for the first set, 925 for the second, and 959 for the third set were retrieved from MEDLINE using PubMed. These documents were re-ranked using SemRep and the CQA-1,0 system. Essie searches were conducted on MEDLINE at the same time. Relevance judgments were generated using pooling strategy developed for TREC. The top ten documents from each system, including PubMed, were evaluated by Dr. CS. The trec_eval-8.0 package was used to evaluate the results. Because of the size of the pool and the number of questions, the results of these pilot experiments (see Table 5.7) are suggestive but not deterministic of the future performance of the system. For 15 questions, the CQA-1.0 improvement

over PubMed is statistically significant ($p < 0.01$). As can be seen, mean average precision is not always improved by semantic re-ranking, particularly, if based solely on the match on the clinical scenario elements (SemRep). The pilot experiment suggests that the other two components of the EBM-based semantic domain model contribute to re-ranking. This hypothesis is tested in the experiments described below. Judging by the reciprocal rank of the top retrieved document, and precision at five and ten documents, semantic re-ranking is necessary when answers are extracted from the top documents retrieved by PubMed. However, using Essie retrieval results might alleviate the need for re-ranking. An in-depth study would be needed to answer this question.

Table 5.8: **Pairwise agreement (Cohen's kappa) of four MDs for document relevance assessment**

	CS-KWF	CS-DF	CS-DN	KWF-DF	KWF-DN	DF-DN
3-point	0.44	0.50	0.51	0.50	0.53	0.50
binary	0.60	0.66	0.66	0.73	0.70	0.53

Two questions with respect to re-ranking need to be answered to the extent to which it is possible in the current work: will the results obtained on a small sample scale to a representative number of topics, and given evaluations are based on opinion of one judge, how well will the proposed semantic re-ranking satisfy information needs of another clinician? To answer the second question four MDs (CS, KWF, DF – the author, and DN) evaluated the top 10 documents retrieved by PubMed to answer five diagnosis questions in the *FPIN* collection. MEDLINE abstracts were evaluated on a three point scale: containing an answer, topically

relevant, not relevant. Pair-wise agreement between annotators for three-point and binary scale evaluation was computed using Cohen's kappa. (see Table 5.8) For the binary evaluation (considering an abstract to be relevant or not) documents containing an answer and topically relevant were grouped together. Despite some variations, agreement between annotators is fairly uniform. It has to be noted that although all MDs have over 15 years of experience and hold an equivalent of Board of Medical Specialties certification in their respective specialties/countries, two were educated and practiced abroad, and all are specialists in different fields, which quite possibly is the reason for only moderate agreement.

The remaining questions:

- Are the CQA-1.0 re-ranking results scalable and repeatable?
- How successful is the CQA-1.0 system re-ranking compared to using well-known state-of-the art search engines? and
- How do individual components of the semantic domain model contribute to the improvement?

will be answered next. Some of the runs and results in these experiments were obtained by Jimmy Lin, and will be marked as such.

5.3.2 Document re-ranking experiments

Re-ranking of retrieval results for 50 *FPIN* questions was compared to In-Query, based on a probabilistic retrieval model (Callan, Croft, & Harding 1992) and Indri, a language-model based search engine (Metzler & Croft 2004). These publicly available search engines consistently perform well in TREC evaluations. All Indri runs were performed by Jimmy Lin (Lin & Demner-Fushman 2006).

Table 5.9: **Results of re-ranking experiments for FPIN-train and FPIN-eval-1 sets.**

FPIN-train	Therapy	Diagnosis	Prognosis	Etiology	All
			MAP		
PubMed	0.354	0.421	0.385	0.608	0.428
InQuery	0.327	0.210	0.123	0.422	0.292
Indri	0.706	0.521	0.502	**0.686**	0.630
CQA-1.0	**0.819**	**0.794**	**0.635**	0.649	**0.754**
			P10		
PubMed	0.300	0.367	0.400	0.533	0.378
InQuery	0.590	0.483	0.433	0.500	0.525
Indri	0.620	0.483	0.467	**0.613**	0.565
CQA-1.0	**0.730**	**0.800**	**0.633**	0.553	**0.699**
			MRR		
PubMed	0.428	0.792	0.733	0.900	0.656
InQuery	0.683	0.750	0.667	**1.000**	0.764
Indri	0.900	0.756	**0.833**	**1.000**	0.876
CQA-1.0	**0.933**	**0.917**	0.667	**1.000**	**0.910**
FPIN-eval-1	Therapy	Diagnosis	Prognosis	Etiology	All
			MAP		
PubMed	0.421	0.279	0.235	0.364	0.356
InQuery	0.391	0.426	0.324	0.468	0.407
Indri	0.595	0.534	0.533	0.439	0.544
CQA-1.0	**0.765**	**0.637**	**0.722**	**0.701**	**0.718**
			P10		
PubMed	0.350	0.150	0.200	0.320	0.281
InQuery	0.692	**0.783**	**0.567**	**0.720**	**0.704**
Indri	0.575	0.500	0.367	0.400	0.500
CQA-1.0	**0.783**	0.583	0.467	0.660	0.677
			MRR		
PubMed	0.579	0.443	0.456	0.540	0.526
InQuery	0.854	**1.000**	**1.000**	**1.000**	0.933
Indri	0.750	0.728	0.833	0.380	0.683
CQA-1.0	**0.917**	0.889	**1.000**	**1.000**	**0.936**

Table 5.10: **Difference in CQA-1.0 re-ranking versus InQuery and Indri for FPIN-train (train) and FPIN-eval-1 (test) collections. All but differences shown in italics are statistically significant ($p < 0.05$.)**

	MAP		P10		MRR	
	train	test	train	test	train	test
CQA-1.0 vs. InQuery	+158.2%	+76.4%	+33.1%	*-3.8%*	+19.1%	*+0.32%*
CQA-1.0 vs. Indri	+19.7%	+32.1%	+23.6%	+35.4%	*+3.8%*	+37.0%

Since the goal was to compare CQA-1.0 re-ranking quality against that of state-of-the art search engines, rather than state-of-the art retrieval against that of PubMed, InQuery and Indri were used for re-ranking in these experiments. First a 10-year subset of MEDLINE that contains all documents retrieved for the experiments using PubMed was indexed, then 10,000 documents were retrieved using the original questions as queries, the results were intersected with PubMed retrieval results, thus achieving InQuery re-ranking of the original set. Indri results were obtained in the same manner, with the exception that PubMed hits not in top 10k Indri hits were added to the end of the results list sorted in reverse chronological order. Table 5.9 presents the results of these experiments.

Table 5.11: **Results of re-ranking experiments for FPIN-train questions using alternative judgments (FPIN-eval-2)**

	MAP		Bpref		P10		MRR	
	CS	KWF	CS	KWF	CS	KWF	CS	KWF
PubMed	0.365	0.336	0.281	0.229	0.258	0.191	0.611	0.577
InQuery	0.142	0.177	0.237	0.318	0.121	0.112	0.346	0.343
Indri	0.335	0.328	0.532	0.487	0.225	0.179	0.440	0.417
CQA-1.0	**0.499**	**0.462**	**0.549**	**0.506**	**0.292**	**0.254**	**0.722**	**0.654**

Statistical significance of the results was determined using the Wilcoxon signed

ranks test. Indri and CQA-1.0 significantly outperform the PubMed baseline on all metrics. InQuery outperforms the baseline on MRR and precision at 10, but not on MAP.

The main question however is whether CQA-1.0 achieves the state-of-the-art re-ranking performance? Differences in CQA-1.0 re-ranking versus InQuery and Indri in percent of change is shown in Table 5.10. In three cases (shown in italics in the table) the differences between the CQA-1.0 re-ranking and that of InQuery or Indri is not statistically significant. These are precision at 10 and MRR vs. InQuery on the test set, and MRR vs. Indri on the development set. On all other metrics for both sets, the CQA-1.0 re-ranking algorithm significantly outperforms Indri and InQuery.

As the relevance judgments for FPIN-train and FPIN-eval-1 were prepared by the author (prior to CQA-1.0 processing), the FPIN-eval-2 relevance judgments were used to control for potential bias in the *FPIN* evaluation. Table 5.11 presents the results of the FPIN-eval-2 evaluation. Because the judgments are incomplete (at most 10 documents per topic), the Bpref results are also used in this evaluation. CQA-1.0 significantly outperforms InQuery on all metrics ($p < 0.05$) according to both doctors. In addition, CQA-1.0 significantly outperforms Indri ($p < 0.05$) on MRR, and MAP-CS; and PubMed ($p < 0.01$) on Bpref. These results suggest that the evaluation based on FPIN-train and FPIN-eval-1 fairly accurately reflects the expected performance of the system.

5.3.3 Domain model components contribution to re-ranking

Table 5.12: Domain model components contribution to re-ranking of the *FPIN* collection compared to linear combination of all components (Percent of relative difference.)

	MAP	vs. EBM	**P10**	vs. EBM	**MRR**	vs. EBM
FPIN-train						
S_{EBM}	0.754		0.699		0.910	
S_{PICO}	0.709	−6.0%	0.657	−6.0%	0.903	−0.8%
S_{SoE}	0.512	−32.2%	0.482	−31.0%	0.674	−25.9%
S_{task}	0.512	−32.2%	0.457	−34.6%	0.714	−21.6%
$S_{SoE} + S_{task}$	0.556	−26.4%	0.528	−24.5%	0.781	−14.2%
FPIN-eval-1						
S_{EBM}	0.718		0.677		0.936	
S_{PICO}	0.646	−10.0%	0.627	−7.4%	0.847	−9.5%
S_{SoE}	0.457	−36.4%	0.427	−36.9%	0.644	−31.1%
S_{task}	0.504	−29.8%	0.435	−35.8%	0.663	−29.2%
$S_{SoE} + S_{task}$	0.538	−25.1%	0.485	−28.4%	0.677	−27.6%

Table 5.13: Domain model components contribution to re-ranking of the CE-eval-1 collection

	MAP	R-prec	MRR	P@5	P@10
PubMed	0.3965	0.3977	0.6462	0.5120	0.3000
S_{EBM}	0.7129	0.6477	0.9213	0.8080	0.4400
S_{PICO}	0.4912	0.4515	0.7161	0.4960	0.3640
S_{SoE}	0.4503	0.3971	0.6180	0.4480	0.3200
S_{task}	0.1926	0.1293	0.2821	0.1040	0.1240
$S_{PICO} + S_{SoE}$	0.6692	0.5907	0.8893	0.7360	0.4320
$S_{PICO} + S_{task}$	0.5020	0.4515	0.7722	0.4800	0.3240
$S_{SoE} + S_{task}$	0.4776	0.4318	0.6147	0.4800	0.3240

The remaining question, which of the three basic components contributes most to re-ranking performance, was answered in a series of ablation experiments. The EBM score of a MEDLINE citation is computed in CQA-1.0 as a sum of three equally weighted components of the semantic domain model. To establish the impact

Table 5.14: **Domain model components contribution (percent of relative difference in MAP) to re-ranking of the *CE-eval-1* collection compared to linear combination of all components and PubMed.**

Component	S_{PICO}	S_{SoE}	S_{task}	$S_{PICO} + S_{SoE}$	$S_{PICO} + S_{task}$	$S_{SoE} + S_{task}$
vs. PubMed	23.9%	13.6%	−51.4%	68.8%	26.6%	20.5%
vs. CQA-1.0	−31.1%	−36.8%	−72.9%	−6.1%	−29.6%	−33.0%

of individual components, computation of the final score (See Equation 4.5) was modified as follows:

$$S_{\text{EBM}} = \lambda_1 S_{\text{PICO}} + \lambda_2 S_{\text{SoE}} + (1 - \lambda_1 - \lambda_2) S_{\text{task}} \tag{5.1}$$

Ablation experiments were conducted setting coefficients for each component to zero without any other changes to the system. These experiments were conducted on the *CE-eval-1* and *FPIN* collections, the latter by Jimmy Lin. Table 5.12 and Table 5.13 present the results of these experiments.

Since the trends in performance are similar for all metrics, differences in performance as compared to PubMed baseline and the complete system are illustrated using MAP in Table 5.14. With the exception of re-ranking by task on the *CE-eval-1* collection, all individual components outperform the baseline at least at the 10% confidence level. The reason for the poor task-scoring performance for *CE-eval-1* probably lies with the PubMed retrieval strategy being already strongly geared towards therapy by restricting search to documents indexed with subheading *drug therapy* and publication type *clinical trial*. However, this weak predictor boosts performance of the other two re-rankers. With the exception of $S_{PICO} + S_{SoE}$ re-ranking for *CE-eval-1*, combining all three components significantly improves the

ranking when compared to individual components or the other pairwise combinations. Given the results of these two sets of experiments, it is reasonable to conclude that a template for formalizing a problem structure common to the majority of the information needs in the domain, and identification of this structure in documents, is the most effective component of semantic re-ranking. The second component that improves the overall semantic re-ranking performance is the strength of evidence. Finally, the task reranker, which performed poorly on its own on the *CE-eval-1* collection, boosts the performance of the other two re-rankers.

Besides playing an important role in re-ranking, the PICO structure is also crucial in answer generation, evaluation of which will be discussed next.

5.4 Answer generation evaluation

Three evaluations of answer generation were conducted:

- manual evaluation of the best answers
- interactive clinical scenario emulation given a multi-tiered answer
- exploration of automatic evaluation

5.4.1 Best answer evaluation

This evaluation focuses only on the evaluation of topical relevance of the answers generated by the system, which deviates from the recommendations described in Section 2.2.3 in that it does not involve verification of reliability of the answer or applicability in a specific clinical setting, but rather concentrates on the answer itself.

Evaluation design

Two assessors, a family practitioner (CS) and a surgeon (KWF), were asked to evaluate whether the texts they were given answer the questions or contain a potentially actionable advice. They were not asked whether they would act on the given answer in the clinical situation described in the question. To enforce this restriction, information that could influence the judgments, such as the strength of evidence, and names of the authors and of the journal was withheld.

Doctors CS and KWF were asked to evaluate answers on a three-point scale:

direct answer A plus (+) indicated that the response directly answers the question and the article from which it was extracted will be examined.

probable answer A check ($\surd$) indicates that the response provides clinically relevant information that may factor into decisions about patient treatment, and that the citation and the article were worth examining in more detail.

not an answer A minus (-) indicates that the response does not provide useful information in answering the clinical question, and that the citation was not worth examining in detail.

This experiment was conducted on the twenty four questions of the FPIN-eval-2 collection. Answers were automatically extracted from the top five citations in the re-ranked list. As a contrastive condition, answers were automatically extracted from the first five citations in the original PubMed result set. Each question, followed by the blinded answers, was presented to each assessor in a paper printout

that contained answers in a randomized order (that was consistent across the two assessors) with duplicates removed. On average, the length of answers generated from the original PubMed list of citations was ninety words; answers generated from the re-ranked list of citations averaged eighty seven words. Answers from both sources achieve approximately 30% compression of the original text: for original PubMed results the average length of an abstract is 250 words and for re-ranked results it's 270 words.

Answers were evaluated using precision, calculated as Instance Precision(IP), defined in the TREC QA track as follows (Voorhees 2004): "Let D be the number of correct, distinct responses returned by the system, and N be the total number of responses returned by the system. Then $IP = D/N$." Because $N = 5$ in this experiment, answer precision is equivalent to P@5. Instance recall, the second metric in the TREC QA list questions evaluation, is computed as the ratio of the number of correct responses returned by the system (D) to the number of known instances (S.) Although the goal of this experiment was to evaluate the quality of the generated best answers through precision (the completeness is hypothesized to be achieved in multi-tiered answers), recall was also computed. Answer precision and recall was calculated for two conditions: under the strict condition, only "plus" judgments were considered good (direct answer); under the lenient condition, both "plus" and "check" judgments were considered good (probable+ answer).

Examples of what the doctors considered to be direct and probable answers follow.

What is the best treatment for analgesic rebound headaches?

Plus judgment (direct answer):

> Medication overuse headache from anti-migraine therapy: clinical features, pathogenesis and management: Because of easy availability and low expense, the greatest problem appears to be associated with barbiturate-containing combination analgesics and over-the-counter caffeine-containing combination analgesics. The best management advice is to raise awareness and strive for prevention. Reduction in headache risk factors should include behavioral modification approaches to headache control earlier in the natural history of migraine.

This answer was accepted by both physicians because it identifies specific analgesics that are most likely to cause the problem, and suggests preventive treatment.

Check judgment (probable answer):

> Does chronic daily headache arise de novo in association with regular use of analgesics? Regular use of analgesics preceded the onset of daily headache in 5 patients by a mean of 5.4 years (range, 2 to 10 years). In 1 patient, the onset of daily headache preceded regular use of analgesics by almost 30 years. These findings suggest that individuals with primary headache, specifically migraine, are predisposed to developing chronic daily headache in association with regular use of analgesics.

Although this passage provides some information about analgesic rebound headaches, it does not discuss treatment options. For these reasons this abstract was marked as potentially leading to an answer, but not as containing one.

Tables 5.15 and 5.16 present the evaluation results for the direct and probable+ answers respectively. The differences between answer precision achieved by CQA-1.0 and the PubMed baseline on all answers is statistically significant (according

Table 5.15: **Direct answer precision and recall for PubMed (Base) and CQA-1.0 (CQA).**

	Family physician				Surgeon			
	Precision		Recall		Precision		Recall	
	Base	CQA	Base	CQA	Base	CQA	Base	CQA
Therapy	0.16	**0.26**	0.40	**0.65**	0.04	**0.20**	0.16	**0.83**
Diagnosis	0.23	**0.37**	0.44	**0.69**	0.23	**0.30**	0.47	**0.60**
Prognosis	**0.33**	**0.33**	**0.56**	**0.56**	0.20	**0.27**	0.50	**0.67**
Etiology	0.48	**0.60**	0.54	**0.68**	0.40	**0.56**	0.50	**0.70**
All	0.27	**0.37**	0.48	(**0.66**)	0.18	**0.31**	0.42	**0.70**

Table 5.16: **Probable+ answer precision for PubMed (Base) and CQA-1.0 (CQA).**

	Family physician				Surgeon			
	Precision		Recall		Precision		Recall	
	Base	CQA	Base	CQA	Base	CQA	Base	CQA
Therapy	0.40	**0.64**	0.39	**0.63**	0.24	**0.52**	0.32	**0.70**
Diagnosis	0.30	**0.57**	0.38	**0.71**	0.27	**0.60**	0.32	**0.72**
Prognosis	**0.53**	0.40	**0.62**	0.46	0.33	**0.40**	0.50	**0.60**
Etiology	0.52	**0.64**	0.54	**0.67**	0.44	**0.56**	0.52	**0.67**
All	0.42	**0.59**	0.45	**0.63**	0.30	**0.53**	0.39	**0.69**

to the Kolmogorov-Smirnov two sample test) for the family physician at the 95% significance level, and for the surgeon at the 99% significance level. According to both assessors, CQA-1.0 significantly outperforms the PubMed baseline (which already utilizes PICO extraction technology) in the "best answer" generation.

5.4.2 Interactive clinical scenario emulation

The multi-tiered answers presented in Section 3.7 are meant to be interactively explored by clinicians, who might follow different interaction paths depending on their background knowledge, familiarity with the system, etc. In this section, a complete analysis of an answer to one question is used to illustrate how the "decision points" in the interactive clinical scenario emulation led to an evaluation design of the CQA-1.0 multi-tiered answer generation described in Section 4.6.1. The evaluation design, results, and analysis are then presented.

Clinical scenario example

> **Question:**
> A woman presents at nine weeks of pregnancy with a large visible uterine/abdominal enlargement. Her beta-human chorionic gonadotropin (BHCG) is reported to be greater than 200,000 IU/L. Does she need an ultrasound to rule out a molar pregnancy, twins or other possibilities?

Given this scenario, the doctor is interested in both the differential diagnosis and diagnostic methods (what could cause patient's symptoms, and is ultrasound indicated to find the cause.) For the diagnostic task question, the PICO frame contains elevated HCG in the Population slot and ultrasound in the Intervention slot. Forty five citations are retrieved using these terms. Clustering on problems should

list disorders that could cause elevated HCG, and clustering on interventions should provide an answer to the question whether ultrasound is indicated and sensitive to the above disorders. Clustering on problems provides two broad categories as the top-tier answer (*Congenital Disorders* and *Maternal complications of pregnancy.*) Specific diseases belonging to these categories, and outcome statements provide context for the two key points(see Figure 5.1).

Congenital Disorders (23)

Down syndrome (trisomy 21) 14698332; 12204419; 9933818; 9313073; 11920896; ...

> Fetal nuchal translucency thickness (NT) at the 11-14 week scan has been combined with maternal age to provide an effective method of screening for trisomy 21; for an invasive testing rate of 5%, about 75% of trisomic pregnancies can be identified.

Trisomy 14755412; 15758613; 15229846; 11981911

> Increased fetal nuchal translucency at 11-14 weeks can also identify a high proportion of other chromosomal abnormalities and is associated with major defects of the heart and great arteries, and a wide range of skeletal dysplasias and genetic syndromes.

Triploidy 10861716; 10739543; 15229844

> Two patterns of first-trimester markers emerged: type I, characterized by extremely high levels of free beta-hCG and elevated NT; and type II, characterized by very low levels of PAPP-A and free beta-hCG with normal NT

Edwards syndrome ...

intrauterine growth restriction ...

Maternal complications of pregnancy (10) ...

Figure 5.1: Semantic clustering on problems.

Clustering on interventions might seem superfluous, because the query contained the intervention of interest; however, because CQA-1.0 clusters on the main intervention, several interesting clusters were formed (see Figure 5.2). These answers

General Medical Procedures and Services Including Histories and Physical Examinations (19)

Screening 14698332; 12204419; 15668845; ...

> Prospective studies have demonstrated that screening by a combination of fetal nuchal translucency (NT) and maternal serum free-beta-human chorionic gonadotropin (hCG) and pregnancy-associated plasma protein-A (PAPP-A) can identify 90% of fetuses with trisomy 21 and other major chromosomal abnormalities for a false-positive rate of 5%. In this, the patients are subdivided into a high-risk group, requiring invasive testing, a low-risk group, which can be reassured that an abnormality is unlikely, and an intermediate-risk group (risk of 1 in 101 to 1 in 1000), in which further assessment is performed by first-trimester ultrasound examination (for presence/absence of the nasal bone or presence/absence of tricuspid regurgitation or normal/abnormal Doppler velocity waveform in the ductus venosus), and chorionic villus sampling is performed if their adjusted risk becomes 1 in 100 or more.

Amniocentesis 15229846; 14755412

> Fetal blood sampling is of a limited value in confirming mosaic trisomy 16 ascertained through amniocentesis.

pelvic examination ...

Imaging/visualization/scanning (11)

transvaginal ultrasound 9159456; 10577398; ...

Hematology investigations (incl blood groups)

...

Figure 5.2: Semantic clustering on interventions.

demonstrate redundancy elimination achieved through clustering. Rather than reading 16 titles of articles devoted to screening interleaved with other interventions, a clinician may view the context of the best representative of the cluster to find out what constitutes screening in this context.

> Although the absolute value of the BHCG is undiagnostic, a level of 200,000 IU/L is indeed elevated for nine weeks of pregnancy. This finding, along with the enlarged abdomen, suggests several possible diagnoses, including: wrong dates, multiple pregnancy, normal pregnancy with pelvic pathology (fibroids or ovarian mass), and molar pregnancy. All of these can be ruled in or out with an ultrasound. As such, at any point early in a pregnancy when you're not quite sure exactly what's going on, an ultrasound is the most reliable diagnostic tool we have, and we should never be afraid to use it.

Figure 5.3: Reference answer in the Parkhurst Database.

Multi-tiered answer evaluation design overview

The above example of the interactive answer examination highlights two "decision points" that depend on the quality of two tiers of the multi-tiered answer: (1) the informativeness of the top-tier answers (cluster labels), which determines the cluster inspection order; and (2) the quality of the evidence support in the cluster, which determines whether the top-tier answers are useful. Another factor that determines the usefulness of the top-tier answer is the quality of the formed clusters. To evaluate the usefulness of the clusters independently of a user's ability to find the right cluster, we previously developed an "oracle"-based cluster selection method (Demner-Fushman, He, & Oard 2004). This method was modified to evaluate the semantic clusters as described below.

The reference answer given to the example question in the Parkhurst database (see Figure 5.3) demonstrates that there is a significant overlap between the reference and the system response in Figure 5.1, but it also illustrates that an automatic evaluation metric based on direct surface representation overlap, such as ROUGE (Lin & Hovy 2003), a metric most widely used in summarization, might have only limited success. Therefore manual evaluations of the system were conducted, and the

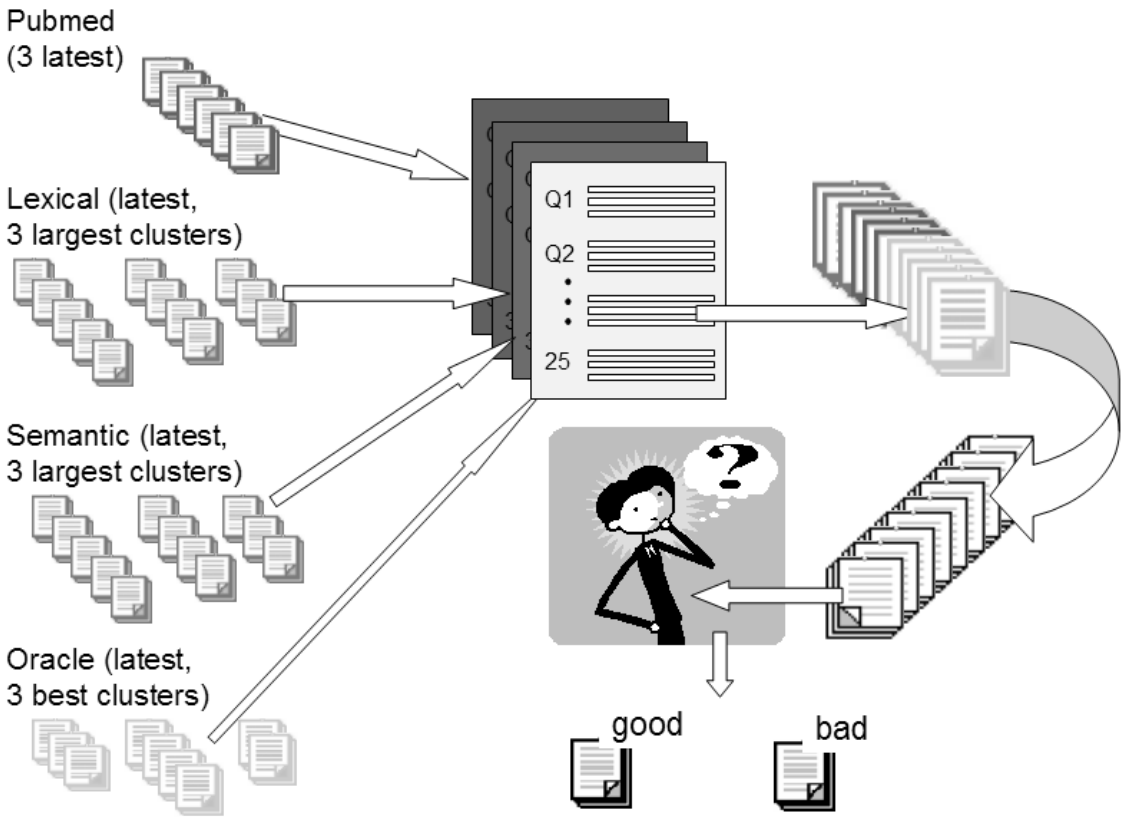

Figure 5.4: Cluster selection for evaluation

automatic evaluation using ROUGE was further explored (see Section 5.4.3.)

The evaluation design was motivated by two facts:

- Doctors are willing to spend 2-4 minutes searching for an answer (Alper *et al.* 2001; Ely *et al.* 2005).
- Titles of up to three articles can be displayed on PDAs that are increasingly used to access information at the point of care (Hauser *et al.* 2004).

Therefore the number of abstracts to evaluate for each cluster inspection order was set to three. The answer inspection order and the usefulness of the semantic clusters was evaluated using precision (the proportion of examined abstracts that are relevant.) Three evaluations of multi-tiered answers were conducted.

In the first evaluation (see Figure5.4), conducted by the author, the usefulness of the semantic clusters is evaluated on twenty two questions pertaining to therapy

and twelve questions pertaining to diagnosis. Twenty therapy and ten diagnosis questions were set aside for testing (the FPIN-eval-3 collection), and the remaining four questions were used for system tuning. As all abstracts in this collection were evaluated in advance of system development, no additional examination of the abstracts was required. The second and third evaluations are based on CE-eval-1, CE-eval-2, and CE-eval-3 collection. These experiments evaluate the two aspects of the answer quality (top level answer informativeness and the quality of answer support) under several potential cluster examination orders.

Cluster usefulness evaluation design (FPIN-eval-3 collection)

The first three articles in the original PubMed presentation order served as a baseline in this experiment. In addition to PubMed baseline, four cluster selection orders were evaluated: one for a contrastive lexical clustering baseline; one for a semantic clustering baseline; and two for clusters selected by an "oracle". Lexical clustering tool based on Ward's algorithm implemented by Anton Leuski (2001), was used to cluster MEDLINE citations purely on their term content (by representing each document as a high dimensional vector in which each unique term served as a feature.) This is the *Lexical* condition. The examination order for the *Lexical* condition presents the first abstract for each of the top three intervention clusters, sorted in descending order by cluster size. That is, clusters with larger numbers of contributing abstracts are evaluated first.

The semantic baseline *Semantic* was established using the cluster size exam-

ination order. As for the *Lexical* examination order, semantic clusters were sorted by their size for examination purposes. Because the labels of semantic clusters are meaningful, a trained professional does not have to rely on the system for cluster selection, but can examine the clusters in the order of expected relevance. Such behavior can be simulated using an "oracle", which selects the three most relevant clusters for the clinician. In this experiment, the *FPIN* reference standard (the original answers provided by experts) served as an oracle. For example, for the question *What are the best medications for panic disorder?* the answer *Drug groups primarily affecting the central nervous system* is examined first, because both antidepressants and benzodiazepines, recommended in the original *FPIN* answer, fall into this broad category in the UMLS. This is the *Oracle-1* evaluation order.

Instead of sampling three abstracts from different clusters, three abstracts from the most promising cluster selected by an "oracle" could be examined. This simulates the scenario when a clinician is interested in only one intervention and wants to learn about it in depth, which would correspond to examining only the *imaging/visualization/scanning* cluster in Figure 5.2, as *ultrasound* is an imaging procedure. This is the *Oracle-2* evaluation order.

Because of the relatively small number of abstracts in the examination order evaluation for each question, the original four-point judgments made on the abstracts were used as binary: marginally relevant citations were considered to be non-relevant, and topically relevant and relevant citations were considered to be relevant.

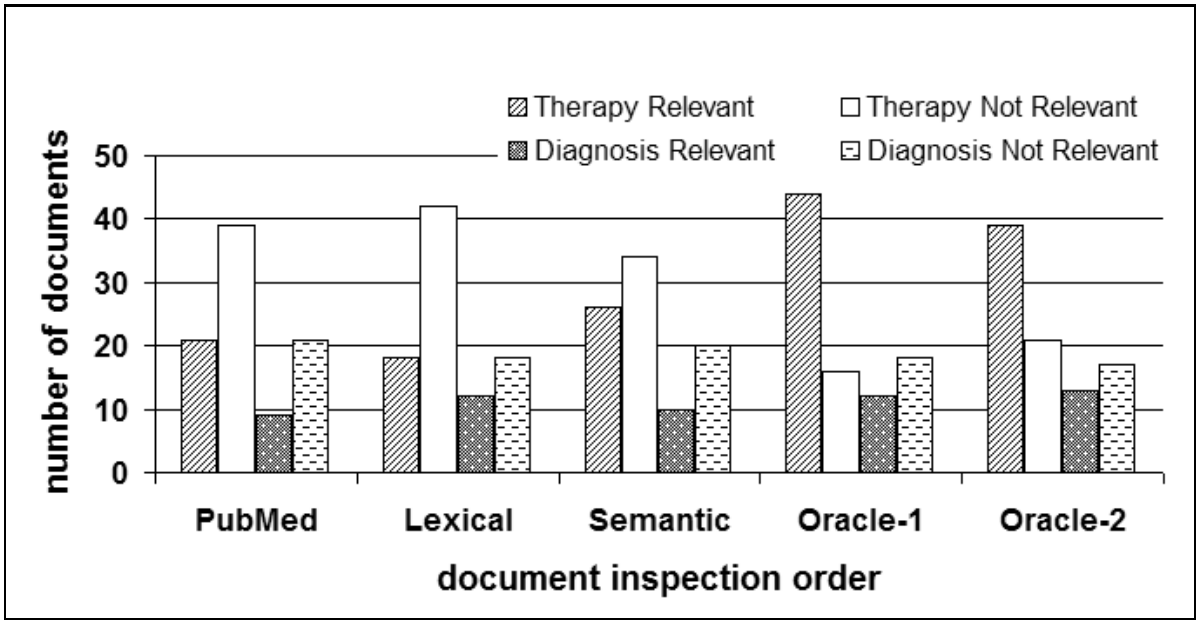

Figure 5.5: Results of answer clustering for the *FPIN* collection.

FPIN-eval-3 collection evaluation results

Table 5.17: **Cluster usefulness evaluation (in precision) for FPIN-eval-3 collection.**

Condition	Rank 1	Rank 2	Rank 3	Total
PubMed	0.33	0.36	0.30	0.33
Lexical	0.30	0.36	0.36	0.33
Semantic	0.40	0.40	0.36	0.38
Oracle-1	0.63	0.70	0.53	0.62
Oracle-2	0.63	0.53	0.56	0.58

The results of the FPIN-eval-3 evaluation are shown in Table 5.17. Despite employing advanced techniques and focusing clinical questions for PubMed queries, only a third of the top three retrieved citations were relevant, as demonstrated by the baseline condition in Table 5.17. Of those abstracts, only 16% were evaluated as containing an answer in the original four-point judgments. Lexical clustering,

or grouping citations based solely on their "bag-of-words" representation, did not improve significantly upon the PubMed baseline. The differences in the results between the *Semantic* condition and two other baselines (PubMed original order and lexical clustering) are not statistically significant. For the *Oracle-1* and *Oracle-2* conditions, the clusters contained more relevant abstracts. Figure 5.5 presents the number of relevant and non-relevant documents suggested for examination over all questions under each experimental condition. Overall, precision of semantic clustering on therapy questions was better than on diagnosis questions. The proportion of relevant documents suggested for examination under the *Oracle-1* and *Oracle-2* conditions is significantly higher ($p < 0.05$) than in the sets suggested by the first two baselines. The poor performance on the diagnosis questions might be explained by the mixture of questions about diagnostic tests and differential diagnosis in this collection. Clustering on disorders shown in Figure 5.1 should yield better results for differential diagnosis questions. This hypothesis can not be tested quantitatively given the existing collections. A special effort would be needed to create a collection of etiology and differential diagnosis questions of a size sufficient to capture statistically significant differences.

In this evaluation, consistent improvements in CQA-1.0 performance over the baselines can be attributed to clustering, since to avoid confounding the influence of clustering and the EBM-based re-ranking, abstracts were listed in the original PubMed order within each cluster. An additional advantage of clustering is in providing an overview of the available information, which is known to speed up finding an answer (Hearst & Pedersen 1996; Dumais, Cutrell, & Chen 2001).

The question remains, whether the semantic oracle condition approximates reality? If the answer is not known in advance, given a small number of choices, would clinicians rank broad intervention categories suggested by the system appropriately, based on their background knowledge? Selection of examination order based on clinician's background knowledge and intuition (instead of an "oracle") is evaluated in CE-eval-1,2,3 experiments.

CE collection evaluation design

Cluster labels derived in semantic clustering on interventions using documents retrieved to answer 25 questions in the CE-eval-1 collection form the top-tier answers. These labels represent classes of interventions, for example, *Antibiotics*. The first step in CE-eval-1 evaluation is the answer inspection order selection. To avoid bias in evaluation and establish a baseline, three different sets of answers were generated. The presentation orders for the CE-eval-1 are: (1) the first three abstracts retrieved by PubMed (the baseline condition); (2) the first abstract from each of the three largest clusters (same as the *Semantic* baseline condition in the FPIN-eval-3 evaluation); and (3) informed cluster selection(similar to the *Oracle-1* condition in the FPIN-eval-3 evaluation.) This is the *User* condition. These experimental conditions led to selection of two overlapping sets of documents: CE-eval-2 and CE-eval-3. CE-eval-2 was generated and further inspected for evidence support by the author. The identical evaluation conducted by Dr. CA resulted in the CE-eval-3 collection.

Due to the nature of the *CE database*, which contains reference judgments in

the form of interventions, the informativeness of the top-tier answers was evaluated based on the interventions extracted from the abstracts, rather than on the whole abstracts. For the PubMed baseline, the main intervention from the first three PubMed citations was used. For the *Semantic* and the *User* condition, the main intervention of the abstracts chosen as described above was added to the pool. This yields interventions that are at the same level of ontological granularity as those extracted from the unclustered PubMed abstracts. This preparation yielded up to nine different intervention names, three from each experimental condition. The interventions in the PubMed baseline and the *Semantic* condition are identical in the CE-eval-2 and the CE-eval-3 collection. The difference is in the interventions from the *User* condition. After blinding the source of these interventions and removing duplicates, each short answer was printed in a random order and evaluated by the author for the CE-eval-2 collection, and by Dr. CA for the CE-eval-3 collection.

Each answer was first evaluated with respect to the CE reference set. Each intervention found by the system was assigned to one of the six categories (beneficial, likely beneficial, tradeoffs, unknown, unlikely beneficial, harmful), based on CE recommendations. Because these recommendations are incomplete, an additional category *not in CE* was introduced. Matching of the system output to the reference standard needs to be done manually and requires domain knowledge to deal with synonymy and differences in ontological granularity. Only drug names were used in this categorization. In the next step of simulated interaction, the quality of the context for answer support was judged on a three-point scale as *good*, *okay* (marginal), or *bad* based on CE where available, and taking into account the title,

the top-scoring outcome sentence, and, if necessary, the entire abstract text.

Table 5.18: **Manual evaluation of short answers for CE-eval-2: distribution of system answers with respect to CE categories (Key: B=beneficial, LB=likely beneficial, T=tradeoffs, U=unknown, UB=unlikely beneficial, H=harmful, N=not in CE)**

	B	LB	T	U	UB	H	N
PubMed	0.200	**0.213**	0.160	0.053	-	**0.013**	**0.360**
Semantic	0.387	0.173	**0.173**	0.027	-	-	0.240
User	**0.400**	0.200	0.133	**0.093**	**0.013**	-	0.160

Table 5.19: **Manual evaluation of short answers for CE-eval-3: distribution of system answers with respect to CE categories (Key: B=beneficial, LB=likely beneficial, T=tradeoffs, U=unknown, UB=unlikely beneficial, H=harmful, N=not in CE)**

	B	LB	T	U	UB	H	N
PubMed	0.133	0.253	0.133	0.013	-	**0.013**	**0.453**
Semantic	0.227	0.267	**0.120**	0.040	-	**0.013**	0.333
User	**0.347**	**0.293**	0.107	**0.067**	-	-	0.187

Table 5.20: **Manual evaluation of contextual support.**

	CE-eval-2			CE-eval-3		
	Good	Okay	Bad	Good	Okay	Bad
PubMed	0.600	**0.227**	**0.173**	0.573	**0.146**	**0.266**
Semantic	0.827	0.133	0.040	0.720	0.093	0.187
User	**0.893**	0.093	0.013	**0.853**	0.080	0.067

CE collection evaluation results

Results of the manual evaluation of the cluster label informativeness are presented in Tables 5.18 and 5.19, which show the distribution of judgments for the three experimental conditions. For the PubMed, 13% of the examined drugs fell into

the beneficial category according to one assessor, and 20% according to the other; the values are 23%(39%) for the *Semantic* condition and 35%(40%) for the *User* condition. Dr. CA judged more interventions to be "not in CE". The differences in assignment of the system answers to *CE* categories were anticipated, as it calls for judgments. Because of that, the absolute values are less important than the distribution patterns for each selection order, and the relative performance of the selection orders with respect to each other.

In terms of top-tiered answers, semantic clustering returns approximately twice as many beneficial drugs as the baseline (for both cluster inspection orders and both assessor.) According to the second tier judgments (judgments on the context), 57%(60%) of PubMed short answers were found to be *good*, compared to 72%(83%) and 85%(89%) for the *Semantic* and *User* conditions, respectively. From a factoid QA point of view, CQA-1.0 clearly outperforms the PubMed baseline.

5.4.3 Exploration of automatic evaluation of abstracts

This evaluation explores a possibility to replace human judgments on MEDLINE abstracts with an automatic estimate whether an abstract answers the information need expressed in the question (at least for system development and tuning.) Stated this way, this research question resembles the Document Understanding Conferences (DUC) system task (Dang 2005), in which 250-word long automatically generated summaries are evaluated both manually and automatically. The latter, using reference summaries generated manually by NIST assessors and a software

package called ROUGE (Recall-Oriented Understudy for Gisting Evaluation (Lin & Hovy 2003).) The automatic evaluation of NLP applications using ROUGE is based on co-occurrences of words and multi-word units in the automatic- and the reference summaries.

The *CE* collection provides an opportunity to explore this automatic evaluation metric and compare the results with human judgments. Leveraging this resource, the cluster usefulness was evaluated automatically, replacing relevance judgments with ROUGE scores. Under an assumption that the references cited in *CE* are high-quality relevant abstracts (since they were used in generating the drug recommendations), those were retrieved and used as the reference summaries. Abstracts found by the CQA-1.0 system in support of the extracted interventions were treated as summaries "generated" by the system and evaluated using the standard DUC ROUGE settings. ROUGE computes a number of scores, of which ROUGE-1 precision (found to correlate better than other ROUGE scores with human judgments (van Kesteren & Kraaij 2006)) was used in this evaluation.

The cluster inspection orders and document selection conditions established for the CE-eval-1 experiments (see Section 5.4.2) were used in this evaluation: the *User-1* condition evaluates the top abstract from the three promising clusters selected by Dr. CA; the *User-2* condition evaluates the top abstract from the three promising clusters selected by the author; and the *User-3* condition evaluates three abstracts from the most promising cluster selected by the author.

The precision was approximated using "cumulative relevance" (gain the user would receive by examining the selected abstracts up to a given rank (Järvelin &

Kekäläinen 2002).) This metric accumulates the ROUGE-1 precision score of the first, second, and third abstract that were examined in manual evaluations.

Table 5.21: **Cumulative relevance after examining the first, second, and third abstracts.** (° denotes significance at the 90% level, ▽ denotes significance at the 95% level.)

	Rank 1		Rank 2		Rank 3	
PubMed	0.170		0.349		0.523	
Semantic	0.181	(+6.3%)	0.356	(+2.1%)	0.526	(+0.5%)
User-1	0.200	(+17.7%)	0.393	(+12.5%)°	0.588	(+12.3%)▽
User-2	0.206	(+21.5%) °	0.396	(+13.6%)°	0.586	(+11.9%)▽
User-3	0.206	(+21.5%) °	0.392	(+12.6%)°	0.597	(+14.0%)▽

Table 5.22: **Overlap between the reference abstracts and the system answers after examining the first, second, and third abstracts (Percent at each rank).**

	Rank 1	Rank 2	Rank 3
PubMed	1.1%	1.1%	
Semantic	2.2%	2.2%	2.2%
User-1	10%	6.7%	5.6%
User-2	11.1%	8.9%	6.9%
User-3	11.1%	6.4%	8.9%

Table 5.21 shows the statistically significant increase in cumulative relevance for the *User* conditions after examining the first (except for *User-1*), second, and third abstract. Although ROUGE-based evaluation agrees with manual evaluation on the overall usefulness of the clusters, judgments on single answers agree with manual judgments only moderately. For example, an abstract containing the following answer:

> In patients with both GER and asthma, antireflux surgery (but not medical therapy with ranitidine 150 mg t.i.d.) has minimal effect on pulmonary function, pulmonary medication requirements, or survival, but significantly improves asthma symptoms and overall clinical status.

for the question about the most effective interventions for asthma, has the highest ROUGE score, but is marginally relevant, because it is highly unlikely a physician will consider a surgical procedure to improve asthma symptoms. Because this answer contains all the "right" terms and semantic types, to correctly evaluate this instance, an automatic metric would need domain knowledge beyond the given context. The overall agreement of the automatic and manual evaluation might be attributed to direct matches (see Table 5.22) on the reference abstracts found by the system. This assumption and its implications for the question answering and summarization evaluation call for a thorough investigation in the future.

Summary

This chapter evaluated the CQA-1.0 system in the context of two clinical question answering goals: finding the best answer for quick reference, and providing a full multi-tiered answer.

Two assessors (Dr. CS and Dr. KWF) evaluated short lists of best answers generated by the CQA-1.0 system and the PubMed baseline representative of the best tools available to clinicians, who seek answers to questions in real-life clinical situations. In this evaluation (presented in Section 5.4.1), CQA-1.0 significantly outperformed the baseline, providing at least one probable+ answer to 23 of 24 questions in the FPIN-eval-2 collection according to Dr. CS, and to all questions according to Dr. KWF.

Section 5.4.2 presents the results of the evaluation of multi-tiered answers.

The multi-tiered answers were evaluated in two independent two-stage emulations of an interactive clinical scenario. In the first stage, the assessor (Dr. CA or the author) selected top-tier answers generated by the CQA-1.0 system through semantic clustering of documents retrieved to answer 25 questions in the CE-eval-1 collection. Interventions from the top abstract in three clusters selected by the assessor for each disease were randomly combined with interventions generated in two baseline conditions (first three abstracts in PubMed retrieval results retrieved to answer the questions, and the top abstract in three largest clusters.) The assessors then used the reference standard to categorize each intervention on a six-point scale ranging from beneficial to harmful. The top-tier answers (cluster labels) helped both assessors find twice as many beneficial drugs as the baseline.

In the second stage of this evaluation, the assessors evaluated the quality of the contextual support for the top-tier answers. According to both assessors, the support provided by CQA-1.0 was higher in quality by a large margin.

In addition to the summative evaluation of the system, this chapter provides evaluations of the system components: Section 5.2 presents the respectable performance of the population, problems, interventions, and outcome extractors. Experiments presented in Section 5.3.2 confirm that utilizing semantic domain model to identify elements necessary for answer generation provides a re-ranking mechanism that outperforms or is on par with two of the best statistical methods. Finally, section 5.3.3 evaluates contribution of each component of the semantic domain model to document re-ranking.

Chapter 6

Conclusions

This study uses a semantic domain model of clinical medicine to encode (a) a clinician's information need expressed as a question on the one hand and (b) the meaning of scientific publications on the other to yield a common representation. It was hypothesized that this approach works well for (1) finding documents that contain answers to clinical questions and (2) extracting these answers from the documents. The "proof by construction" approach was taken to test this hypothesis.

6.1 Recapitulation

A clinical question answering system was implemented and evaluated to corroborate the above hypothesis in the medical domain. The medical informatics domain provides unparalleled resources that permitted implementation of a working prototype of a clinical question answering system. Having this prototype is important to the domain of medical informatics itself, since one of the major goals of research in this domain is how to provide support for informed clinical decision making.

Several steps needed to be undertaken prior to system implementation: definitions of medical domain models were researched and analyzed, finally settling on the model defined within the paradigm of Evidence Based Medicine (Sackett

et al. 2000), as opposed to, for example, the Clinical Process model (Asp & Petersen 2003). The advantage of the EBM approach to analysis of clinical medicine is that unlike many other domain models it is not concerned with identifying entities, roles and relationships, but rather with articulating principles and guidelines for providing the best known solutions to individual problems.

The next step in developing a question answering system based on the semantic domain model involved identification of the domain model components needed for development of an end-to-end question answering system. With the goal of preserving as high level of abstraction as possible, three basic components were identified:

- clinical tasks that shape information needs and predetermine types of clinical studies that are best suited for satisfaction of information needs
- a synopsis of the clinical scenario that generated an information need, which EBM formalizes as a framework for asking focused questions (PICO) and also extends to the analysis of the results of clinical studies
- reliability of the answer, which is known as Strength of Evidence.

Although the semantic domain model does not specify the desirable structure of a well-formed answer, user studies, such as Ely et al. (2005) and specialized manually created answer databases served as examples for multi-tiered answer format.

Once the basic components were identified, medical domain knowledge provided information about entities, roles, and relationships within the components. Much of this information is available in machine readable form through UMLS and

tools for medical text processing. Specific design decisions, and development of the question answering system architecture that utilizes the basic components in a principled way, were presented in Chapter 4.

In many cases, utilization of the semantic domain model relies upon available resources, such as indexing of MEDLINE citations and UMLS concept identification, combined with simple manually constructed rules. However, the complexity of patient outcome statements that play an important role in answer generation called for application of statistical methods, which, in turn, required development of an annotated collection of MEDLINE citations and development of methods for combining information extraction based on statistical methods and knowledge-based methods derived from the domain knowledge formalized in the UMLS.

The applicability of the system architecture for complex question answering in the medical domain is demonstrated through several evaluations described in Chapter 5. These evaluations required creation of test collections and adaptation of the question answering evaluation methods.

6.2 Implications

This dissertation demonstrates how a semantic domain model of evidence based medicine is used to implement an end-to-end question answering system. Overall, the above-outlined steps, which were instrumental in achieving this goal, should serve as a good starting point for development of systems capable to answer complex domain specific questions.

Practical implications of this work for medical informatics are threefold:

- The prototype question answering system will permit researching issues concerning user preferences directly, such as utility of primary literature sources and answers derived from these sources at the point of care, or verification of the practicality of asking a structured question as opposed to a free form natural language question, or a short set of key terms.

- Knowledge extraction modules will be used to create a repository of facts, such as a database of patient oriented outcomes that could be used for informed decision making serving as a starting point for meta-analysis of clinical trials, or reviews for patients.

- Principles discovered in development of the system may be used to improve document retrieval and ranking. For example, indicators of the clinical task orientation of a MEDLINE citation are used in Essie to re-score documents for task-specific searches.

6.3 Limitations

There are several limiting factors to this research. The main focus of the work is on the novel approach to complex question answering that involves use of domain knowledge, which does not permit an in-depth research of all components of the question answering system. The design and implementation of the prototype system leaves many open questions, starting with the user interface issues that were

not addressed in the current work, that will ultimately determine the success or failure of a system, should this prototype be implemented in a publicly used system.

Another equally important issue not addressed in the current implementation is the form in which a question is submitted to the system. In the evaluations conducted in this work, the questions were presented in PICO frames which required different degrees of manual effort for different evaluations. The same limitation extends to PubMed queries. Controversial evidence exists for the utility of a query interface based on structured PICO frames (Booth, O'Rourke, & Ford 2000). For example, only 8% of search requests using the MD on Tap client (Hauser *et al.* 2004) were submitted through the PICO form in July 2006. Automatic post-processing of the queries to generate question frames might degrade the quality of the answers.

Another limitation of the currently implemented prototype, common to all question answering systems that rely upon an information retrieval step performed by an "of-the-shelf" search engine, is that the quality of the set of documents given to the system depends on the quality of the query to the search engine and the search engine effectiveness. There is a demonstrable difference in the quality of the results between PubMed searches performed by average and experienced users. For example, given 30 clinical questions that resident physicians could not answer using MEDLINE, SH, an experienced MEDLINE indexer, found answers to 26 of those using complex PubMed search strategies. Such individually crafted queries were used in the *FPIN* collection experiments to evaluate the answer quality against a strong baseline. However, comparable quality of the initial document set and improvements over the baseline were achieved in the *CE* collection experiments, in

which documents were retrieved using a standard therapy-oriented search strategy. The generalizability of this approach (using predetermined query templates such as PubMed Clinical Queries or search hedges developed by FPIN librarians) needs to be verified in the future. Another possibility is to replace PubMed with a search engine that achieves good performance even given short queries from inexperienced users, for example, Essie, a probabilistic search engine developed at the National Library of Medicine. Essie was not used in the prototype implementation because its API is not publicly available. The pilot experiments were performed on the internal NLM servers.

Another possibility further down the line is to directly use information from the Semantic Medline databases that will store semantic information, complementary to the information available through PubMed, for every MEDLINE citation. Creation of such databases has been proposed within the long-term planning initiative for the NLM.

Several limitations stem from the absence of test collections that need to be developed to replace a mostly *ad hoc* nature of parameter setting in the citation scoring algorithm and within its three components with weights learned from data, as well as to improve largely manually-crafted indicators for clinical task determination.

Redundancy removal achieved through semantic clustering is also somewhat limited in that it is not propagated to deeper levels of the answer. The single-document extractive approach to providing context for an answer should be replaced with multi-document context generation in the future.

A practical limitation of the prototype is in the time requirements for semantic

pre-processing of the documents. Due to complex natural language processing, the prototype is capable of processing on the order of a 100 new abstracts for approximately 15 queries per day, and it is not capable of processing a new question and providing an answer in real time. At present, this limitation cannot be resolved purely algorithmically. Information for modules that rely on mapping of text to UMLS concepts will need to be pre-computed.

Finally, a limitation of this work is in that although it demonstrates the feasibility of complex question answering based on a semantic domain model, it does so only for the domain of clinical medicine. Possible applicaton of the basic principles of this approach to other domains is outlined in the future work section.

6.4 Future work

The goal of the current work was to investigate a new approach to complex question answering. This approach realizes some of the early visions of using automatically populated frame systems. The availability of suitable resources made this approach feasible in the domain of clinical medicine. Applicability of the proposed approach in another domain needs to be verified. The legal domain is perhaps best suited to test how the implementation steps presented in this dissertation could be applied to a different domain. This domain has a potential for development of question answering frameworks based on consistent application of a domain model, since many of the essential components identified in this dissertation are well developed in legal information research. A legal domain ques-

tion answering system should encompass the roles enacted by lawyers. The roles, such as drafting, advocacy, negotiating, and counseling each address different tasks that are formalized similarly to clinical tasks (Leckie, Pettigrew, & Sylvain 1996; Komlodi & Soergel 2002). The domain has two databases analogous to MEDLINE: LEXIS and WESTLAW; core ontologies (Breuker 2004); and beginnings of an adequate domain model (Hafner & Berman 2002). These resources, and the existing specialized search engines and domain-specific knowledge-based systems (Moens & Spyns 2005) are well suited to satisfy information needs requiring verification of known cases. What seems to be missing in this domain, is a framework for formulation of focused questions in difficult cases, when an attorney may not even know where to begin the search. In the absence of such framework, a user studies applicable material hoping that a previously unknown connection or unseen pattern emerges (Kuhlthau & Tama 2001). In satisfying their information needs, lawyers prefer informal sources internal to their organizations rather than external sources (Wilkinson 2001). The difference in users' behavior in the law and clinical domain might be consequential to the fundamental differences between structured queries, widely available for specialized law domain searches and well suited for conveying well-specified requests, and a framework for formulating specific requests based on analysis of an underlying scenario. The EBM PICO framework was developed exactly for the purposes of untangling a mass of confusing facts and focusing on what is important in a given scenario, as opposed to, for example, structured queries available through PubMed advanced search options that provide an opportunity to somewhat focus a search, but are not a substitute for the PICO framework. It seems

that a similar framework for legal domain is a final missing component that, when developed, will enable application of a model proposed in this work to legal domain. It might be possible to test this hypothesis within the new TREC legal track, which explores searches on *topics* approximating real legal requests in litigation.

The genomics track evaluation in TREC is another venue for testing the proposed approach. Based on information needs and questions collected in several biochemical laboratories, a group of volunteers with biomedical background and knowledge of information retrieval developed generic topic types and identified question patterns corresponding to each type. After the 2006 evaluation, a test collection containing questions based on generic topic types and relevant passages that answer these questions will be available. Another research direction facilitated by this collection will be answering questions using full text of the articles rather than just the abstracts.

In addition to branching out to new domains and different media types for answer generation, future directions include the in-depth development of the system, such as learning component weights from clinical questions-relevant citations pairs that are being collected by clinical residents at the John A. Burns School of Medicine (Demner-Fushman *et al.* 2006); and extension of multi-document answer generation to outcome statements. Imrovements in determining the strength of evidence might be acchieved using the ISI journal impact factor and other Journal Citation Reports metrics. It is not clear without a further study if these metrics could be used to classify articles published in a journal currently not rated as a high-impact medical journal. The frequency with which the "average article" in a journal

has been cited might or might not correlate with its value to practitioners. Another research direction is to use weighted ontological distance in clustering, which should eliminate the need for manual truncation of the top ontology categories. The ongoing exploration of multi-document answer generation based on comparative relations identified using SemRep will also continue in the future (see Appendix D for an in-depth analysis of the intermediate results of SemRep experiments.)

Other complex and interesting research areas that have not been addressed in the present work include development of a better frame structure for clinical questions; development of user interfaces; automatic transformation of natural language questions into question frames, and translation of question frames into successful search strategies; handling of contradictory and inconsistent evidence; development of storage, access, and data mining tools for the facts extracted from biomedical text. Finally, the created test collections provide an opportunity to evaluate feasibility of automatic concept-based evaluation in complex question answering.

In conclusion, this dissertation posed an important research question, whether consistent application of a semantic domain model of clinical medicine to all steps of the question answering process would result in successful generation of answers to complex clinical questions. This question was answered positively in several evaluations enabled by the working prototype of an end-to-end clinical question answering system developed in this work. The development of the system involved combining the classic AI frame-based approach to question answering with statistical NLP methods, resulting in a novel hybrid approach to question answering. The work on this dissertation posed many additional questions and opened a broad extent of

future venues ranging from application of the proposed schema in a new domain to in-depth studies of many components of the prototype system.

Appendix A

Search strategies for PICO-annotated collection

1. ((("arthritis, rheumatoid" [MeSH Terms] OR RHEUMATOID ARTHRITIS [Text Word]) OR ("migraine" [MeSH Terms] OR MIGRAINE [Text Word])) OR ("breast neoplasms" [MeSH Terms] OR BREAST CANCER [Text Word])) AND (randomized controlled trial [ptyp] OR ((randomized [Title/Abstract] AND controlled [Title/Abstract]) AND trial [Title/Abstract])) AND jsubsetaim [text] AND ("1999/1/1" [PDat] : "2004/1/1"[PDat]))

2.
   - ("tuberculosis, pulmonary" [MeSH Terms] OR pulmonary tuberculosis [Text Word]) AND hasabstract [text] AND Randomized Controlled Trial [ptyp] AND English [Lang] AND ("human" [MeSH Terms] OR "hominidae" [MeSH Terms]) AND ("1999/01/01" [PDAT] : "2004/01/01" [PDAT]))
   - ("hypertension, renal" [MeSH Terms] OR renal hypertension [Text Word]) AND hasabstract [text] AND Randomized Controlled Trial [ptyp] AND English [Lang] AND ("human" [MeSH Terms] OR "hominidae" [MeSH Terms]) AND ("1999/01/01" [PDAT] : "2004/01/01" [PDAT])
   - ("asthma, exercise-induced" [MeSH Terms] OR asthma, exercise-induced [Text Word]) AND hasabstract [text] AND Randomized Controlled Trial [ptyp] AND English [Lang] AND ("human" [MeSH Terms] OR "hominidae" [MeSH Terms]) AND ("1999/01/01" [PDAT] : "2004/01/01" [PDAT])

3. "treatment outcome" [MeSH Terms] AND "loattrfree full text"[sb] AND hasabstract [text] AND Randomized Controlled Trial [ptyp] AND English [Lang] AND "humans" [MeSH Terms]

4. (immunizations [Text Word] OR immunisations [Text Word] OR "immunization" [MeSH Terms]) AND hasabstract [text] AND English [Lang] AND ("infant, newborn" [MeSH Terms] OR "child, preschool" [MeSH Terms] OR "infant" [MeSH Terms]) AND ("adverse effects" [Subheading] OR adverse effects [Text Word])

5. (diabetes mellitus [Text Word] OR "diabetes mellitus" [MeSH Terms] OR diabetes insipidus [Text Word] OR "diabetes insipidus" [MeSH Terms] OR diabetes [Text Word]) AND hasabstract [text] AND English [Lang] AND ("human" [MeSH Terms] OR "hominidae" [MeSH Terms])

Appendix B
Questions in the FPIN collection

Questions marked with (P) are from the Parkhurst Exchange Forum. The remaining questions are from FPIN.

B.1 Training

B.1.1 Therapy

1. What is the best treatment for analgesic rebound headaches?
2. What is the interval for monitoring warfarin therapy once therapeutic levels are achieved?
3. Does quinine reduce leg cramps for young athletes?
4. Does acyclovir help herpes simplex virus cold sores if treatment is delayed?
5. How effective is prophylactic therapy for gout in people with prior attacks?
6. First- or second-generation antihistamines: which are more effective at controlling pruritus?
7. Does combining aspirin and warfarin decrease the risk of stroke for patients with nonvalvular atrial fibrillation?
8. What is the most effective nicotine replacement therapy?
9. (P) Is it safe to follow a 40-year-old male patient with a 19% spontaneous pneumothorax conservatively?
10. (P) What are the best medications for panic disorder?

B.1.2 Diagnosis

1. What is the diagnostic approach to a 1-year-old with chronic cough?
2. For knee pain, how predictive is physical examination for meniscal injury?
3. Does a Short Symptom Checklist accurately diagnose ADHD?
4. What is the differential diagnosis of chronic diarrhea in immunocompetent patients?
5. How often is coughing the presenting complaint in patients with gastroesophageal reflux disease?

6. (P) How would you manage a woman with brownish discharge from one of her breasts? She is premenopausal (less than 50 years old).

B.1.3 Prognosis

1. Should we screen for bacterial vaginosis in those at risk for preterm labor?
2. (P) What's the prognosis of lupoid sclerosis?
3. (P) What is the risk for chronic active hepatitis, cirrhosis, and hepatoadenocarcinoma in an asymptomatic 45-year-old male hepatitis B virus (HBV) carrier with no history of illness, strongly positive result for HBsAg and practically none for HBsAb.

B.1.4 Etiology

1. What are the causes of hypomagnesemia?
2. (P) Does maternal smoking and second-hand smoke cause ADHD?
3. (P) Can diverticulosis cause repeatedly positive occult fecal blood?
4. (P) Can selective serotonin reuptake inhibitor (SSRI) use cause impulsive suicidal or homicidal behaviour?
5. (P) Can topiramate cause kidney stones?

B.2 Evaluation

B.2.1 Therapy

1. What is the most effective treatment for ADHD in children?
2. What are effective treatments for oppositional and defiant behaviors in preadolescents?
3. Is antibiotic prophylaxis effective for recurrent acute otitis media?
4. Other than anticoagulation, what is the best therapy for those with atrial fibrillation?
5. Do acetaminophen and an NSAID combined relieve osteoarthritis pain better than either alone?
6. What regimens eradicate Helicobacter pylori?
7. (P) What's the current success rate of electroconvulsive therapy (ECT)?
8. (P) What's the best treatment for epididymitis?

9. (P) Are there any new drugs for the treatment of slowly progressive mS?
10. (P) What's your opinion of eye movement desensitization and reprocessing (EMDR) as a treatment for posttraumatic stress disorder (PTSD)?
11. (P) Could stimulants be useful for chronic fatigue syndrome?
12. (P) Do TCAs or SSRIs have any effect on decreasing tinnitus?

B.2.2 Diagnosis

1. What are the indications for evaluating a patient with cough for pertussis?
2. Can transvaginal ultrasound detect endometrial disease among asymptomatic postmenopausal patients?
3. Is the ThinPrep better than conventional Pap smear at detecting cervical cancer?
4. How accurate is stress radionuclide imaging for diagnosis of CAD?
5. (P) What's the significance of splinter hemorrhages in a healthy 40-year-old woman?
6. (P) In a patient with back pain, is parasthesia caused by a spinal cord compression or a disk problem?

B.2.3 Prognosis

1. What is the prognosis for acute low back pain?
2. (P) What's the prognosis of human papillomavirus of the throat?
3. (P) Does a polyp in the gallbladder pose any risk of becoming malignant?

B.2.4 Etiology

1. (P) Can a very low serum iron cause fatigue in a patient whose hemoglobin and red blood cell (RBC) counts are normal?
2. (P) Can measles-mumps-rubella (MMR) vaccine cause autism in children?
3. (P) Can a short course of steroids (one or two weeks) result in avascular hip necrosis?
4. (P) Could a flare-up of the Crohn's cause miscarriage at eight weeks?
5. (P) What are the causes of myokymia?

Appendix C
Diseases in the CE collection
C.1 Tuning set

1. acute asthma
2. chronic prostatitis
3. community acquired pneumonia
4. erectile dysfunction
5. osteoporosis

C.2 Evaluation set

1. COPD (Chronic Obstructive Pulmonary Disease)
2. PTSD (Post-Traumatic Stress Disorder)
3. Parkinson's disease
4. acute bronchitis
5. acute ischaemic stroke
6. acute sinusitis
7. atrial fibrillation (heart rate control)
8. bacterial conjunctivitis
9. chlamydia
10. chronic asthma
11. chronic psoriasis
12. generalised anxiety disorder
13. genital warts
14. glaucoma
15. gonorrhea
16. leg cramps

17. mania
18. meningitis
19. obesity
20. osteoarthritis
21. otitis externa
22. pneumocystis carinii pneumonia
23. rheumatoid arthritis
24. scabies
25. schizophrenia

Appendix D
Extracting comparative structures

SemRep identifies two types of comparison relations: the first asserts that two drugs are compared; the second provides additional information about the scale on which the drugs are compared, and the relative position of the drugs on the scale. The second type is used to rank extracted interventions based on their comparative effectiveness (rather than frequency and rank of the documents from which they were extracted, as in the semantic clustering method described in Section 4.6.1.) Because of the sparseness of comparative studies, this processing cannot be applied to each semantic cluster, but rather presents an alternative approach to semantic clustering and redundancy removal. This approach is illustrated here using a manual analysis of SemRep output for the question, *What are the effects of treatments for acute bronchitis in people without chronic respiratory disease?* SemRep output for 256 citations retrieved to answer this question was filtered to find the second type of comparative relations. The majority of the type 2 comparative structures found in thirty-one citations are meaningful (see Figure D.1.)

Manual analysis shows that in 19 of 31 citations comparisons were recognized correctly. All but three of the comparisons are either of two antibiotics, or between different regimens for an antibiotic, or an antibiotic compared to placebo, or other drug classes. Antibiotics fell into four disjoint graphs. Except for the last relation *albuterol is higher than antibiotics on the effectiveness scale*, the manual process presented in Figure D.1 would be easily automated using simple rules. However,

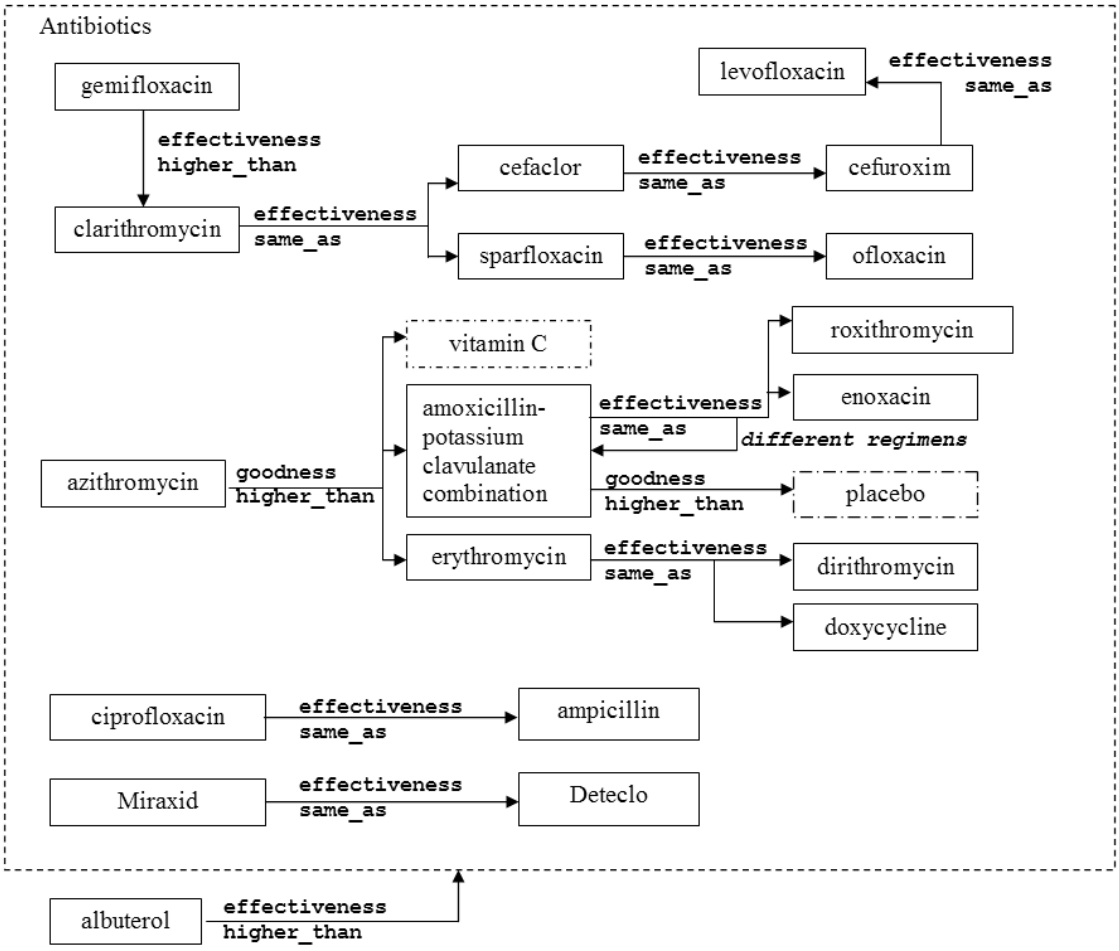

Figure D.1: Drug comparisons identified using SemRep.

there are several reasons to keep investigating this approach without integrating it with the system yet. Since the comparatives identification is in the early stages of its development, the confidence in the comparatives output is not as high as in the other relations extracted by SemRep. For example, the assertion *azithromycin is higher than vitamin C in goodness* is incorrect, the actual outcome is *azithromycin is no better than low-dose vitamin C for acute bronchitis*. It is also not clear if the statement *azithromycin was as effective as and better tolerated than erythromycin or amoxicillin* means a *higher_than* relationship on the *goodness* scale. Other 12 relations were excluded from being presented in the graph for various reasons. One of those was excluded because of the fairly meaningless scale *less*. The context from which it was extracted was not comparing interventions, so it did not clarify the meaning of one drug being less than the other. Another one was excluded because of the relation *frequency*. This relation did not quite capture the complex comparisons:

> The incidence of diarrhea and other gastrointestinal symptoms was significantly more frequent in the amoxicillin/clavulanate group (13.5% and 5.6%) than in the loracarbef group (4.5% and 1.7%), while the incidence of severe headaches was significantly more frequent in the loracarbef than the amoxicillin/clavulanate group (7.2% vs 3.1%).

Ideally, we should capture that the drugs are compared on their side effects, which involves knowledge and reasoning not currently available. In addition, one equivalence relation was processed incorrectly because it required anaphora resolution: the text refers to *both treatments*, and then names only one of them. Another complex comparison that was captured only partially was *goodness* in the following sentence: *The acute success was similar for doxycycline with the other antibacterials and was superior to cefaclor*. The *antibacterial higher_than cefaclor* was captured,

which unfortunately does not make sense because, according to at least one UMLS hierarchy, *cefaclor* is an *antibacterial*. Another nuance that was not captured properly is the duration of an infection-free period in the following statement: *The infection-free period was longer after doxycycline than with the other four antimicrobials*. The relation was identified as *doxycycline* being *higher_than antimicrobial* on the *length* scale. It is not clear how one drug is longer than the other, and *doxycycline* is an *antimicrobial*. There are several cases that are not exactly wrong, but involve concepts on a level that is not practically useful, these are: *regimen seems to be as effective as conventional treatment*; *fluoroquinolones appeared to be as effective as standard antibiotic regimens*; *Therapeutic procedure same_as Therapeutic procedure* on *effectiveness* scale, etc. Two errors come from abbreviation expansion. The first article says: *grepafloxacin 400 mg or 600 mg od is as effective as amoxycillin 500 mg tds in the treatment of ABECB*. UMLS associates the following concepts with the term *tds*:

Chad, Trinidad and Tobago

tyramine-deoxysorbitol

Tetanus and diphtheria toxoid adsorbed for adult use

but not the sense used in the article: three-times daily. Since *tyramine-deoxysorbitol* is an organic chemical, it is identified as being compared to the clinical drug *grepafloxacin*, and the comparison says *grepafloxacin is same_as tyramine-deoxysorbitol* on *effectiveness* scale. Another error occurs in the following sentence: *cefuroxime axetil 250 mg BID is as effective as amoxicillin/clavulanate 500 mg TID in the treatment of patients with acute bronchitis*. Here *BID* is recognized as *BIDS,*

$4-benzamido-4'-isothiocyanostilbene-2,2'-disulfonate$, as opposed to the *twice a day* sense intended in the article. Despite the still unresolved issues, the prevailing positive examples provided above show comparative relations to be a very powerful tool that will advance answer generation to a new level, once it is mature enough.

Appendix E

Coefficients for document ranking and answer generation

With few exceptions, the coefficient values were set heuristically based on recommendations for critical appraisal of medical literature, intuition, and observations on a training set. The differences between the results obtained using the heuristically set coefficients and the automatically optimized values were not statistically significant.

Table E.1: Coefficients and values.

Coefficient	Heuristic values	Automatically learned values
PICO, SoE, Task (equation 4.5	1, 1, 1	0.38, 0.34, 0.28 (Demner-Fushman & Lin 2006 in press)
Problem (equation 4.6)	$1\|-1\|0.5\|-0.5$	
Population (equation 4.6)	$1\|0$	
Intervention (equation 4.6)	$1\|0\|-0.5$	
Outcome (equation 4.6)	1	
Study type (equation 4.3)	$0.5\|0.3\|0.2$	
Journal type (equation 4.3)	$0.6\|0.3$	
Outcome classifiers (equation 4.1)		0.03, 0.20, 0.08, 0.30, 0.20, 0.07

References

Ad Hoc Working Group for Critical Appraisal of the Medical Literature. 1987. A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine* 106:595–604.

AHRQ. 2002. Systems to rate the strength of scientific evidence. Technical Report No. 02-P0022, Agency for Healthcare Research and Quality.

Alper, B. S.; Stevermer, J. J.; White, D. S.; and Ewigman, B. G. 2001. Answering family physicians clinical questions using electronic medical databases. *The Journal of Family Practice* 50(11):960–965.

Alper, B. S.; White, D. S.; and Ge, B. 2005. Physicians answer more clinical questions and change clinical decisions more often with synthesized evidence: a randomized trial in primary care. *Annals of Family Medicine* 3(6):507–513.

Aronson, A. R., and Rindflesch, T. C. 1997. Query expansion using the umls metathesaurus. In *Proceeding of the 1997 Annual Symposium of the American Medical Informatics Association (AMIA 1997)*, 485–489.

Aronson, A. R.; Demner-Fushman, D.; Humphrey, S. H.; Ide, N. C.; Kim, W.; Liu, H.; Loane, R. R.; Mork, J. G.; Smith, L. H.; Tanabe, L. K.; Wilbur, W. J.; and Xie, N. 2004a. Knowledge-intensive and statistical approaches to the retrieval and annotation of genomics medline citations. In Voorhees, E. M., and Buckland, L. P., eds., *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004), November 2004, Gaithersburg, Maryland*. National Institute of Standards and Technology.

Aronson, A. R.; Mork, J. G.; Gay, C. W.; Humphrey, S. M.; and Rogers, W. J. 2004b. The nlm indexing initiative‘s medical text indexer. In *Proceedings of the 11th World Congress on Medical Informatics (MEDINFO 2004)*, 268–272.

Aronson, A. R.; Demner-Fushman, D.; Humphrey, S. H.; Lin, J.; Liu, H.; Ruch, P.; Ruiz, M. E.; Smith, L. H.; Tanabe, L. K.; and Wilbur, W. J. 2005. Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. In Voorhees, E. M., and Buckland, L. P., eds., *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005), November 2005, Gaithersburg, Maryland*. National Institute of Standards and Technology.

Aronson, A. R. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proceeding of the 2001 Annual Symposium of the American Medical Informatics Association (AMIA 2001)*, 17–21.

Asp, L., and Petersen, J. 2003. A conceptual model for documentation of clinical information in the ehr. *Studies in health technology and informatics* 95:239–244.

Baeza-Yates, R., and Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. Addison Wesley Longman Publishing Co. Inc.

Benamara, F. 2004. Cooperative question answering in restricted domains: The WEBCOOP experiment. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*.

Bergus, G. R.; Randall, C. S.; Sinift, S. D.; and Rosenthal, D. M. 2000. Does the structure of clinical questions affect the outcome of curbside consultations with specialty colleagues? *Archives of Family Medicine* 9(6):541–547.

Bishop, A. P. 1999. Document structure and digital libraries: how researchers mobilize information in journal articles. *Information Processing and Management* 35(3):255–279.

Blair-Goldensohn, S.; McKeown, K. R.; and Schlaikjer, A. H. 2004. *New Directions In Question Answering*. Menlo Park, CA: AAAI Press. chapter 4, 47–58.

Booth, A., and O'Rourke, A. 1997. The value of structured abstracts in information retrieval from medline. *Health Libraries Review* 14(3):157–166.

Booth, A.; O'Rourke, A.; and Ford, N. J. 2000. Structuring the pre-search reference interview: a useful technique for handling clinical questions. *Bulletin of the Medical Library Association* 88(3):239–246.

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

Breuker, J. 2004. Constructing a legal core ontology: Lri-core. In Freitas, F.; Stuckenschmidt, H.; ; and Volz, R., eds., *Workshop on Ontologies and their Applications WONTO'2004, September 28th 2004, Sao Luis, Maranhao, Brazil*, 115–126. Porto Alegre, BR: LivroRapido.

Brill, E.; Lin, J.; Banko, M.; Dumais, S. T.; and Ng, A. Y. 2001. Data-intensive question answering. In *Text REtrieval Conference*.

Browne, A. C.; Divita, G.; Aronson, A. R.; and McCray, A. T. 2003. Umls language and vocabulary tools. In *Proceeding of the 2003 Annual Symposium of the American Medical Informatics Association (AMIA 2003)*, 798.

Bryant, S. L. 2005. User preferences for clinical question answering services. Available from: http://www.tvsha.nhs.uk/libraries/doc/$TVOL_CQAS_User_surveys.pdf$. cited 2006 Jul 26.

Buckley, C., and Voorhees, E. M. 2004. Retrieval evaluation with incomplete information. In *SIGIR*, 25–32. ACM.

Callan, J. P.; Croft, W. B.; and Harding, S. M. 1992. The INQUERY retrieval system. In *Proceedings of DEXA-92, 3rd International Conference on Database and Expert Systems Applications*, 78–83.

Card, S. K.; Mackinlay, J. D.; and Shneiderman, B., eds. 1999. *Readings in Information Visualization: Using Vision to Think*. San Francisco, CA, USA: Morgan Kaufmann Publishers.

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

Chen, H.; Lally, A. M.; Zhu, B.; and Chau, M. 2003. Helpfulmed: Intelligent searching for medical information over the internet. *Journal of the American Society for Information Science and Technology (JASIST)* 54(7):683–694.

Chung, H.; Song, Y.-I.; Han, K.-S.; Yoon, D.-S.; Lee, J.-Y.; Rim, H.-C.; and Kim, S.-H. 2004. A practical QA system in restricted domains. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*, 39–45.

Cimino, J. J.; Clayton, P. D.; Hripcsak, G.; and Johnson, S. B. 1994. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *Journal of the American Medical Informatics Association (JAMIA)* 1(1):35–50.

Cogdill, K. W., and Moore, M. E. 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.

Cohen, A. M.; Stavri, P. Z.; and Hersh, W. R. 2004. A categorization and analysis of the criticisms of evidence-based medicine. *International Journal of Medical Informatics* 73(1):35–43.

Cook, R. J., and Sackett, D. L. 1995. The number needed to treat: a clinically useful measure of treatment effect. *BMJ (Clinical research)* 310(6977):452–454.

Cooper, W. S. 1964. Fact retrieval and deductive question-answering information retrieval systems. *Journal of the ACM (JACM)* 11(2):117–137.

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

Cui, H.; Kan, M.-Y.; and Chua, T.-S. 2005. Generic soft pattern models for definitional question answering. In Baeza-Yates, R. A.; Ziviani, N.; Marchionini, G.; Moffat, A.; and Tait, J., eds., *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development of Information Retrieval*, 384–391. ACM.

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

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

Dee, C., and Blazek, R. 1993. Information needs of the rural physician: A descriptive study. *Bulletin of the Medical Library Association* 8(1):259–264.

Demner-Fushman, D., and Lin, J. 2006a. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*.

Demner-Fushman, D., and Lin, J. 2006b. Situated question answering in the clinical domain: Selecting the best drug treatment for diseases. In *Proceedings of COLING/ACL 2006 Workshop on Task-Focused Summarization and Question Answering*.

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

Demner-Fushman, D.; Hauser, S. E.; Ford, G.; and Thoma, G. R. 2004. Organizing literature information for clinical decision support. In *Proceedings of 11th World Congress on Medical Informatics (MEDINFO 2004)*, 602–606.

Demner-Fushman, D.; Hauser, S. E.; Humphrey, S. M.; Ford, G.; Jacobs, J. L.; and Thoma, G. R. 2006. Medline as a source of just-in-time answers to clinical questions. In *Proceeding of the 2006 Annual Symposium of the American Medical Informatics Association (AMIA 2006)*.

Demner-Fushman, D.; He, D.; and Oard, D. W. 2004. Exploring interactive relevance feedback with a two-pass study design. Technical Report LAMP-TR-116, CAR-TR-1001, CS-TR-4621, UMIACS-TR-2004-63, University of Maryland, College Park.

Denny, J. C.; Smithers, J. D.; Spickard, A.; and Miller, R. A. 2002. A new tool to identify key biomedical concepts in text documents, with special application to curriculum content. In *Proceeding of the 1997 Annual Symposium of the American Medical Informatics Association (AMIA 1997)*, 1007.

Diekema, A. R.; Yilmazel, O.; and Liddy, E. D. 2004. Evaluation of restricted domain question-answering systems. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*.

Doan-Nguyen, H., and Kosseim, L. 2004. The problem of precision in restricted-domain question answering. some proposed methods of improvement. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*.

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

Ebell, M. H.; Siwek, J.; Weiss, B. D.; Woolf, S. H.; Susman, J.; Ewigman, B.; and Bowman, M. 2004. Strength of Recommendation Taxonomy (SORT): A patient-centered approach to grading evidence in the medical literature. *The Journal of the American Board of Family Practice* 17(1):59–67.

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

Ely, J. W.; Osheroff, J. A.; Gorman, P. N.; Ebell, M. H.; Chambliss, M. L.; Pifer, E. A.; and Stavri, P. Z. 2000. A taxonomy of generic clinical questions: classification study. *BMJ* 321:429–432.

Ely, J. W.; Osheroff, J. A.; Chambliss, M. L.; Ebell, M. H.; and Rosenbaum, M. E. 2005. Answering physicians‘ clinical questions: Obstacles and potential solutions. *Journal of the American Medical Informatics Association* 12(2):217–224.

Evidence-Based Medicine Working Group. 1992. Evidence-based medicine. a new approach to teaching the practice of medicine. *The Journal of the American Medical Association* 268(17):2420–2425.

Fiszman, M.; Rindflesch, T. C.; and Kilicoglu, H. 2004. Abstraction summarization for managing the biomedical research literature. In *Proceedings of the HLT/NAACL 2004 Workshop on Computational Lexical Semantics*.

Flaherty, R. J. 2004. A simple method for evaluating the clinical literature. *Family Practice Management* 11(5):47–52.

Florance, V.; Giuse, N. B.; and Ketchell, D. S. 2002. Information in context: integrating information specialists into practice settings. *Journal of the Medical Library Association* 90(1):49–58.

Focht, 3rd, D. R.; Spicer, C.; and Fairchok, M. P. 2002. The efficacy of duct tape vs cryotherapy in the treatment of verruca vulgaris (the common wart). *Archives of Pediatrics and Adolescent Medicine* 156(10):971–974.

Fontelo, P.; Nahin, A.; Liu, F.; Kim, G.; and Ackerman, M. 2005. Accessing medline/pubmed with handheld devices: Developments and new search portals. In *Proceedings of the 38th Hawaii International Conference on System Sciences*, 1–5.

Fox, E. A., and Shaw, J. A. 1994. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC-2)*, 243–252.

Friedman, C.; Kra, P.; and Rzhetsky, A. 2002. Two biomedical sublanguages: a description based on the theories of zellig harris. *Journal of Biomedical Informatics* 35(4):222–235.

Fukumoto, J.; Kato, T.; and Masui, F. 2004. An evaluation of question answering challenge (qac-1) at the ntcir workshop 3. *SIGIR Forum* 38(1):25–28.

Gabbay, I., and Sutcliffe, R. F. 2004. A qualitative comparison of scientific and journalistic texts from the perspective of extracting definitions. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*.

Garg, A. X.; Adhikari, N. K. J.; McDonald, H.; Rosas-Arellano, M. P.; Devereaux, P. J.; Beyene, J.; Sam, J.; and Haynes, R. B. 2005. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. *The Journal of the American Medical Association* 293(10):1223–1238.

Giuse, N. B.; Huber, J. T.; Giuse, D. A.; Brown Jr., C. W.; Bankowitz, R. A.; and Hunt, S. 1994. Information needs of health care professionals in an aids outpatient clinic as determined by chart review. *The Journal of the American Medical Informatics Association* 1(5):395–403.

Gorman, P. N., and Helfand, M. 1995. Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. *Medical Decision Making* 15(2):113–119.

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

Guyatt, G. H.; Sackett, D.; and Cook, D. J. 1994. Users‘ guides to the medical literature. ii. how to use an article about therapy or prevention. b. what were the

results and will they help me in caring for my patients? evidence-based medicine working group. *The Journal of the American Medical Association* 271(1):59–63.

Hafner, C. D., and Berman, D. H. 2002. The role of context in case-based legal reasoning: teleological, temporal, and procedural. *Artificial Intelligence and Law* 10(1-3):19–64.

Harabagiu, S., and Hickl, A. 2006. Using scenario knowledge in automatic question answering. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering at COLING-ACL 2006*, 32–39.

Harabagiu, S., and Lacatusu, F. 2004. Strategies for advanced question answering. In *Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004*, 1–9.

Harman, D. 1992. Relevance feedback revisited. In *SIGIR 1992: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark*, 1–10.

Harman, D. 1996. Evaluation techniques and measures. In *Proceedings of the 4th Text REtrieval Conference (TREC-4).*, A6–A14.

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

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

Hazell, P. L., and Stuart, J. E. 2003. A randomized controlled trial of clonidine added to psychostimulant medication for hyperactive and aggressive children. *Journal of the American Academy of Child and Adolescent Psychiatry* 42(8):886–894.

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

Hersh, W. R., and Greenes, R. A. 1990. Saphire - an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. *Computers and biomedical research, an international journal* 23(5):410–425.

Hersh, W. R.; Hickam, D. H.; Haynes, R. B.; and McKibbon, K. A. 1994. A performance and failure analysis of saphire with a medline test collection. *Journal of the American Medical Informatics Association* 1(1):51–60.

Hersh, W.; Bhupatiraju, R. T.; and Corley, S. 2004. Enhancing access to the bibliome: The TREC genomics track. In *Proceedings of the 11th World Congress on Medical Informatics (MEDINFO 2004)*, 773–777.

Hildebrandt, W.; Katz, B.; and Lin, J. 2004. Answering definition questions with multiple knowledge sources. In *Proceedings of the 2004 Human Language*

Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004).

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

Huang, X.; Lin, J.; and Demner-Fushman, D. 2006. Evaluation of pico as a knowledge representation for clinical questions. In *Proceeding of the 2006 Annual Symposium of the American Medical Informatics Association (AMIA 2006).*

Humphrey, S. M.; Rogers, W. J.; Kilicoglu, H.; Demner-Fushman, D.; and Rindflesch, T. C. 2006. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology* 57(5):726–726.

Jacquemart, P., and Zweigenbaum, P. 2003. Towards a medical question-answering system: A feasibility study. In Baud, R.; Fieschi, M.; Beux, P. L.; and Ruch, P., eds., *The New Navigators: From Professionals to Patients*, volume 95 of *Actes Medical Informatics Europe, Studies in Health Technology and Informatics*. Amsterdam: IOS Press. 463–468.

Jaeschke, R.; Guyatt, G. H.; and Sackett, D. L. 1994. Users‘ guides to the medical literature. iii. how to use an article about a diagnostic test. b. what are the results and will they help me in caring for my patients? the evidence-based medicine working group. *The Journal of the American Medical Association* 271(9):703–707.

Järvelin, K., and Kekäläinen, J. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems* 20(4):422–446.

Jurafsky, D., and Martin, J. H., eds. 2000. *Speech and Language Processing*. Upper Saddle River, New Jersey, USA: Prenyice-Hall, Inc.

Komlodi, A., and Soergel, D. 2002. Attorneys interacting with legal information systems: Tools for mental model building and task integration. In *ASIST 2002: American Society for Information Science and Technology Annual Meeting, Philadelphia, PA, November 18-21, 2002*.

Koonce, T. Y.; Giuse, N. B.; and Todd, P. 2004. Evidence-based databases versus primary medical literature: an in-house investigation on their optimal use. *Journal of the Medical Library Association* 92(4):407–411.

Krauthammer, M., and Nenadic, G. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics* 37(6):512–526.

Kuhlthau, C. C., and Tama, S. L. 2001. Information search process of lawyers: a call for "just for me" information. *Journal of Documentation* 57(1):25–43.

Laupacis, A.; Wells, G.; Richardson, W. S.; and Tugwell, P. 1994. Users‘ guides to the medical literature. v. how to use an article about prognosis. evidence-based medicine working group. *The Journal of the American Medical Association* 272(3):234–237.

Leckie, G. J.; Pettigrew, K.; and Sylvain, C. 1996. Modelling the information-seeking of professionals: a general model derived from research on engineers, health care professionals and lawyers. *Library Quarterly* 46(2):161–193.

Lehnert, W. G. 1977. A conceptual theory of question answering. In *International Joint Conference on Artificial Intelligence (IJCAI 1977)*, 158–164.

Leuski, A. 2001. Evaluating document clustering for interactive information retrieval. In *CIKM*, 33–40.

Levine, M.; Walter, S.; Lee, H.; Haines, T.; Holbrook, A.; and Moyer, V. 1994. Users‘ guides to the medical literature. iv. how to use an article about harm. evidence-based medicine working group. *The Journal of the American Medical Association* 271(20):1615–1619.

Lin, J., and Demner-Fushman, D. 2006. The role of knowledge in conceptual retrieval: A study in the domain of clinical medicine. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*.

Lin, C.-Y., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACL2003: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 71–78. Morristown, NJ, USA: Association for Computational Linguistics.

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

Lindlof, T. R., and Taylor, B. C. 2002. *Qualitative Communication Research Methods*. Thousand Oaks, CA USA: SAGE Publications, second edition.

Lita, L. V., and Carbonell, J. G. 2004. Unsupervised question answering data acquisition from local corpora. In Grossman, D.; Gravano, L.; Zhai, C.; Herzog, O.; and Evans, D. A., eds., *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management*, 607–614. ACM.

Magnini, B.; Vallin, A.; Ayache, C.; Erbach, G.; nas, A. P.; de Rijke, M.; Rocha, P.; Simov, K. I.; and Sutcliffe, R. F. E. 2004. Overview of the clef 2004 multilingual question answering track. In *Cross-Language Evaluation Forum (CLEF)*, 371–391.

Manning, C. D., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

McCallum, A. K. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

McKeown, K.; Elhadad, N.; and Hatzivassiloglou, V. 2003. Leveraging a common representation for personalized search and summarization in a medical digital library. In *Proceedings 3rd ACM/IEEE Joint Conference on Digital Libraries (JCDL 2003)*.

Mendonça, E. A., and Cimino, J. J. 2001. Building a knowledge base to support a digital library. In *Proceedings of 10th World Congress on Medical Informatics (MEDINFO 2001)*, 222–225.

Metzler, D., and Croft, W. B. 2004. Combining the language model and inference network approaches to retrieval. *Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval* 40(5):735–750.

Minsky, M. 1975. *A framework for representing knowledge.* New York, NY: McGraw-Hill. Also available at http://web.media.mit.edu/ minsky/papers/Frames/frames.html.

Mladenic, D., and Grobelnik, M. 1999. Feature selection for unbalanced class distribution and Naïve Bayes. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999)*, 258–267.

Moens, M.-F., and Spyns, P., eds. 2005. *Legal Knowledge and Information Systems*, volume 134 of *Frontiers in Artificial Intelligence and Applications.* Amsterdam, The Netherlands: IOS Press.

Moldovan, D. I.; Pasca, M.; Harabagiu, S. M.; and Surdeanu, M. 2003. Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems (TOIS)* 21(2):133–154.

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

Perry, C. A. 1990. Knowledge bases in medicine: a review. *Bulletin of the Medical Library Association* 78(3):271–282.

Prager, J.; Brown, E.; Coden, A.; and Radev, D. 2000. Question-answering by predictive annotation. In Belkin, N. J.; Ingwersen, P.; and Leong, M.-K., eds., *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2000, Athens, Greece*, 184–191. ACM.

Pratt, W., and Wasserman, H. 2000. Querycat: Automatic categorization of medline queries. In *Proceeding of the 2000 Annual Symposium of the American Medical Informatics Association (AMIA 2000)*, 655–659.

Pratt, W., and Yetisgen-Yildiz, M. 2003. A study of biomedical concept identification: MetaMap vs. people. In *Proceeding of the 2003 Annual Symposium of the American Medical Informatics Association (AMIA 2003)*, 529–533.

Pratt, W.; Hearst, M. A.; and Fagan, L. M. 1999. A knowledge-based approach to organizing retrieved documents. In *AAAI/IAAI*, 80–85.

Purcell, G. P.; Rennels, G. D.; and Shortliffe, E. H. 1997. Development and evaluation of a context-based document representation for searching the medical literature. *International Journal of Digital Libraries* 1(3):288–296.

Ravichandran, D., and Hovy, E. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, 41–47.

Resnik, P. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* 11:95–130.

Richardson, W. S., and Wilson, M. C. 1997. On questions, background and foreground. *Evidence Based Health Care Newsletter* 17:8–9.

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

Riloff, E. 1996. Automatically generating extraction patterns from untagged text. In *AAAI/IAAI, Vol. 2*, 1044–1049.

Rinaldi, F.; Dowdall, J.; Schneider, G.; and Persidis, A. 2004. Answering questions in the genomics domain. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*.

Rindflesch, T. C., and Fiszman, M. 2003. 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.

Rosenberg, W., and Donald, A. 1995. Evidence based medicine: an approach to clinical problem-solving. *British Medical Journal* 310(6987):1122–1126.

Russell, S., and Norvig, P. 2003. *Artificial Intelligence: A Modern Approach*. Prentice Hall, second edition.

Sackett, D. L.; Straus, S. E.; Richardson, W. S.; Rosenberg, W.; and Haynes, R. B. 2000. *Evidence-Based Medicine: How to Practice and Teach EBM*. Edinburgh: Churchill Livingstone, second edition.

Seideman, P.; Samuelson, P.; and Neander, G. 1993. Naproxen and paracetamol compared with naproxen only in coxarthrosis. increased effect of the combination in 18 patients. *Acta orthopaedica Scandinavica* 64(3):285–288.

Seol, Y.-H.; Kaufman, D. R.; Mendonça, E. A.; Cimino, J. J.; and Johnson, S. B. 2004. Scenario-based assessment of physicians information needs. In Fieschi, M.; Coiera, E.; and Li, Y.-C. J., eds., *Proceedings of the 11th World Congress on Medical Informatics*, 306–310. Amsterdam: IOS Press.

Sewell, W. 1964. Medical subject headings in medlars. *Bulletin of the Medical Library Association* 52(1):164–170.

Shatkay, H., and Wilbur, W. J. 2000. Finding themes in medline documents: Probabilistic similarity search. In *Advances in Digital Libraries*, 183–192.

Shneiderman, B. 1997. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., third edition.

Shortliffe, E. H. 1981. Consultation systems for physicians: The role of artificial intelligence techniques. In Webber, B. L., and Nilsson, N. J., eds., *Readings in Artificial Intelligence.*, 323–333. Tioga Publishing Company.

Siegel, S., and Castellan, N. J., eds. 1988. *Nonparametric statistics for the behavioral sciences*. New York, NY, USA: McGraw-Hill, second edition.

Singer, M. 2003. Processes of question answering. In Rickheit, G.; Deutsch, W.; and Hermann, T., eds., *Psycholinguistik / Psycholinguistics*, 422–431. Walter de Gruyter, Inc.

Slawson, D. C., and Shaughnessy, A. F. 2000. Becoming an information master: Using poems to change practice with confidence. patient-oriented evidence that matters. *Family Practice Management* 11(5):47–52.

Smith, R., and Chalmers, I. 2001. Britain‘s gift: a "medline" of synthesised evidence. *BMJ* 323:1437–1438.

Smith, R. 1996. What clinical information do doctors need? *BMJ* 313:1062–1068.

Sneiderman, C.; Demner-Fushman, D.; Fiszman, M.; and Rindflesch, T. C. 2005. Semantic characteristics of MEDLINE citations useful for therapeutic decision-making. In *Proceeding of the 2005 Annual Symposium of the American Medical Informatics Association (AMIA 2005)*.

Soubbotin, M. M. 2001. Patterns of potential answer expressions as clues to the right answers. In *Text REtrieval Conference*.

Srinivasan, P. 1996. Query expansion and medline. *Information Processing and Management* 32(4):431–443.

Stinson, R. E., and Mueller, D. A. 1980. Survey of health professionals‘ information habits and needs: conducted through personal interviews. *JAMA* 3910:140–143.

Strasser, T. C. 1978. The information needs of practicing physicians in northeastern new york state. *Bulletin of the Medical Library Association* 66:200–209.

Straus, S. E., and Sackett, D. L. 1998. Using research findings in clinical practice. *BMJ* 317(7154):339–342.

Strzalkowski, T.; Small, S.; Hardy, H.; Yamrom, B.; Liu, T.; Kantor, P.; Ng, K.; and Wacholder, N. 2005. Hitiqa: A question answering analytical tool. In *Proceedings of International Conference On Intelligence Analysis*.

Tbahriti, I.; Chichester, C.; Lisacek, F.; and Ruch, P. 2004. Using argumentation to retrieve articles with similar citations from medline. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA COLING 2004)*, 8–14.

Timpka, T.; Ekström, M.; and Bjurulf, P. 1989. Information needs and information seeking behaviour in primary health care. *Scandinavian Journal of Primary Health Care* 7(2):105–109.

Ting, K. M., and Witten, I. H. 1999. Issues in stacked generalization. *Journal of Artificial Intelligence Research* 10:271–289.

Tsur, O.; de Rijke, M.; and Sima'an, K. 2004. BioGrapher: Biography questions as a restricted domain question answering task. In *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*.

van Kesteren, F., and Kraaij, W. 2006. Measuring the quality of multi-document cluster headlines. In *Proceedings of the IIIA 2006 workshop*.

van Rijsbergen, C. J. 1979. *Information Retrieval*. Department of Computer Science, University of Glasgow, second edition.

Vargas-Vera, M., and Motta, E. 2004. Aqua - ontology-based question answering system. In Monroy, R.; Arroyo-Figueroa, G.; Sucar, L. E.; and Azuela, J. H. S., eds., *MICAI 2004: Third Mexican International Conference on Artificial Intelligence, Mexico City, Mexico, April 26-30, 2004, Proceedings*, volume 2972 of *Lecture Notes in Computer Science*. Springer.

Villanueva, E. V.; Burrows, E. A.; Fennessy, P. A.; Rajendran, M.; and Anderson, J. N. 2001. Improving question formulation for use in evidence appraisal in a tertiary care setting: a randomised controlled trial. *BMC Medical Informatics and Decision Making* 1(4).

Voorhees, E. M., and Tice, D. M. 1999. The TREC-8 question answering track evaluation. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*.

Voorhees, E. M. 2003. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*.

Voorhees, E. M. 2004. Overview of the trec 2004 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2004)*.

Ward, D.; Meadows, S. E.; and Nashelsky, J. E. 2005. The role of expert searching in the family physicians' inquiries network (fpin). *Journal of the Medical Library Association* 93(1):88–96.

Wilczynski, N.; McKibbon, K. A.; and Haynes, R. B. 2001. Enhancing retrieval of best evidence for health care from bibliographic databases: Calibration of the hand search of the literature. In *Proceedings of 10th World Congress on Medical Informatics (MEDINFO 2001)*, 390–393.

Wilkinson, M. A. 2001. Information sources used by lawyers in problem-solving: An empirical exploration. *Library and Information Science Research* 23(3):257–276.

Yang, Y., and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, 412–420.

Zakowski, L.; Seibert, C.; and VanEyck, S. 2004. Evidence-based medicine: Answering questions of diagnosis. *Clinical Medicine & Research* 2(1):63–69.

Zhang, J.; Ga, J.; Zhou, M.; and Wang, J. 2001. Improving the effectiveness of information retrieval with clustering and fusion. *Computational Linguistics and Chinese Language Processing* 6(1):109–125.

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

Zieman, Y. L., and Bleich, H. L. 1997. Conceptual mapping of users queries to medical subject headings. In *Proceeding of the 1997 Annual Symposium of the American Medical Informatics Association (AMIA 1997)*, 519–522.

Zou, Q.; Chu, W. W.; Morioka, C.; Leazer, G. H.; and Kangarloo, H. 2003. Indexfinder: A method of extracting key concepts from clinical texts for indexing. In *Proceeding of the 2003 Annual Symposium of the American Medical Informatics Association (AMIA 2003)*, 763–767.