THE UNIVERSITY *of* EDINBURGH

Automated Question Answering for Clinical Comparison Questions

Annette C. Leonhard

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2012

Abstract

This thesis describes the development and evaluation of new automated Question Answering (QA) methods tailored to clinical comparison questions that give clinicians a rank-ordered list of MEDLINE® abstracts targeted to natural language clinical drug comparison questions (e.g. ”Have any studies directly compared the effects of Pioglitazone and Rosiglitazone on the liver?”).

Three corpora were created to develop and evaluate a new QA system for clinical comparison questions called RetroRank. RetroRank takes the clinician's plain text question as input, processes it and outputs a rank-ordered list of potential answer candidates, i.e. MEDLINE® abstracts, that is reordered using new post-retrieval ranking strategies to ensure the most topically-relevant abstracts are displayed as high in the result set as possible.

RetroRank achieves a significant improvement over the PubMed recency baseline and performs equal to or better than previous approaches to post-retrieval ranking relying on query frames and annotated data such as the approach by Demner-Fushman and Lin (2007).

The performance of RetroRank shows that it is possible to successfully use natural language input and a fully automated approach to obtain answers to clinical drug comparison questions. This thesis also introduces two new evaluation corpora of clinical comparison questions with “gold standard” references that are freely available and are a valuable resource for future research in medical QA.

Acknowledgements

I would like to thank my supervisor Bonnie Webber for her support and guidance throughout my PhD. Her input has been invaluable. I would also like to thank my co-supervisor Claudia Pagliari for providing insight into the medical domain.

I am also grateful to Claire Grover for her advice on adapting LT-TTT2 to clinical questions.

I thank the School of Informatics and my friends and colleagues for providing an inspiring research and social environment. Special thanks go to my friends Sharon Givon and Silke Scheible and my office mates Sarah Luger and Philipp Petrenz for their friendship, support and inspiration.

On a personal level, I would like to thank my husband Donald MacDonald for being there for me when it mattered most and for his infinite patience and support. I would also like to thank my parents Martin and Marie-Luise and my grandmother Brunhilde for their encouragement and support throughout my life.

In conclusion, I recognize that this research would not have been possible without the financial assistance from the MRC and ESRC and I express my gratitude to both agencies.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Annette C. Leonhard*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Clinicians wishing to practice evidence-based medicine need to keep up with a vast amount of ever changing research to be able to use the current best evidence in individual patient care (Sackett et al., 1996). This can be difficult for time-pressed clinicians, although methods such as systematic reviews, evidence summaries and clinical guidelines can help to translate research into practice. Computer technology, in the form of clinical search engines or electronic clinical decision support systems, can be used to facilitate the retrieval and presentation of clinical evidence, but there are still limits concerning its usability and accessibility when timely guidance is of the essence. In a survey commissioned by Doctors.net.uk, 97% of doctors and nurses said that they would find a Question Answering (QA) Service useful, where they can ask questions in their own words (Bryant and Ringrose, 2005). Studies have also shown that clinicians often want answers to particular questions, rather than getting information on broad topics (Chambliss and Conley, 1996; Ely et al., 1999, 2005). Clinicians commonly want to know how one thing compares with another. In the initial corpus of clinical questions collected from the National Library of Health (NLH) Question Answering Service (`http://www.clinicalanswers.nhs.uk`), a manual QA service with the aim of providing answers in a clinically relevant time frame using the best available evidence, approximately 17% of the 4580 questions in their repository on the 20th of July 2007 concerned comparisons of different drugs, treatment methods or interventions as in (1.1).

(1.1) Have any studies directly compared the effects of Pioglitazone and Rosiglitazone on the liver?

Answering such comparison questions automatically presents an interesting challenge as it is not enough to simply search for the drug names or interventions in the question. First, relevance also demands a basis of the comparison in regards to which the drugs or interventions are compared. Second, a post-retrieval ranking step is needed to find out which of the retrieved abstracts are the most relevant for providing a concise answer to a clinician.

Despite the frequency of comparison questions one cannot yet call upon a method especially designed to answer them, as one can for non-comparison clinical queries. Existing general medical search engines such as *askMEDLINE* (Fontelo et al., 2005) or Entrez PubMED do not perform well on the retrieval of relevant abstracts for clinical comparison questions (cf. Section 6.3). This shows that a different approach is needed.

Manual QA services can only handle a limited number of questions at one time and there is no real-time response. It is also too expensive to run manual clinical QA services on a large scale. Unlike manual QA services such as the now defunct NLH QA service, automated QA methods have the potential advantages of cost-effective, real-time answers and no limit on the number of questions that can be asked and answered.

The aim of this PhD project is to develop and test new automated QA methods tailored to clinical comparison questions that give clinicians a rank-ordered list of MEDLINE® abstracts targeted to clinical questions framed in natural language. Three corpora were created to develop and evaluate a new QA system for clinical comparison questions called RetroRank. RetroRank takes the clinician's plain text question as input, processes it and outputs a rank-ordered list of potential answer candidates, i.e. MEDLINE® abstracts. The rank-ordered list is reordered using several post-retrieval ranking strategies to ensure the most topically-relevant abstracts are displayed as high in the ranking as possible. While it would be possible to generate answers in form of extractive summaries from the top-ranked results, displaying a rank-ordered list of abstracts rather than answers consisting of extractive summaries will give clinicians the flexibility to make an informed decision based on their medical knowledge and experience.

The main contribution of this thesis is a new automated QA system, RetroRank, for natural language queries, which is tailored to clinical comparison questions and implements new post-retrieval ranking strategies. RetroRank achieves a significant improvement over the PubMed baseline and performs equal to or better than previous approaches to post-retrieval ranking relying on query frames and annotated MEDLINE® data such as the approach by Demner-Fushman and Lin (2007). In addition, two new

evaluation corpora of clinical comparison questions with "gold standard" references were created that are freely available for use in future research in medical QA.

The structure of the thesis is as follows:

Chapter 2 provides background knowledge about the information needs of clinicians and the domain of Evidence Based Medicine (EBM) as well as an overview of Information Retrieval (IR) and Question Answering (QA) techniques and existing clinical QA services and applications that provide a background for the current research.

Chapter 3 gives an overview of the characteristics of clinical comparison questions and shows the types of different comparative constructions that appear in clinical questions. Also the lexical items indicative of comparison questions are introduced, which will be used for the creation of Corpus 1 (Section 4.1).

Chapter 4 describes the creation and preprocessing of the three corpora of clinical questions. Corpus 1 (NLH QAS) is used for the manual exploration of the best strategy for abstract retrieval in Chapter 5 and for the system development of the automated abstract retrieval component of RetroRank in Chapter 6. Corpus 2 (Essential Evidence Plus POEMs) is used for the evaluation of the retrieval component in Chapter 6 according to the same criteria used for the evaluation of *ask*Medline by Fontelo et al. (2005) and for developing the post-retrieval ranking strategies described in Chapter 7. Because Corpus 2 only has one "gold-standard" reference for each clinical question, it cannot be used for calculating standard information retrieval (IR) metrics such as Mean Average Precision (MAP). To evaluate the final version of RetroRank with standard IR metrics, Corpus 3 (Cochrane Systematic Reviews) was collected, which provides a number of "gold standard" references for each question.

Chapter 5 presents the initial experiment on manual query construction, abstract retrieval and evaluation. The aim of this chapter is to describe different search strategies for clinical comparison questions that were tried to determine the best one and to evaluate it using human judges. The findings in this chapter were used to develop the automatic QA system RetroRank introduced in Chapter 6.

Chapter 6 describes the implementation and evaluation of the query construction and abstract retrieval component of RetroRank. The retrieval component of RetroRank was developed on Corpus 1 (Section 4.1) and evaluated in terms of retrieval accuracy using Corpus 2 (Section 4.2).

Chapter 7 introduces different post-retrieval ranking strategies which were used to rerank the MEDLINE® abstracts retrieved for each question in Corpus 2 and Corpus 3. The goal is to display the most relevant abstracts as high in the result set as possible and to outperform the recency ordering performed by PubMed. Because recency does not equate to relevance, post-retrieval ranking is an important feature in a system that is geared towards providing the most relevant abstracts at the top of the result list to enable clinicians to find the most relevant information in a quick and reliable way.

Chapter 8 describes the implementation and evaluation of the post-retrieval ranking module of RetroRank. The post-retrieval ranking component was developed and tested on Corpus 2 (Section 4.2) and fully evaluated on Corpus 3 (Section 4.3). The evaluation shows how the different post-retrieval ranking strategies perform on the different corpora and in comparison to the post-retrieval ranking system developed in Demner-Fushman and Lin (2007). It is shown that the automatic *ISI-citation* based strategies and the *Expert Voting* strategy are a significant improvement over the PubMed recency baseline and are a strong contender to the strategies based on query frames and annotated data developed by Demner-Fushman and Lin (2007).

Chapter 9 gives a summary and conclusion of the work described in this thesis and an outlook on future work.

Chapter 2

Background

Because of the interdisciplinary nature of this research the disciplines of Natural Language Processing and Clinical Medicine need to be described, but a comprehensive overview of each is beyond the scope of this work. To provide a foundation and motivation for the research undertaken in this thesis, the goal of this chapter is to provide background knowledge about the information needs of clinicians and the domain of Evidence Based Medicine (EBM) as well as an overview of Information Retrieval (IR) and Question Answering (QA) techniques and existing clinical QA services and applications.

2.1 Information Needs of Clinicians

The information needs of clinicians have been a topic of research for decades. One of the earliest works is the study of Covell et al. (1985) that researched the information needs of doctors during a half day, or four hours, of typical office practice. His findings are similar to the findings of other studies reported in this section. Since then a variety of methodologies such as interviews, self reports and observation have been used to determine the information-seeking behaviour of clinicians. Studies mainly focus on the type of information need, the number of questions arising while tending to patients, the preferred source type for information, i.e., printed resources, electronic resources, advice from colleagues and the percentage of questions pursued and answered (Demner-Fushman, 2006; Davies, 2011).

Health professionals have different types of information needs. A review by Smith (1996) identified the following six different categories of information needs:

- 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.
- Information on scientific, political, legal, social, management, and ethical changes that will affect both how medicine is practised in a society and how doctors will interact with individual patients.

In this research the need for medical knowledge will be addressed. Medical knowledge can be obtained from a number of sources ranging from textbooks to electronic databases such as MEDLINE®, and the challenge lies in finding the relevant information and applying it to the individual patient (Smith, 1996). According to the comprehensive survey by Davies (2007) the top categories for medical knowledge are knowledge about treatment or therapy (average 38%), diagnostic methods (average 24%) and drug therapy or about drugs themselves (average 11%).

Medical knowledge falls into two categories, namely background knowledge and foreground knowledge (Richardson and Wilson, 1997). The term "background knowledge" refers to general knowledge on a condition or disease and leads to wh-type questions such as *What are the symptoms of liver failure?*, *What treatment options exist for joint pain?* or *What causes fever and rash?* Background knowledge questions can be answered well from textbooks or systematic reviews.

Foreground knowledge is directly related to a patient and concerns issues such as diagnosis, treatment and prognosis, e.g., the choice of therapeutic interventions or determining the best test for a condition (Davies, 2011; Demner-Fushman, 2006). These kinds of question can also be answered by secondary sources such as up-to-date systematic reviews if the question is frequent enough to warrant one. If the question does not address a common enough problem, electronic sources such as MEDLINE® that index clinical trials and observations might provide an answer. The average number of foreground knowledge questions per patient is 0.24 according to a recent survey using clinical librarians in the UK as data collectors during clinical meetings (Davies, 2009). This number is similar to the number reported by Ely et al. (1999) which was 0.32 questions per patient considering only foreground questions. Clinical librarians believe the real number of questions is higher, though, because there are more questions asked via email and on the phone than during the clinical meetings (Davies, 2009).

Printed resources such as drug handbooks or medical textbooks are very popular in the medical community and often still preferred over electronic resources. When physicians were asked to rank the aids they use in clinical decision making, a survey from 2007 reports that text sources were ranked first, humans, i.e., colleagues, were ranked second, and electronic resources were ranked third (Davies, 2007). A new survey from 2011 shows that US and Canadian clinicians ranked electronic resources first, while clinicians in the UK still rank them third and prefer asking colleagues or consulting printed resources (Davies, 2011). However, printed resources are not the best choice where up-to-date information is concerned because there is a time delay between editing and publication of at least six months (Davies, 2007; Ebell, 2009). Also, the amount of time it takes to read through a wealth of printed resources to find the relevant information is generally not compatible with a clinician's busy schedule.

The electronic resources clinicians most frequently use are MEDLINE® (81.4% in the US and 76.5% in the UK) and Cochrane Systematic Reviews (70.2% in the US and 74.5% in the UK). MEDLINE® contains unfiltered information (Grandage et al., 2002), whereas the Cochrane Library provides filtered evidence-based systematic reviews (Davies, 2011). Previous studies found that between 71% and 88% of clinical questions are appropriate for MEDLINE® and that for approximately 50% of the questions information can be found which clinicians deem relevant (Demner-Fushman, 2006). Evidence Based Medicine (EBM) resources such as the Cochrane Library were found to answer only about 20% of more complex clinical questions that involve more than one concept such as a drug dose (Davies, 2007).

There is a reluctance towards using electronic resources because they present multiple problems for clinicians. Converting a clinical question into a searchable strategy can be challenging and the use of inappropriate search terms, spelling errors, wrong connectors or drug brand names rather than generic names leads to the retrieval of incomplete or non-useful information (Verhoeven et al., 1995; Davies, 2007). Ely et al. (2002) summarises the obstacles clinicians face when searching for evidence-based answers related to patient care as "inadequate time to search for information, failure of the resource to address the topic, and inadequate synthesis of multiple bits of evidence into a clinically useful statement" (Ely et al., 2002). Because of the problems encountered "doctors often decided not to pursue their questions because they doubted the existence of useful information in available resources" (Ely et al., 2002). Gorman and Helfand (1995) found that clinicians pursued less than a third of their questions. A study by Ely et al. (2005) found that answers to 55% of questions were pursued by

clinicians.

There is a general consensus that the most useful information is relevant, valid, and easy to access and apply (Slawson et al., 1994; Smith, 2002). As Ely et al. (2002) states "Practising doctors do not have time to search multiple sites or scroll through long text. Nor do they have time to search multiple textbooks or perform literature searches for most of their questions. They need to pick the right resource the first time, the information in that resource needs to be readily found, and all the information must be there. Although it remains to be shown, we believe that systems designed to overcome the obstacles we identified will improve the asking and answering of questions and potentially patient outcomes."

The need for a service or system to aid in the question answering process is confirmed in a survey commissioned by Doctors.net.uk, in which 97% of doctors and nurses said that they would find a Question Answering (QA) service that allows them to ask questions in their own words useful (Bryant and Ringrose, 2005). This research is a step towards an automated way of fulfilling that need and increasing the usefulness of MEDLINE®.

2.2 Evidence Based Medicine

Evidence Based Medicine (EBM) is "the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett et al., 2000). Practising EBM involves integrating one's individual clinical expertise with the best available clinical evidence gained from systematic research to decide upon the best treatment for a patient. The best clinical evidence is up-to-date information from EBM sources such as systematic reviews which are overviews of synthesized primary research for particular clinical research questions.

In order to apply the best evidence while making decisions, the question needs to be clearly defined, the necessary information located and checked for validity and relevance, and the information summarized (Demner-Fushman, 2006). The four main clinical tasks in EBM concern etiology, diagnosis, therapy and prognosis. Richardson et al. (1995) identified four components that are key elements in a clinical question:

(**P**atient/**P**roblem): What is the patient/problem being addressed?
(**I**ntervention): What is the intended intervention?
(**C**omparison): What is the intervention compared to?

(**O**utcome): What are the outcomes?

These components are known as the PICO framework and are important for question and search strategy formulation as well as for assessing clinical information.

Another important component of the EBM model is the quality and strength of the clinical evidence. A grading scale was developed by Ebell et al. (2004) to assess the quality, quantity, and consistency of clinical evidence for outcomes that help patients to live longer and better lives, e.g., improvement of symptoms, better quality of life or reduced mortality. This is known as "patient-oriented evidence". The taxonomy differentiates between the strength of a body of evidence and the quality of individual research studies. According to Ebell et al. (2004) a body of evidence can have one of three grades ranging from A to C with A being the best.

Grade A evidence is good quality, consistent patient-oriented evidence. Grade A evidence usually comes from high quality double- or triple-blind randomized clinical trials or from meta-analysis of controlled studies. Grade B evidence stems from inconsistent or limited-quality studies such as non-randomized controlled studies. Grade C evidence is disease-oriented rather than patient-oriented evidence that is mainly based on consensus, usual practice or opinion.

In addition, three levels of evidence in individual studies were defined. Level 1 denotes good-quality patient-oriented evidence which can be found in systematic reviews, meta-analysis and high quality randomized controlled trials and cohort studies. Level 2 denotes limited-quality patient-oriented evidence from less stringent clinical trials and cohort studies, and Level 3 consists of other evidence from Grade C resources (Ebell et al., 2004; Demner-Fushman, 2006).

2.3 MEDLINE®, PubMed and Entrez PubMed

2.3.1 MEDLINE®

MEDLINE® is the US National Library of Medicine's (NLM) database of life sciences and biomedical information. It currently contains approximately 18 million citations and abstracts from approximately 5,500 biomedical journals worldwide since 1950 and is viewed as the authoritative source of clinical evidence by clinicians, biomedical researchers, and other professionals in the field. Nearly 70,000 new citations were added in 2010 [1]. Its status as the authoritative source for clinical evidence makes it an

[1] `http://www.nlm.nih.gov/pubs/factsheets/medline.html`

ideal source for abstracts suitable for answering clinical questions.

Each MEDLINE® citation includes basic information about the citation such as the article title, the author name(s), the publication name and type, the publication date of the article, the date it was indexed in MEDLINE®, the publication language and the abstract of the indexed paper. About 88% of all abstracts are in English. These abstracts serve as potential answer candidates for the RetroRank system.

Each MEDLINE® citation is also indexed with Medical Subject Headings[2] (MeSH) from NLM's controlled vocabulary thesaurus. It contains approximately 26,000 descriptors stored in a hierarchical structure so that it can be searched at different levels of detail. There are also over 177,000 entry terms that help in locating the most appropriate MeSH Heading, e.g., "Acetylsalicylic Acid" is an entry term for "Aspirin". MeSH terms are assigned by over 100 indexers with degrees in the life sciences and extensive training by the NLM. The MeSH vocabulary is continually updated and revised to reflect the most accurate information possible. An example for the use of MeSH terms is given in Section 5.1.

MEDLINE® can be searched via PubMed, the NLM's gateway, or other third party search engines.

2.3.2 PubMed

PubMed[3] is a boolean search engine that allows users to search the abstract text and metadata fields such as MeSH terms. PubMed can be searched using MeSH terms, author names, title words, text words or phrases, journal names or any combination of these. It also provides "Clinical Queries" search filter templates for narrowing down a query to studies based on etiology, diagnosis, prognosis, or treatment of a particular disease (Haynes et al., 1994). These templates are fixed boolean query fragments such as MeSH term restrictions that are appended to the user's query. To use the "Clinical Query" templates, the search keywords are entered on the PubMed Clinical Queries website `http://www.ncbi.nlm.nih.gov/pubmed/clinical/` and the appropritate filter and the scope of the query (sensitive/broad or specific/narrow) are chosen. Example (2.1) shows the Clinical Queries query translation for the therapy question"What are the effects of using beta interferon in the treatment of multiple sclerosis?" using the "Therapy" and "Narrow" filters.

[2] `http://www.nlm.nih.gov/pubs/factsheets/mesh.html`
[3] `http://www.nlm.nih.gov/pubs/factsheets/pubmed.html`

(2.1) (multiple sclerosis beta interferon) AND (Therapy/Narrow[filter])

Therapy/Narrow corresponds to the PubMed equivalent:
(randomized controlled trial[Publication Type] OR (randomized[Title/Abstract] AND controlled[Title/Abstract] AND trial[Title/Abstract])).

2.3.3 Entrez PubMed

Entrez PubMed[4] is a text-based search and retrieval system for PubMed®. Entrez PubMed provides access to MEDLINE® and other bio(medical) databases. It was developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) which is part of the U.S. National Institutes of Health (NIH) [5]. While it is not a QA system, it can be used to answer questions, which are simply treated as text strings. Figure 2.1 shows the search interface and an extract of the answers for the question:

(2.2) How safe and effective are aspirin and warfarin therapy in the prevention of stroke in patients with atrial fibrillation?

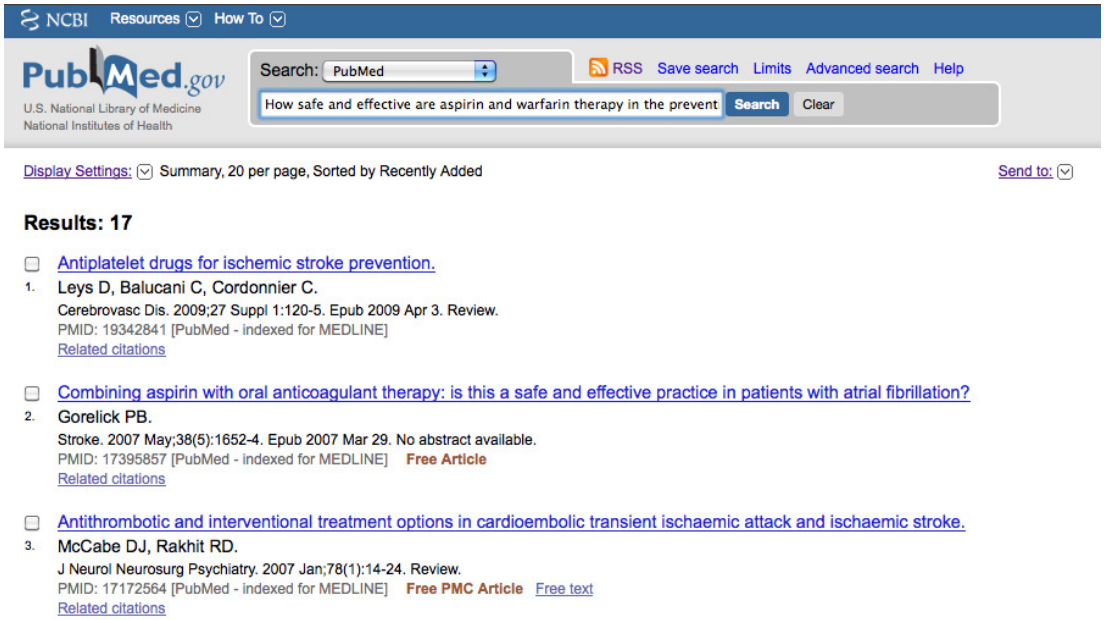


Figure 2.1: Entrez PubMed interface

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

The results page of Entrez PubMed also offers a search for related articles and shows the query translation from the free text question to PubMed® search terms. The query translation for Example (2.2) is shown in Example (2.3), which illustrates the standard way PubMed® performs query translations.

(2.3)

safe[All Fields] AND effective[All Fields] AND (“aspirin”[MeSH Terms] OR “aspirin”[All Fields]) AND (“warfarin”[MeSH Terms] OR “warfarin”[All Fields]) AND (“therapy”[Subheading] OR “therapy”[All Fields] OR “therapeutics”[MeSH Terms] OR “therapeutics”[All Fields]) AND (“prevention and control”[Subheading] OR (“prevention”[All Fields] AND “control”[All Fields]) OR “prevention and control”[All Fields] OR “prevention”[All Fields]) AND (“stroke”[MeSH Terms] OR “stroke”[All Fields]) AND (“patients”[MeSH Terms] OR “patients”[All Fields]) AND (“atrial fibrillation”[MeSH Terms] OR (“atrial”[All Fields] AND “fibrillation”[All Fields]) OR “atrial fibrillation”[All Fields])

Entrez PubMed searches for all words in an input text string except for those found on the PubMed stop word list that includes question words such as “how”, “when”, “which”, etc. (see Appendix C). This approach leads to low accuracy as will be discussed in the evaluation of *ask*MEDLINE against Entrez PubMED (Section 2.3.3) and in the evaluation of Entrez PubMed on clinical comparison questions (Section 6.3).

The NCBI also offers a set of Entrez programming tools called E-Utilities[6]. The two main components of the E-Utilities tools are ESearch, which searches and retrieves primary PubMed IDs (PMIDs) and term translations and EFetch, which produces abstracts from Entrez PubMed. A full list of E-Utilities tools is available at: `http://eutils.ncbi.nlm.nih.gov`. The EUtilities are used by *ask*MEDLINE (Fontelo et al., 2005) and in the retrieval component of RetroRank described in Chapter 6.

2.4 Clinicial Information Retrieval and Question Answering

There are several services for and approaches to clinical question answering as well as research on identifying and extracting comparative constructions, but so far no methods have been developed for automatically answering clinical comparison questions. This section presents work on comparative sentence mining by Jindal and Liu (2006a,b), work by Fiszman et al. (2007) concerning the interpretation of comparative structures

[6] `http://eutils.ncbi.nlm.nih.gov/`

in biomedical texts, and the work on clinical question answering by Demner-Fushman and Lin (2005, 2006, 2007) and Demner-Fushman (2006). It also introduces existing clinical question answering services and applications to give a background for my research.

2.4.1 Comparative Sentence Mining

In 2006, Jindal and Liu published two papers describing the study of *comparative sentence mining*. By "comparative sentence" Jindal and Liu denote sentences that express "an ordering relation between two sets of entities with respect to some common features" (2006b). Jindal and Liu focus on evaluative texts giving subjective opinions from the internet such as customer reviews, forum discussions and news articles. They define comparative sentence mining as a two-step task. The first step is to identify comparative sentences in the text. The second step is to extract comparative relations from the sentences that have been identified. The term *comparative relation* is defined as follows (Jindal and Liu, 2006b):

> A *comparative relation* captures the essence of a comparative sentence and is represented with the following:
>
> (relationWord, features, entityS1, entityS2, type)
>
> where *relationWord*: The keyword used to express a comparative relation in a sentence.
>
> *features*: a set of features being compared.
>
> *entityS1* and *entityS2*: Sets of entities being compared. Entities in entityS1 appear to the left of the relation word and entities in entityS2 appear to the right of the relation word.
>
> type: *non-equal gradable, equative or superlative*.

An *entity* is defined as the name of a person, a product brand, a company, a location, or similar, which is compared in a comparative sentence, and a *feature* is defined as a part or property of the compared entity. Entities and features can only be nouns or pronouns. This means that for a sentence like *"Canon's optics are better than those of Sony and Nikon"*, the system is expected to extract the following relation:

(better, {optics}, {Canon}, {Sony, Nikon})

The *relationWord* is *better*, the feature in this example is optics and the *entities* being compared are Canon optics compared to Sony's and Nikon's.

Jindal and Liu distinguish between four different types of comparison: *non-equal gradable*, *equative*, *superlative*, and *non-gradable*, with the following definitions Jindal and Liu (2006b):

1. *Non-equal Gradable*: Relations of the type **greater** or **less than** that express a total ordering of some entities with regard to certain features. This type also includes user preferences.

2. *Equative*: Relations of the type **equal to** that state two entities are equal with respect to some features.

3. *Superlative*: Relations of the type **greater** or **less than all others** that rank one entity over all others.

4. *Non-Gradable*: Sentences which compare features of two or more entities, but do not explicitly grade them.

Jindal and Liu only focus on the first three types, which they call *gradable comparatives*, because they express an explicit ordering between the comparison entities (2006b). These correspond to scalable adjectives introduced in Section 3.1.

Jindal and Liu identify comparative sentences by using an approach which involves *class sequential rules* (CSR) and *naïve Bayesian classification* (Jindal and Liu, 2006a). For extracting the comparative relations described above, they propose a type of rules called *label sequential rules* (LSR) of the form X $\Rightarrow$ Y. Both X and Y are sequences where X is a sequence produced from Y by replacing some of its items with wildcards which can match any item (Jindal and Liu, 2006b). The rules are based on POS tags and a small set of additional keywords generated by training on hand-labelled data. All nouns and pronouns in comparative sentences are marked up as belonging to one of the categories *entityS1*, *entityS2*, *feature*, and *non-entity feature* (i.e. nouns or pronouns that do not refer to features or entities).

The data used by Jindal and Liu (2006b) was labelled by two annotators. Out of a total of 3248 sentences, 285 were labelled as *non-equal gradable*, 110 as *equative*, and 169 as *superlative*. The other sentences did not contain comparisons. In total, the annotators labelled 488 instances of *entityS1*, 300 instances of *entityS2*, and 348 features. The overall F-score for the extraction task is 72%, which shows a considerable

improvement to the F-score of 58% achieved by the baseline CRF system developed by Sarawagi (2004). Jindal and Liu (2006b) claim that LSR achieved around 80% F-score for *entityS1*, about 70% for *entityS2*, and around 60% for *features*. The *relationWords* were not extracted by rules and therefore the extraction results are not reported.

Jindal and Liu's work (2006b) is not in the clinical domain and focuses only on comparative sentences, not on comparative questions but their method could be applied to questions as well. There are problems with their approach to extracting comparative entities, though. Jindal and Liu's method only extracts both entities correctly when they appear on the left and right of a *relationWord* instead of determining the entities by a semantic approach. *Entity1* and *entity2* are also only determined by their relative position in a sentence assuming the argument to the left (*entity1*) is greater than the argument to the right (*entity2*) as it would be in a mathematical "greater than" relation. This is problematic, however, because sentences in natural language do not adhere to a strict ordering but can have a variety of forms. Just because an *entity* is mentioned first, it does not have to be the one that is higher on a comparative scale. The same is true for clinical comparison questions. Jindal and Liu's approach could be applied to a question like Example (2.4).

(2.4) Is Ibuprofen better than aspirin for treating a headache?

In which case the following could be extracted:

$$(\text{better}, \{\text{headache}\}, \{\text{Ibuprofen}\}, \{\text{Aspirin}\})$$

However, for a question like Example (2.5), the proposed method would yield an incorrect result, because the left side of the *relationWord* is empty and therefore Ibuprofen would not be recognised as *entity1*.

(2.5) What is better for treating a headache: Ibuprofen or Aspirin?

A similar problem arises in Example (2.6):

(2.6) Is Ibuprofen or Aspirin better for treating a headache?

In this case both Ibuprofen and Aspirin are on the left side and the right side of the *relationWord* is empty, meaning that Aspirin would not be recognised as *entity2*.

Another problem with Jindal's and Liu's approach (2006a) is that Jindal and Liu only look for adjectives or adverbs as *relationWords*, i.e. *better*, *faster*, *quicker*, but

comparisons can also be expressed by comparative cues such as *compare* as illustrated in Example (2.7) and introduced in Section 3.2.

(2.7) How does Ibuprofen compare to Aspirin for treating headaches?

Because *compare* is not one of the *relationWords* that is searched for, neither drug would be recognized as an *entity*.

2.4.2 Interpretation of Comparative Structures

(Fiszman et al., 2007) describes work on automatically interpreting comparative constructions in MEDLINE® abstracts. They use an extension of an existing semantic processor, SemRep (Rindflesch and Fiszman, 2003; Rindflesch et al., 2005), from the Unified Medical Language System resources to construct semantic predications for the extracted comparative expressions. In this paper, Fiszman et al. (2007) concentrate on extracting two different "comparative structures in which two drugs are compared with respect to a shared attribute", which frequently occur in reports on clinical trials for drug therapies. A shared attribute is, for example, a drug's efficacy in treating a certain condition. The drugs' relative merits in achieving their purpose is expressed by positions on a scale which is denoted by adjectives or nouns. The compared terms are expressed by co-joined noun phrases and their shared characteristic is a predicate outside of the comparative structure. Words like *than*, *as*, *with* and *to* are cues for identifying compared terms, the comparison scale and the relative position of the compared entities on the scale.

The first comparative structure (*comp1*) identified by Fiszman et al. (2007) includes a form of the word *compare* and is a comparison between a primary therapy and a secondary therapy. The second structure (*comp2*) compares the relative merits of a primary and secondary therapy using scalable adjectives. The first step is to generate a semantic interpretation of these comparative structures and the second step is to identify the elements of the comparative structures based on their semantic predications.

Fiszman et al.'s *comp1* structures can be identified by a form of the word *compare*. They compare two terms mostly without ranking them on a scale. If a scale is mentioned, it is indicated by a noun such as efficacy. However, Fiszman et al. do not extract a scale for *comp1* structures because the position on a scale is never mentioned. Example (2.8) illustrates a *comp1* structure.

(2.8) To compare **misoprostol** with **dinoprostone** for cervical ripening and labor induction. (Example (3) in (Fiszman et al., 2007))

Comp2 structures are more complex, comparing two terms using scalable adjectives which indicate the ranking on a scale. Example (2.9) illustrates both terms at an equal position on the scale:

(2.9) **Azithromycin** is as effective as **erythromycin estolate** for the treatment of pertussis in children. (Example (5) in (Fiszman et al., 2007))

Comp2 structures express equality or inequality. For structures expressing inequality a distinction can be made between the expression of superiority, where the primary term is ranked higher than the secondary term (Example (2.10)), and inferiority, where the primary term is ranked lower than the secondary term (Example (2.11)).

(2.10) **Naproxen** is safer than **aspirin** in the treatment of the arthritis of rheumatic fever. (Example (6) in (Fiszman et al., 2007))

(2.11) **Sodium valproate** was significantly less effective than **prochlorperazine** in reducing pain or nausea. (Example (7) in (Fiszman et al., 2007))

To develop the comparative processing, Fiszman et al. extracted sentences from 10,000 MEDLINE® citations reporting the results of clinical trials. From these sentences, the most frequent patterns were extracted and used to enhance SemRep argument identification. These patterns only list the obligatory components but allow modifiers and qualifiers. Here is an example pattern for *comp1*:

C2: *compare* Term1 with/to Term2 (Fiszman et al., 2007)

A full list of these patterns can be found in Appendix B.

SemRep finds underspecified semantic propositions in biomedical text based on a syntactic analysis and domain knowledge from the UMLS including the Metathesaurus and the Semantic Network [7]. Fiszman et al. focus on the group Chemicals & Drugs from the Semantic Network. SemRep assumes a *comp1* structure when it encounters a form of the word *compare* and looks to the noun phrase on the right preceded by *with*, *to*, *and*, or *versus*, which serve as cues for indicating the compared terms. If the head has a corresponding concept having a semantic type in the Chemical & Drugs

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

group, it is identified as Term2. Then the algorithm looks to the left of the identified term to find a noun phrase that also has a semantic type in the same group and if such a noun phrase is found it identifies it as Term1. The predicate for *comp1* structures is COMPARED_WITH. Example (2.13) shows the processed sentence from Example (2.12):

(2.12) To compare the efficacy and tolerability of Hypericum perforatum with imipramine in patients with mild to moderate depression. (Example (13) in (Fiszman et al., 2007))

(2.13) Hypericum perforatum COMPARED_WITH Imipramine (Example (14) in (Fiszman et al., 2007))

In addition to identifying the compared terms, a scale must be identified in *comp2* structures as well as the positions of the terms on the scale. The algorithm for *comp2* patterns looks for one of the cues such as *than* and maps the heads of the noun phrases to the right and left the same way as for *comp1* patterns. The scale name is found by looking at the secondary compared term and locating the first adjective to its left of which the nominalization according to the SPECIALIST lexicon [8] is used as the name of the scale, e.g., *effective (include arrow) Effectiveness*. The relative positions on the scale are determined by contrasting equality and inequality. The primary compared term is considered to be higher on the scale unless it is preceded by *less* or *is inferior*. The representation for *comp2* structures is shown in Example (2.15) for Example (2.14).

(2.14) **Losartan** *was more effective than* **atenolol** in reducing cardiovascular morbidity and mortality in patients with hyptertension, diabetes, and LVH. (Example (20) in (Fiszman et al., 2007))

(2.15) Losartan COMPARED_WITH Atenolol
Scale: Effectivness
Losartan HIGHER_THAN Atenolol (Example (21) in (Fiszman et al., 2007))

A test set of 300 sentences containing comparative structures were extracted from MEDLINE® abstracts published later than the ones that were used to develop the methodology. The sentences were annotated with their PubMed ID, names of the two

[8] `http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html`

drugs, the scale and their relative position on the scale. After removing duplicates, 287 sentences with 288 comparative structures remained. There are 203 *comp1* structures and 85 *comp2* structures. The overall F-score for SemRep's performance on the test set is 81%. The F-score for drug extraction is also 81%, *comp1* structures are retrieved with 98% precision but only 74% recall, and *comp2* structures with 92% precision and only 62% recall. The results show a high precision for all tasks, but 26% of *comp1* structures and 38% of *comp2* structures remain unidentified. This is mainly due to empty heads and word sense ambiguity.

Like Jindal and Liu's work (2006a; 2006b), Fiszman et al. (2007) do not focus on questions. While their method could also be applied to questions, the same problem mentioned in the discussion of Jindal and Liu (2006b) applies. The algorithm takes its targets by looking at the noun phrases to left and right of a cue indicating a comparison. Therefore, it would also misinterpret questions in which both compared entities occur on the right side of the cue or both entities occur on the left side of the cue as shown in Examples (2.16) and (2.17).

(2.16) What is *better* for stroke prevention: **Aspirin** or **Warfarin**?

(2.17) Is **Aspirin** or **Warfarin** *more efficient* for preventing stroke?

In addition, comparisons that are split across clauses or sentences, such as Examples (2.18) and (2.19) which do not fit the pattern for recognizing *comp1* and *comp2* structures, cause problems for the algorithm:

(2.18) Although Warfarin may be used for preventing stroke, it is not as good as aspirin.

(2.19) Warfarin can be used for preventing stroke. However Aspirin is better.

There is also an essential limitation to SemRep. SemRep's comparative module is based on *scalar* comparative constructions. *Non-scalar* comparisons, e.g., comparisons like "Is x the same intervention as y" or "How does drug x differ from drug y" cannot be extracted using SemRep. This means that a different method is necessary in order to process *non-scalar* comparisons as well as *scalar* comparisons that cannot be recognized because of their structure, e.g., both compared entities are to the right side of the comparative cue.

Problems also occur for "Wh-" or "anything" questions. "Wh-words" or "anything" do not have a type that can be mapped by the SemRep algorithm.

(2.20) What drug is better than X for treating Y?

(2.21) Is there anything better than X for treating Y?

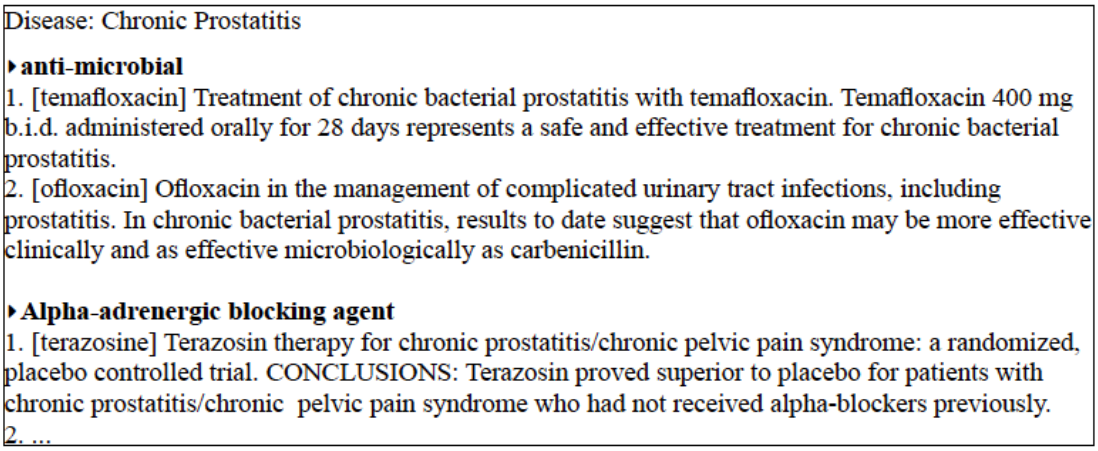
2.4.3 Addressing Clinical Questions

In 2006, Demner-Fushman and Lin published a paper on answering clinical questions of the type "What is the best treatment for *X*" by using a hybrid approach consisting of information retrieval and summarization. This question type was chosen because studies of clinician's behaviour have shown that this class of question frequently occurs in the clinical setting (Ely et al., 1999). My own data confirms this finding (cf. Section 4.1). Demner-Fushman and Lin use answer extraction for identifying short answers, semantic clustering for grouping results, and extractive summaries to generate supporting evidence. The first step is to identify the drugs that are searched for. The second step is to cluster abstracts for these drugs using semantic classes from the UMLS ontology. As a third and final step, a short summary is generated for each abstract, which gives supporting evidence.

Demner-Fushman and Lin's research follows the paradigm of EBM (Sackett et al., 2000). The PICO framework introduced in Section 2.2 underlies the question answering research by Demner-Fushman and Lin (2005, 2006) and Demner-Fushman (2006). Clinical questions as well as MEDLINE® abstracts are translated into the PICO format using specific purpose-designed extractors. There is one for the *Population*, one for the *Problem*, one for the *Intervention/Comparison* and one for the *Outcome*. The translation of abstracts to the PICO format relies on the availability of an annotated corpus of MEDLINE® abstracts (Demner-Fushman and Lin, 2005).

Demner-Fushman and Lin (2006) try to balance clinicians' need for conciseness and completeness, which results from time pressure and the necessity to completely examine all relevant evidence, by giving hierarchical answers which support multiple levels of drill-down. An example for the question "What is the best treatment for chronic prostatitis?" is shown in Figure 2.2. "Best treatment" is used in the sense of "most studied treatment". It lists two drug categories that are relevant to the treatment of the disease with associated clusters of extractive summaries of MEDLINE® abstracts. If needed the full abstract text or even the electronic version of the study can be retrieved.

Demner-Fushman and Lin focus primarily on synthesising correct answers from a set of search results consisting of MEDLINE® citations. Given this set of search


Disease: Chronic Prostatitis

anti-microbial

1. [temafloxacin] Treatment of chronic bacterial prostatitis with temafloxacin. Temafloxacin 400 mg b.i.d. administered orally for 28 days represents a safe and effective treatment for chronic bacterial prostatitis.
2. [ofloxacin] Ofloxacin in the management of complicated urinary tract infections, including prostatitis. In chronic bacterial prostatitis, results to date suggest that ofloxacin may be more effective clinically and as effective microbiologically as carbenicillin.

Alpha-adrenergic blocking agent

1. [terazosine] Terazosin therapy for chronic prostatitis/chronic pelvic pain syndrome: a randomized, placebo controlled trial. CONCLUSIONS: Terazosin proved superior to placebo for patients with chronic prostatitis/chronic pelvic pain syndrome who had not received alpha-blockers previously.
2. ...

Figure 2.2: System response to the question "What is the best treatment for chronic prostatitis?" cited from Demner-Fushman and Lin (2006).

results, answer generation is a three step process.

The first step is answer extraction. Demner-Fushman and Lin created an extractor to identify the drugs, or interventions in EBM terminology, under study which is based on MetaMap (Aronson, 2001), a programme that automatically identifies entities which correspond to UMLS concepts. Drugs mainly fall under the semantic type of PHARMACOLOGICAL SUBSTANCE. All entities with a corresponding UMLS concept are marked as candidates and scored on their position in the abstract, the frequency and other features.

In a second step, the retrieved MEDLINE® abstracts for the identified interventions are grouped into clusters based on the main interventions in the abstracts. This is done with a variant of the hierarchical agglomerative clustering algorithm by Zhao and Karypis (2002). This algorithm uses semantic relationships from the UMLS (Unified Medical Language System) for computing similarities between interventions. In Figure 2.2, *temafloxacin* and *ofloxacin* are in the same cluster because they are both hyponyms of *anti-microbials* according to the UMLS ontology.

During the third step, a short extractive summary is generated for each MEDLINE® abstract in the cluster. It has three elements: The main intervention, the abstract title and the outcome sentence with the highest score. The outcome sentence gives the findings of a clinical study. The "Outcome Extractor" used is based on former work by Demner-Fushman and Lin (2005, 2006).

The system was tested on a set of 30 questions from the June 2004 edition of *Clinical Evidence* (CE), a periodic report from the British Medical Journal (BMJ), which summarises the best know drugs for some dozen diseases. The first author

used PubMed to retrieve the best possible MEDLINE® abstracts by generating manual queries which were formulated to take advantage of MeSH (Medical Subject Heading) terms, a manually assigned controlled vocabulary that encodes a large amount of knowledge about the content of an abstract. Abstracts were limited to "drug therapy" and clinical trials.

Demner-Fushman and Lin evaluate the system in two ways. The first evaluation is a manual, factoid-style evaluation with a focus on short answers. The second evaluation is an automatic one with ROUGE[9] based on the CE abstracts which are taken as reference summaries. The baseline for both evaluations are the main interventions from the first three MEDLINE® abstracts retrieved by the manual PubMed queries (*PubMed*).

The two test conditions for the first evaluation are the three main interventions from the first abstract from the largest clusters (*Cluster*), and the main interventions from the first abstract from three selected by an oracle, the first author (*Oracle*). The oracle condition is the upper-bound, given the results of expert manual querying. For the baseline *PubMed*®, 20% of the drugs were evaluated as beneficial, for the *Cluster* condition the number is 39%, and for the *Oracle* condition 40%. 60% of the PubMed® answers were judge as good in comparison to 83% for the *Cluster* condition and 89% for the *Oracle* condition.

The test conditions for the automatic evaluation consist of the baseline from the first evaluation. The three other conditions are a *cluster round robin* selecting the first abstract by size from the top three clusters, an *oracle cluster order*, selecting three abstracts from the best cluster, and an *oracle round-robin*, selecting the first abstract from each of the top three clusters. The results for the cumulative evaluation after three rounds are 52.3% for the baseline, 52.6% for the *cluster round robin*, 59.7% for the *oracle cluster order* and 58.6% for the *oracle round-robin*.

Both evaluations show Demner-Fushman and Lin's system outperform the *PubMed*® baseline. There a certain limitations however. The system depends on a set of high quality search results, based on manually generated queries. The system also does not perform semantic processing for determining the efficacy of drugs, but can only recognise topics by matching terms to the UMLS ontology and find outcome statements in the retrieved abstracts. The clusters are ordered by size, which means the most commonly discussed drug is selected as the best drug. This assumption is not valid but the authors have observed that drugs that are studied more are more likely to be beneficial.

Selecting the most studied drug as the best drug might work as far as useful an-

[9] `http://berouge.com/default.aspx`

swers for the best drug are concerned, but "best" does not necessarily equate to "most studied" in clinical superlative questions. Demner-Fushman and Lin (2006)'s approach cannot be used to answer questions such as Examples (2.22) and (2.23), because looking for the most studied drugs will not provide an answer to the question of which drug has the fewest side effects or is safest to use.

(2.22) Which drug for treating X has the fewest side effects?

(2.23) Which drug is safest to use for treating X?

Demner-Fushman and Lin (2006) and Demner-Fushman (2006) deal with small sets of well-formed clinical questions, 30 and 50 questions respectively. In order to deal with a large corpus of clinical questions of the comparative form, which are not asked in compliance with the PICO format, a different approach than Demner-Fushman and Lin's might be preferred to translating the comparative questions into the PICO format for identifying all relevant parts of the comparison.

2.4.4 Answering Clinical Questions with Knowledge-Based and Statistical Techniques

Demner-Fushman and Lin (2007) present a system based on the knowledge extractors developed in Demner-Fushman and Lin (2005, 2006), which use knowledge-based and statistical techniques to identify different elements of MEDLINE® abstracts. These elements are the input for an algorithm that scores "the relevance of citations with respect to structured representations of information needs" within the framework of Evidence Based Medicine (Demner-Fushman and Lin, 2007). The system reorders an initial list of abstracts retrieved by PubMed to bring relevant abstracts into higher ranks. It is evaluated on real-word clinical questions and performs significantly better than previous systems used by clinicians up to the date of Demner-Fushman and Lin (2007)'s research. The same task will be performed by RetroRank as described in Chapter 8, and an evaluation of the results compared to the results of Demner-Fushman and Lin (2007) will be shown.

Demner-Fushman and Lin (2007) use carefully hand-crafted queries and structured PICO frames (c.f. Section 2.4.3) instead of natural language questions as the input to their clinical QA system. Example (2.24) shows a clinical question and Example (2.25) the extracted query frame.

(2.24) In children with an acute febrile illness, what is the efficacy of single-medication therapy with acetaminophen or ibuprofen in reducing fever?

(2.25)

Search Task: therapy selection
Problem/Population: acute febrile illness/in children
Intervention: acetaminophen
Comparison: ibuprofen
Outcome: reducing fever

The query encodes the search task and the corresponding PICO structure.

In a frame-based query interface the burden of translating a query lies with the clinician. A possible advantage is that the query needs to be well thought through to make sure all important elements are captured. This avoids poorly worded queries, which are one of the obstacles to finding anwers (Ely et al., 2005). It also eradicates the need for a linguistic analysis of the queries.

To generate an answer, the MEDLINE® abstracts retrieved for a query are translated into the PICO format as well using the specific purpose-designed extractors for the *Population*, the *Problem*, the *Intervention/Comparison* and for the *Outcome*. The "Outcome" serves as the basis for the answer, because clinicians are mostly interested in the outcome that states the finding of a study. The translation of abstracts into the PICO format relies on the availability of an annotated corpus of MEDLINE® abstracts (Demner-Fushman and Lin, 2005).

The system architecture is as follows:

- **Query formulator**: Converts a clinical question in form of a PICO frame into a PubMed query. PubMed returns a list of MEDLINE® abstracts which is analysed using the knowledge extractors.
- **Semantic matcher**: Takes the PICO query frames and the annotated MEDLINE® abstracts as input and implements the EBM scoring algorithm. The output is a ranked list of abstracts.
- **Answer Generation**: Takes the ranked list of abstracts and generates an extractive summary from the "Outcome" section.

In order to develop the scoring algorithm, a corpus of 50 clinical questions was manually created from clinical questions from the *Journal of Family Practice* and the

Parkhurst Exchange. All questions were manually classified into one of the four clinical tasks (Therapy, Diagnosis, Prognosis and Etiology). In the final preparation step the questions were translated into PICO frames. Table 2.1 shows the composition of the corpus.

	Therapy	Diagnosis	Prognosis	Etiology	Total
Development	10	6	3	5	24
Test	12	26	3	5	26

Table 2.1: Composition of the clinical question collection (Demner-Fushman and Lin, 2007).

Example (2.26) shows an example therapy question. Therapy questions are the type of questions the research in this thesis is concerned with.

(2.26) Does quinine reduce leg cramps for young athletes? (Therapy)

search task: therapy selection
primary problem: leg cramps
co-occurring problems: muscle cramps, cramps
population: young adult
intervention: quinine

The P in PICO was broken up into population, primary problem, and co-occurring problems. This plays an important role for the scoring algorithm which treats those three facets differently. The relevance of an article includes "contributions from matching PICO structures, the strength of evidence of the citation, and factors specically associated with the search tasks (and indirectly, the clinical tasks)" (Demner-Fushman and Lin, 2007). Each score reflects the factors a clinician takes into consideration when examining a MEDLINE® abstract. The assignment of numeric scores and weights is based on intuition.

The first component of the EBM scoring algorithm is based on the score of an abstract which is based on the extracted PICO elements. S_{PICO} is broken into components according to the following formula:

$$S_{PICO} = S_{problem} + S_{population} + S_{intervention} + S_{outcome}$$

(Demner-Fushman and Lin, 2007)

- $\mathbf{S}_{problem}$: The score equals 1 if the problems of the query frame and the primary problem in the abstract match exactly based on their unique UMLS concept ID. A partial string match gets a score of 0.5. If there is no overlap, the score is -1. If the problem extractor cannot identify a problem in the abstract but the query contains one, a score of -0.5 is given.

- $\mathbf{S}_{population}$ & $\mathbf{S}_{intervention}$: The overlap between the query frame elements and the corresponding elements from the abstract is measured and a point is given for each matching intervention and population.

- $\mathbf{S}_{outcome}$: The value assigned to the highest-scoring outcome sentence in the abstract as given by the "Outcome Extractor". Outcomes are omitted in the query representation because they are rarely specified in the corpus. The inherent quality of the outcome statements in an abstract is considered independent of the query, because it is assumed that "given a match on the primary problem, all clinical outcomes are likely to be of interest to the physician" (Demner-Fushman and Lin, 2007).

The second component of the EBM scores is based on the strength of evidence, which is calculated in the following way:

$$S_{SoE} = S_{journal} + S_{study} + S_{date}$$

(Demner-Fushman and Lin, 2007)

- $S_{journal}$: A score of 0.6 is assigned to articles published in a core and high-impact journal, otherwise the score is 0.

- S_{study}: A score of 0.5 is assigned for clinical trials, a score of 0.3 is assigned for observational studies, non-clinical publications receive a score of -1.5 and 0 otherwise.

- S_{date}: $S_{date} = (year_{publication} - year_{current}/100)$, which favours more recent articles.

The third and final component of the scoring algorithm is based on manually assigned MeSH terms for each search task. For each clinical task a list of positive and negative relevance indicators was collected. The score S_{task} is assigned by:

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

(Demner-Fushman and Lin, 2007)

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. A list of indicators can be found in Demner-Fushman and Lin (2007).

The EBM scoring algorithm is evaluated in terms of a document reranking task. For each question in the test collection, PubMed queries were manually crafted to fetch an initial set of MEDLINE® abstracts. The query formulation was performed by the first author, who is a medical doctor. She also verified that each set of retrieved abstracts contained at least some relevant documents. The top 50 results for each of the 50 queries were retained. Some queries retrieved less than 50 abstracts so that the total number of retrieved abstracts was 2,309. Generating a "good" PubMed query is not trivial and took on average 40 minutes per question. Determining the relevance of the retrieved abstracts is also not trivial and requires a medical degree according to Demner-Fushman and Lin (2007). Only topical relevance has been assessed for the abstract set. This process and the associated problems have been solved in the research in this thesis, which uses natural language queries and does not require a doctor to evaluate the relevance of the retrieved abstracts because the corpora used for evaluation have "gold standard" references for each question. Using an automated approach has no negative effects on system performance as will be shown in Chapter 8.

Each citation in (Demner-Fushman and Lin, 2007) was assigned one of four labels:

- **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.
- **Partially relevant**: The citation provides information that is marginally relevant.
- **Not relevant**: The citation does not provide any topically relevant information.

(Demner-Fushman and Lin, 2007)

The relevance assessment process took about 2 hours per questions or 100 hours in total.

Four different systems were compared:

- PubMed baseline (simple ordering by recency)
- A term-based reranker computing overlap between the question and citation weighted by the outcome score from the sentence where the overlap occurs.
- The EBM Scorer described above.
- A combination of the term-based reranker and the EBM scorer normalised using weighted linear interpolation.

The development questions were used for debugging and for tuning weights by trying all possible values for the combination system.

The system was evaluated using the following metrics:

- **Mean Average Precision (MAP)**: The average of the precision values after each relevant document is retrieved (Baeza-Yates, Ricardo A. and Ribeiro-Neto, B., 1999).
- **P@10**: The fraction of relevant documents in the first ten results.
- **Mean Reciprocal Rank (MRR)**: Measures how far down on a list the first relevant abstract is.
- **Total Document Reciprocal Rank (TDRR)**: The sum of the reciprocal ranks of all relevant documents. Unlike MRR it captures the ranks of all relevant documents.

The systems were evaluated under a lenient and a strict condition. Under the lenient condition documents from the categories “contains answer” and “relevant” were considered relevant. Under the strict condition only documents from the category “contains answer” were considered. Here only the tables for the strict evaluation are reported because they correspond to the evaluation of the RetroRank system in this thesis.

Dev	P@10	MAP	MRR	TDRR
PubMed	0.153	0.105	0.385	0.653
Term	0.24	0.183	0.527	0.974
EBM	0.328	0.264	0.693	1.371
Combo	0.34	0.26	0.656	1.315

Test	P@10	MAP	MRR	TDRR
PubMed	0.069	0.045	0.19	0.328
Term	0.15	0.092	0.346	0.632
EBM	0.196	0.129	0.433	0.765
Combo	0.219	0.138	0.494	0.851

Table 2.2: Performance of all systems across all clinical tasks (Demner-Fushman and Lin, 2007).

Table 2.2 contains the strict evaluation for performance across all clinical tasks for the development and test set.

Table 2.3 shows the results for both sets and for all systems on the therapy task under the strict condition. This task is comparable to the task for the QA system in this thesis.

The EBM-based reranker and combination reranker significantly outperforms the PubMed baseline for all metrics. In almost all cases, the EBM and Combo ranking algorithms perform significantly better than the term-based one. The results of both tables will be further discussed in the evaluation of RetroRank in Chapter 8.

While the results of (Demner-Fushman and Lin, 2007) show a significant improvement over the PubMed baseline, the system involves a lot of manual labour in terms of query frame generation, annotating MEDLINE® abstracts to train the PICO extractors and assessment of the query results by a doctor to determine the relevant abstracts, while also leaving the burden of translating a clinical question into PICO frames to a clinician if the system was used later on. These problems have been addressed and an automated solution has been implemented in the RetroRank system developed in this research.

Therapy Dev

	P@10	MAP	MRR	TDRR
PubMed	0.13	0.088	0.35	0.61
Term	0.23	0.205	0.409	0.872
EBM	0.35	0.314	0.675	1.434
Combo	0.35	0.301	0.569	1.282

Therapy Test

	P@10	MAP	MRR	TDRR
PubMed	0.18	0.061	0.282	0.495
Term	0.192	0.082	0.368	0.7
EBM	0.233	0.109	0.397	0.807
Combo	0.258	0.12	0.556	0.969

Table 2.3: Performance of all systems across on the therapy task (Demner-Fushman and Lin, 2007).

2.5 Clinical Question Answering Services and Applications

There are several clinical QA services and applications. Some of the services such as ATTRACT, which is provided by the Welsh National Public Health Service, are manual and the questions are primarily dealt with by clinical librarians or health-care professionals, while others are automated systems with web interfaces such as *ask*MEDLINE. This section describes the systems that are most relevant to the current research.

2.5.1 The National Library of Health Question Answering Service (NLH QAS)

The NLH Question Answering Service (QAS) was an on-line service that clinicians in the UK could use to ask questions, that were then answered by a team of clinical librarians from TRIP Database Ltd.[10], founded by Jon Brassey and Dr Chris Price. The

[10] http://www.tripdatabase.com/index.html

NHS QAS service was discontinued in 2008 but its archive of questions and answers was integrated into ATTRACT [11] run by Jon Brassey. The questions and their answers from the NLH QAS were retained at the website and indexed by major clinical topics (e.g., cancer, cardiovascular disease, diabetes, etc.) so that clinicians could consult the QA archive to check whether information relevant to their own clinical question was already available and if not to pose a new question on-line.

Clinical librarians responded to a question with a list and/or summary of articles that may address it, found via searches of PubMed®, Cochrane or TRIP. Although the service provided useful answers there are some limitations to manual QA services provided by professional staff. For example, they can only handle a limited number of questions at one time and there is no real-time response. QA services delivered by librarians also do not provide interpretative summaries or evidence-based guidance because of lack of clinical knowledge and time (Vincent, 2006; Ward, 2005), so the responsibility for selecting the right information and judging the validity of the answers still lies with the clinicians. Section 4.1 discusses the use of the NLH QAS as a source for collecting an initial corpus of clinical comparison questions (Corpus 1).

2.5.2 The Essential Evidence Plus POEM Archive

Essential Evidence Plus[12] is owned by the publisher John Wiley & Sons and its purpose it to provide tools to health care professionals that give them the most relevant and valid information currently available. One of their services is InfoPOEMs founded by Drs. Barry, Ebell and Slawson in 1990. POEM stands for "Patient-Oriented Evidence that Matters". POEMs are summaries of essential evidence-based research filtered for their relevance to patient care. They consist of a clinical question, a bottom line which summarises the main point of the synopsis, a citation for a reference article, the study type and setting, and a summary of the reference article by a physician. Figure 2.3 shows an example POEM. The reference incuded in a POEM serves as the "gold standard" reference for evaluating the accuracy of RetroRank.

In Section 4.2 the use of the POEM archive for creating Corpus 2 is discussed.

[11] `http://www.attract.wales.nhs.uk/`

[12] `http://www.essentialevidenceplus.com/`

Budesonide > mesalamine for mild-mod Crohn's exacerbation

Daily POEMs

November 1998

Clinical question
Which is more effective for the treatment of Crohn's disease, a controlled release form of budesonide or a slow release form of mesalamine?

Bottom line
Budesonide offers better protection from recurrence, reduced disease severity, and improved quality of life compared with mesalamine over a 4 month period. The decision to use this steroid should of course be made with the patient in light of the increased risks associated with long-term steroid use. (LOE = 1b)

Reference
Thomsen OO, Cortot A, Jewell D, et al. A comparison of budesonide and mesalamine for active Crohn's disease. N Engl J Med 1998; 339: 370-4.

Study design: Randomized controlled trial (double-blinded)

Setting: Outpatient (specialty)

Synopsis
A total of 182 adult patients with mild to moderately active Crohn's disease (score 200 - 400 on a Crohn's Disease Activity Index of 0 to 700) were randomized to either mesalamine (Pentasa) 2 gm po bid or slow release budesonide (Entocort) 9 mg po qd, with matching placebos. The primary outcomes were remission rates, disease activity measured using the above index and quality of life by the Psychological General Well-Being index, measured throughout the 16 week study period. Analysis was by inteition-to-treat. The remission rate was higher at 16 weeks for the budesonide group (62% vs 36%, p < 0.001), which corresponds to a number needed to treat (NNT) of 4. Scores for disease activity and quality of life were both significantly better both clinically and statistically for budesonide compared with mesalamine. While adverse effects were common, serious adverse events were more common in the mesalamine group (17 vs 10). Adrenal function was not surprisingly suppressed in the budesonide group, although typical steroid-related adverse effects such as moon facies, hirsutism, and acne were rare.

Printer Friendly

Outside Links

View article via PubMed

Earn CME credit for this search

Search data is stored for 7 days. Click here to earn credit.

(Don't show this message again)

Figure 2.3: Example POEM.

2.5.3 The Cochrane Database of Systematic Reviews (CDSR) from the Cochrane Library

The Cochrane Library[13] is an online collection of six databases that contain high-quality, independent evidence on the effectiveness of healthcare treatments and interventions, as well as methodology and diagnostic tests to inform healthcare decision-

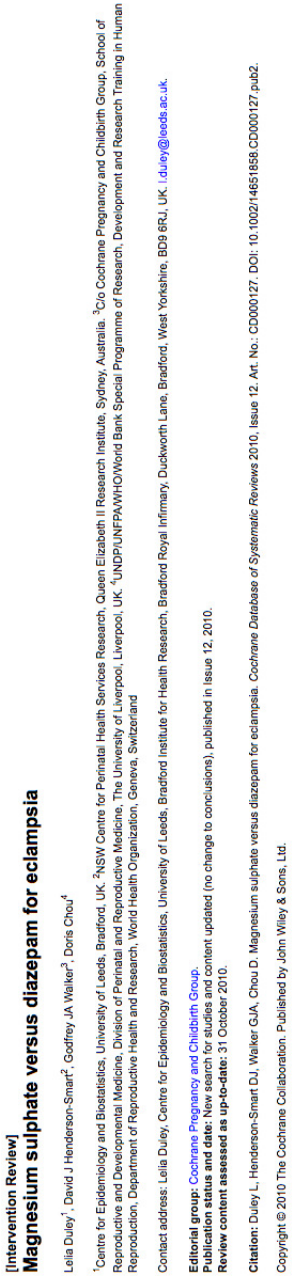
[13] http://www.thecochranelibrary.com/view/0/index.html

making (Collaboration, 2011). One of these databases is the Cochrane Database of Systematic Reviews (CDSR), the leading resource for systematic reviews in evidence-based health care. It contains over 4,500 systematic reviews and 2,000 protocols describing the research methods and objectives for reviews in progress. The Cochrane reviews are high-quality, peer-reviewed systematic reviews of primary research in human health care and health policy. Systematic reviews are critical assessments of clinical evidence for a particular clinical question such as "How efficient and safe are corticosteroids in the treatment of pneumonia?"[14]. They include a comprehensive literature search, assess the quality of studies and report the results in a systematic way (Ebell et al., 2004). The reviews are updated regularly to ensure they present the most up-to-date evidence.

Each systematic review is published in several versions that differ in the amount of detail. The shortest form of a systematic review is called the "Summary". The "Summary" has three sections. The first section names the type of review, e.g., *Intervention Review* and gives the title of the review. It also shows the authors and their affiliations. It lists the editorial group, e.g., "Cochrane Pregnancy and Childbirth Group", the publication status and the date when the review was last assessed as being up-to-date. It also gives the citation key for the review. An example is presented in Figure 2.4 showing the systematic review with the title "Magnesium sulphate versus diazepam for eclampsia".

The second section of a "Summary" is called the "Abstract", and it has seven parts. The first part is a "Background" section detailing the topic of the review. The second part is an "Objectives" section, which explains the purpose of the review and says which drugs or interventions are compared to each other and in what regard, i.e., safety, efficacy and cost-effectiveness. The third part is a "Search Strategy" section, which describes which sources were used to find relevant studies. The fourth part is a "Selection Criteria" section, which says which study types were included, e.g., clinical randomised trials comparing the two drugs in question for the relevant disease. The fifth part is a "Data Collection and Analysis" section, which describes the criteria used in collecting and analysing the data used for the review. The sixth part is the "Main Results" section, which gives the results of all the studies that were considered relevant for the drug comparison. The seventh part is the "Author's Conclusion", which contains a summary of the main results and draws a conclusion. An example of the

[14] `http://onlinelibrary.wiley.com/o/cochrane/clsysrev/articles/CD007720/frame.html`



[Intervention Review]

Magnesium sulphate versus diazepam for eclampsia

Lelia Duley[1], David J Henderson-Smart[2], Godfrey JA Walker[3], Doris Chou[4]

[1]Centre for Epidemiology and Biostatistics, University of Leeds, Bradford, UK. [2]NSW Centre for Perinatal Health Services Research, Queen Elizabeth II Research Institute, Sydney, Australia. [3]C/o Cochrane Pregnancy and Childbirth Group, School of Reproductive and Developmental Medicine, Division of Perinatal and Reproductive Medicine, The University of Liverpool, Liverpool, UK. [4]UNDP/UNFPA/WHO/World Bank Special Programme of Research, Development and Research Training in Human Reproduction, Department of Reproductive Health and Research, World Health Organization, Geneva, Switzerland

Contact address: Lelia Duley, Centre for Epidemiology and Biostatistics, University of Leeds, Bradford Institute for Health Research, Bradford Royal Infirmary, Duckworth Lane, Bradford, West Yorkshire, BD9 6RJ, UK. l.duley@leeds.ac.uk.

Editorial group: Cochrane Pregnancy and Childbirth Group.
Publication status and date: New search for studies and content updated (no change to conclusions), published in Issue 12, 2010.
Review content assessed as up-to-date: 31 October 2010.

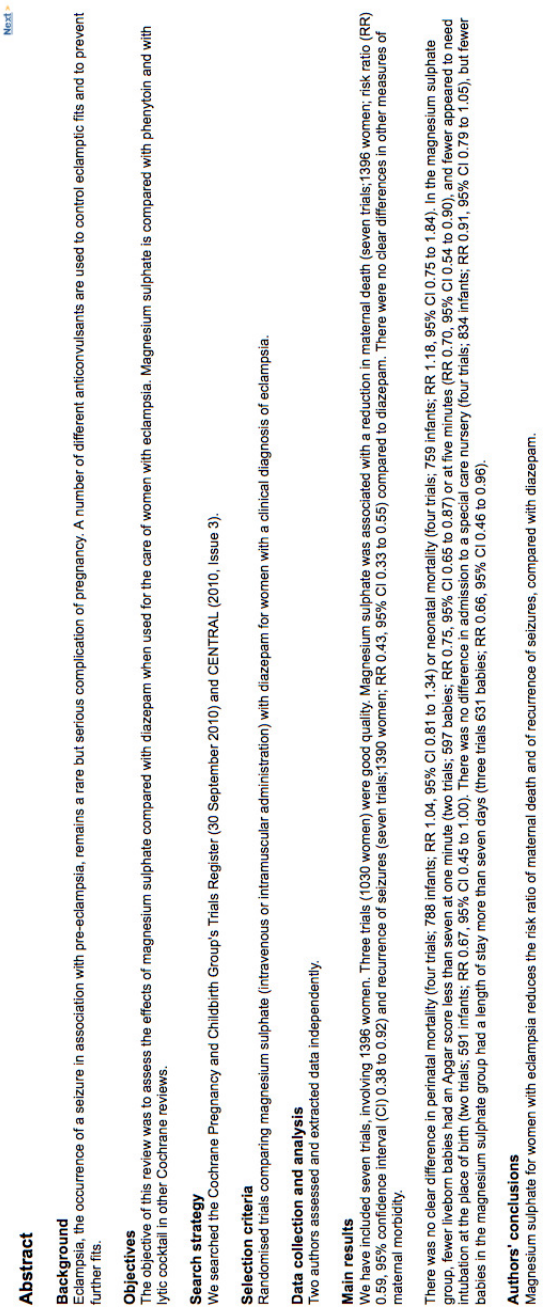
Citation: Duley L, Henderson-Smart DJ, Walker GJA, Chou D. Magnesium sulphate versus diazepam for eclampsia. *Cochrane Database of Systematic Reviews* 2010, Issue 12. Art. No.: CD000127. DOI: 10.1002/14651858.CD000127.pub2.

Figure 2.4: Header of a Systematic Review.

"Abstract" section is shown in Figure 2.5.

At the end of every "Summary" review is a "Plain Language Summary" section. This section repeats the title of the review, describes the disease(s), contains a short description of the compared drugs and describes the main benefits and side effects of the drugs. An example is shown in Figure 2.6.

The second version available of each systematic review is called the "Standard" version. The "Standard" version provides a greater level of detail. In addition to all



Next

Abstract

Background
Eclampsia, the occurrence of a seizure in association with pre-eclampsia, remains a rare but serious complication of pregnancy. A number of different anticonvulsants are used to control eclamptic fits and to prevent further fits.

Objectives
The objective of this review was to assess the effects of magnesium sulphate compared with diazepam when used for the care of women with eclampsia. Magnesium sulphate is compared with phenytoin and with lytic cocktail in other Cochrane reviews.

Search strategy
We searched the Cochrane Pregnancy and Childbirth Group's Trials Register (30 September 2010) and CENTRAL (2010, Issue 3).

Selection criteria
Randomised trials comparing magnesium sulphate (intravenous or intramuscular administration) with diazepam for women with a clinical diagnosis of eclampsia.

Data collection and analysis
Two authors assessed and extracted data independently.

Main results
We have included seven trials, involving 1396 women. Three trials (1030 women) were good quality. Magnesium sulphate was associated with a reduction in maternal death (seven trials;1396 women; risk ratio (RR) 0.59, 95% confidence interval (CI) 0.38 to 0.92) and recurrence of seizures (seven trials;1390 women; RR 0.43, 95% CI 0.33 to 0.55) compared to diazepam. There were no clear differences in other measures of maternal morbidity.

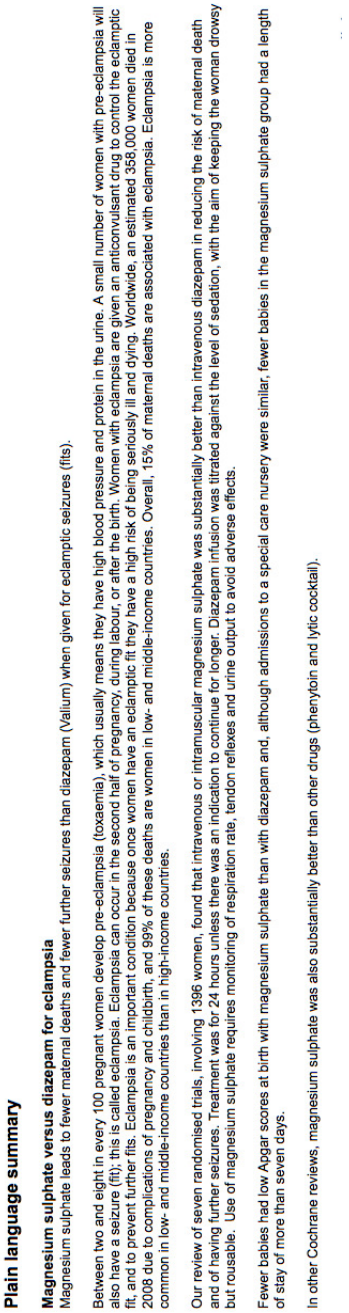
There was no clear difference in perinatal mortality (four trials; 788 infants; RR 1.04, 95% CI 0.81 to 1.34) or neonatal mortality (four trials; 759 infants; RR 1.18, 95% CI 0.75 to 1.84). In the magnesium sulphate group, fewer liveborn babies had an Apgar score less than seven at one minute (two trials; 597 babies; RR 0.75, 95% CI 0.65 to 0.87) or at five minutes (RR 0.70, 95% CI 0.54 to 0.90), and fewer appeared to need intubation at the place of birth (two trials; 591 infants; RR 0.67, 95% CI 0.45 to 1.00). There was no difference in admission to a special care nursery (four trials; 834 infants; RR 0.91, 95% CI 0.79 to 1.05), but fewer babies in the magnesium sulphate group had a length of stay more than seven days (three trials 631 babies; RR 0.66, 95% CI 0.46 to 0.96).

Authors' conclusions
Magnesium sulphate for women with eclampsia reduces the risk ratio of maternal death and of recurrence of seizures, compared with diazepam.

Figure 2.5: Abstract section of a Systematic Review.

the information in the "Summary" version, it also contains the objectives and methods used in the included studies, a description of each study and the results, a discussion section, a detailed data analysis section and a full list of all references retrieved for conducting the systematic review, as well as a table detailing the characteristics of each study and an explanation of why the study is or is not relevant for the systematic review.

The references mentioned in the reviews fall into three categories:



Plain language summary

Magnesium sulphate versus diazepam for eclampsia

Magnesium sulphate leads to fewer maternal deaths and fewer further seizures than diazepam (Valium) when given for eclamptic seizures (fits).

Between two and eight in every 100 pregnant women develop pre-eclampsia (toxaemia), which usually means they have high blood pressure and protein in the urine. A small number of women with pre-eclampsia will also have a seizure (fit); this is called eclampsia. Eclampsia can occur in the second half of pregnancy, during labour, or after the birth. Women with eclampsia are given an anticonvulsant drug to control the eclamptic fit, and to prevent further fits. Eclampsia is an important condition because once women have an eclamptic fit they have a high risk of being seriously ill and dying. Worldwide, an estimated 358,000 women died in 2008 due to complications of pregnancy and childbirth, and 99% of these deaths are women in low- and middle-income countries. Overall, 15% of maternal deaths are associated with eclampsia. Eclampsia is more common in low- and middle-income countries than in high-income countries.

Our review of seven randomised trials, involving 1396 women, found that intravenous or intramuscular magnesium sulphate was substantially better than intravenous diazepam in reducing the risk of maternal death and of having further seizures. Treatment was for 24 hours unless there was an indication to continue for longer. Diazepam infusion was titrated against the level of sedation, with the aim of keeping the woman drowsy but rousable. Use of magnesium sulphate requires monitoring of respiration rate, tendon reflexes and urine output to avoid adverse effects.

Fewer babies had low Apgar scores at birth with magnesium sulphate than with diazepam and, although admissions to a special care nursery were similar, fewer babies in the magnesium sulphate group had a length of stay of more than seven days.

In other Cochrane reviews, magnesium sulphate was also substantially better than other drugs (phenytoin and lytic cocktail).

Figure 2.6: Plain language summary of a Systematic Review.

1. References to studies included in the review.

2. References to studies excluded from the review.

3. References to studies awaiting assessment.

The third version of each systematic review is an even more detailed, extended version of the “Standard” review, containing additional detail in all sections except for the abstract.

For the research reported in this thesis, the references from the category of "References included" in the review will be used as "gold standard" data for development and testing of the post-retrieval ranking system described in Chapter 8 in this work. The corpus collection is described in Section 4.3.

2.5.4 *ask*MEDLINE

*ask*MEDLINE[15] by Fontelo et al. (2005) is a search tool that allows clinicians and researchers to search MEDLINE® using free-text natural language questions via a web interface similar to that of Entrez PubMed and shown in Figure 2.7. *ask*MEDLINE retrieves relevant MEDLINE® articles and provides links to journal abstracts, full-text articles and related items as illustrated in Figure 2.8. It evolved from a search tool in the PICO format which was more cumbersome and difficult to use by busy clinicians and uses the MeSH vocabulary, which provides a reliable means of retrieving information that uses different terms for the same concept.

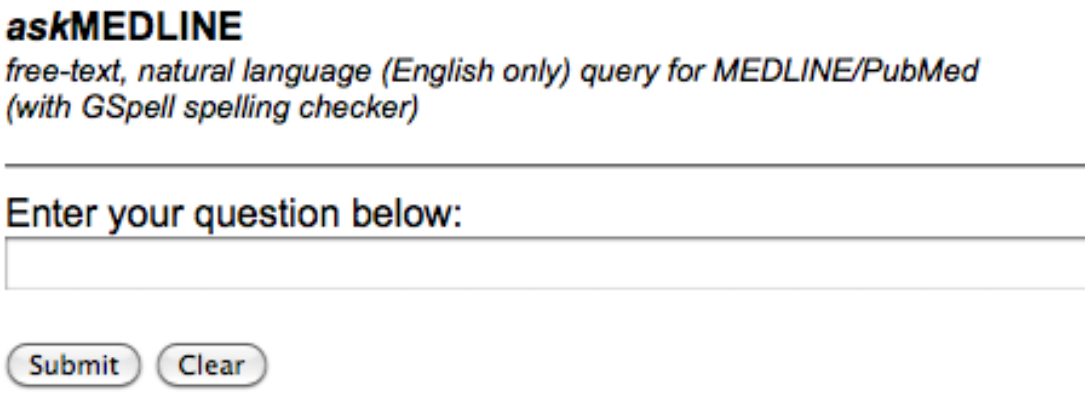
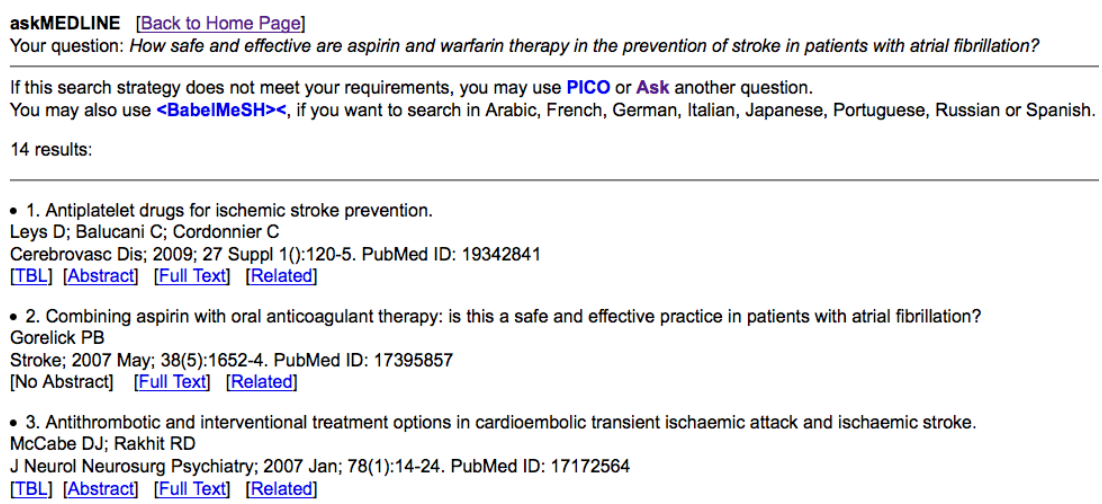


Figure 2.7: *ask*Medline search interface.

Unlike Entrez PubMED, which simply searches for all words in the input text string except for those found on the PubMed stop word list (Appendix C), *ask*MEDLINE employs a multi-round search strategy. During the first round the parser ignores punctuation marks and deletes words from a stopword list that contains PubMed stop words and other words that Fontelo et al. (2005) found detrimental to the search. The list of PubMed stop words can be found in Appendix C. The additional list of stopwords that Fontelo et al. use does not appear to have been published.

After the stopwords have been removed, the modified query is sent to the PubMed Entrez' E-Utilities described in Section 2.3.3. Terms with the label "All Fields" which

[15] http://askmedline.nlm.nih.gov/ask/ask.php



askMEDLINE [Back to Home Page]
Your question: *How safe and effective are aspirin and warfarin therapy in the prevention of stroke in patients with atrial fibrillation?*

If this search strategy does not meet your requirements, you may use **PICO** or **Ask** another question.
You may also use **<BabelMeSH><**, if you want to search in Arabic, French, German, Italian, Japanese, Portuguese, Russian or Spanish.

14 results:

- 1. Antiplatelet drugs for ischemic stroke prevention.
Leys D; Balucani C; Cordonnier C
Cerebrovasc Dis; 2009; 27 Suppl 1():120-5. PubMed ID: 19342841
[TBL] [Abstract] [Full Text] [Related]

- 2. Combining aspirin with oral anticoagulant therapy: is this a safe and effective practice in patients with atrial fibrillation?
Gorelick PB
Stroke; 2007 May; 38(5):1652-4. PubMed ID: 17395857
[No Abstract] [Full Text] [Related]

- 3. Antithrombotic and interventional treatment options in cardioembolic transient ischaemic attack and ischaemic stroke.
McCabe DJ; Rakhit RD
J Neurol Neurosurg Psychiatry; 2007 Jan; 78(1):14-24. PubMed ID: 17172564
[TBL] [Abstract] [Full Text] [Related]

Figure 2.8: Extract of *ask*MEDLINE results for the question: "How safe and effective are aspirin and warfarin therapy in the prevention of stroke in patients with atrial fibrillation?"

are defined as neither MeSH subject nor MeSH subheadings are checked against the "MeSH Backup vocabulary" which includes words classified as "other eligible entries" such as MeSH descriptors. If an "All Fields" word is in the backup vocabulary, it remains in the query; if it is not, it is deleted.

The remaining terms are sent back to PubMed, again through E-Utilities with "Human" and "English" language limits added, restricting the search to studies about human subjects published in English. If the journal retrieval count after the first round is between 1 and 50,000, the first 20 results are displayed in the browser and the search process ends. Results 21 to n can be found on subsequent pages. If no references were found in the first round or if the number of retrieved references was larger than 50,000, the search process enters a second round.

In the second round, two different strategies are employed depending on whether the search was too narrow because too many terms were included in the search and/or too many filters used, or whether the search was too broad because too few terms were included in the search.

In the first case, the "All Fields" words are removed from the query even though they are in the backup vocabulary and only MeSH Terms and Subheadings remain. In the second case, the "All Fields" words that were removed from the search in the first round, because they were not found in the backup vocabulary, are put back into the query, which means the query now contains all MeSH terms and all "All Fields" terms from the original question.

For both cases, the updated query is sent to Entrez E-Utilities again and the retrieved journal articles are displayed in the browser. The search process stops when the retrieval count is between 1 and 50,000 as in the first round. If the second round still does not produce search results, the process enters a third round.

During the third round, terms from another list of "No-Go Terms" are added to the query. The list of "No-Go Terms" includes common MeSH abbreviations, acronyms and words like, "method", "affect", and "lead", and other terms that could result in a successful search and is updated with new terms as they are encountered.

The modified query is again sent to E-Utilities and the retrieved journal articles are displayed. A retrieval count between 1 to 50,000 ends the search process and the first 20 articles are displayed to the user on the first page. If more than 20 articles were retrieved, the rest can be found on subsequent pages. If *ask*MEDLINE only retrieves one to four articles, a search is automatically done for articles related to the top two articles. All original articles and the first 25 related articles of the first two are retrieved and the first 20 are displayed in the browser.

(Fontelo et al., 2005) have compared *ask*MEDLINE's performance to the performance of Entrez PubMed introduced in Section 2.1. In this assessment, the accuracy and relevance of retrieved citations was determined using the "gold standard" reference from POEMs (see Section 2.5.2) and CATs (Critically Appraised Topics) from the University of Michigan, Department of Pediatrics, Evidence-Based Pediatrics website[16]. Unlike POEMs, CATs can have more than one cited reference that can be used as a "gold standard".

(Fontelo et al., 2005) used 95 POEM questions in this comparative assessment. The first pass checks whether the gold standard reference is among the retrieved abstracts. Here, *ask*MEDLINE found 62.1% (59/95) of the articles cited in POEMs, while Entrez retrieved 13.7% (13/95). For CATs, *ask*MEDLINE found 64.2% (18/28) and Entrez retrieved 3.6% (1/28). After including related articles, *ask*MEDLINE found 11.6% more gold standard references and Entrez another 8.4%. For CATs, 10.6% (Entrez 3.6%) more gold standard references were retrieved after including related articles. After rephrasing three questions, *ask*MEDLINE found two more matches while Entrez retrieved none. Rephrasing added 14.3% to *ask*MEDLINE's accuracy for CATs (7.1% to Entrez). In total, *ask*MEDLINE correctly matched 75.8% of the cited references in POEMs and 89.2% of the references cited in response to CATs questions after the first three steps, while Entrez correctly matched 22.1% of the POEM references and 14.3%

[16] `http://www.med.umich.edu/pediatrics/ebm/cat.htm`

of the CAT references.

In cases, where *ask*MEDLINE did not find the specific cited reference for 20 POEM questions, it found journal citations that were deemed relevant and would be useful in answering the question. For Entrez, references for an additional 16 POEM questions were considered relevant. For CATs *ask*MEDLINE found 2 citations that were deemed relevant, while Entrez found none. The overall POEM retrieval failure for *ask*MEDLINE is 3.1% and 61% for Entrez. The overall CAT retrieval failure for*ask*MEDLINE is 3.6% and 85.7% for Entrez.

Both *ask*MEDLINE and Entrez PubMed (c.f. Section 2.3.3) will serve as the basis for comparison for my abstract retrieval system as two of the few free-text, natural query search engines specifically developed for MEDLINE®/PubMed queries. Neither systems was developed with clinical medication comparison questions in mind, and it will be shown in Section 6.3 that their performance drops significantly when dealing with comparison questions. A new clinical QA system, AskHermes[17], was made available online in late 2010. Due to its unavailability at the time this research was undertaken, it was not considered in the system comparison.

2.6 Summary

This chapter gave an overview of the disciplines involved in this research and provided knowledge about the information needs of clinicians and the domain of Evidence Based Medicine (EBM), as well as an overview of Information Retrieval (IR) and Question Answering (QA) techniques in the clinical domain and existing clinical QA services and applications that provide the background for the current research. The retrieval and post-retrieval ranking performance of RetroRank will be evaluated against *ask*Medline, Entrez PubMed and the ranking system developed by Demner-Fushman and Lin (2007).

[17] http://www.askhermes.org/

Chapter 3

Comparative Constructions

The goal of this chapter is to introduce the concept of comparative constructions and the lexical items and phrases indicative of comparison questions, which are used for the creation of Corpus 1 as described in Section 4.1.

3.1 General Purpose and Form in Questions

Comparative constructions are a common phenomenon in the English language and a description can be found in all books on English grammar. This chapter provides a general description of comparisons based on *The Cambridge Grammar of the English Language* by Huddleston and Pullum (2002) as well as examples of the different comparative constructions occurring in clinical questions.

Comparative questions express relations based on similarities or differences between entities. In this research, the term entity refers to drugs and treatment methods or interventions. Comparisons of different entities often occur in questions asked by physicians and in medical literature reporting results from clinical trials, comparative studies, and systematic reviews.

Comparatives can be scalable or non-scalable and both groups can express equality or inequality between the compared entities. Scalable adjectives and adverbs describe attributes that can be measured in degrees and scalability refers to the possibility to place an adjective or adverb on a scale to express the degree to which it applies. Non-scalable adjectives and adverbs cannot be measured in degrees. Equality refers to constructs where two or more compared entities are equal in respect to a shared quality, whereas inequality emphasises the difference between entities in respect to a certain quality. My clinical comparison question corpora contain questions of all of the above

Scalability	Equality	Example
+	+	As efficient as x
-	+	Same intervention as x
+	-	Better treatment than x
-	-	Drug x differs from drug y

Table 3.1: Features of comparatives.

mentioned combinations of scalability and equality. Table 3.1 gives an example showing the four possibilities for drugs and interventions:

Comparison can take a comparative form, a superlative form, or neither as in the "same" and "different from" examples. The comparative form is used to compare two entities with respect to a certain attribute. The superlative form compares or contrasts one entity with a set of other entities and expresses the end of a spectrum. The following examples illustrate the difference:

Comparative form: Is Ibuprofen **better** than Paracetamol for treating pain?
Superlative form: Is Ibuprofen the **best** treatment for pain?

3.2 Lexical Items Indicating Comparative Constructions

Comparisons are conveyed in many ways by lexical items and phrases. These lexical items and their respective part-of-speech tags were used to extract a subset of comparison questions from a clinical QA corpus. (In chapter 4.2, it is described how.) In the three corpora used in this research, the following lexical items and phrases occur as indicators of comparison questions:

1. Comparative adjectives and adverbs:
Regular adjectives and adverbs:
ADJ/ADV -er (e.g. *safer* [than][1] drug/intervention x] [for y]
Irregular adjectives and adverbs:
e.g. *worse/better* [than] or *as good as* drug/intervention x] [for y]
Analytical adjectives and adverbs:

[1] *Than* is optional. For example see A or B: What is safer?

e.g. *less/more* ADJ/ADV [than drug/intervention x][for y]

2. Superlative adjectives and adverbs:
Regular adjectives and adverbs: ADJ/ADV *-est* (eg.safest) x [for y]
Irregular adjectives and adverbs: e.g. *worst/best* x [for y]
Analytical adjectives and adverbs: e.g. *least/most* ADJ/ADV x [for y]

3. Verbs, nouns and coordinating conjunctions:
Verbs: compared to/with, differ from
Nouns: comparison, difference
Coordinating conjunctions: versus/vs, or, and, instead of

3.3 Summary

This chapter introduced comparative constructions and the lexical items and phrases indicative of comparison questions, which are used for the creation of Corpus 1 described in Section 4.1.

Chapter 4

Creating Three Corpora of Clinical Comparison Questions

This chapter describes the creation and preprocessing of the three corpora of clinical questions. The three corpora described in this chapter serve different purposes. The NLH QAS corpus (Corpus 1) described in Section 4.1 was used for the initial MEDLINE® retrieval experiment to be described in Chapter 5 and for developing and testing the automated retrieval component of RetroRank to be described in Chapter 6. However, a different set of questions was necessary for evaluating the automated system according to the same criteria used in the *ask*MEDLINE and Entrez PubMed evaluation published in Fontelo et al. (2005). Therefore, the corpus of the Essential Evidence Plus POEM questions (Corpus 2) in Section 4.2 was collected, along with the "gold standard" reference for each question. Corpus 2 was also used for developing the post-retrieval ranking strategies described in Chapter 7. Because Corpus 2 only has one "gold-standard" reference for each clinical question, it could not be used to calculate Mean Average Precision (MAP) or Mean Rank Precision (MRR) for the post-retrieval ranking module of RetroRank. Therefore, a third corpus, the Cochrane Systematic Review Corpus (Corpus 3), was collected. This corpus provides multiple "gold-standard" references for each clinical question, which allows a system evaluation using the standard IR metrics.

4.1 Corpus 1

A programme was developed to automatically extract all questions and answers available on the 20th of July 2007 from the NLH clinical question answering service (QAS)

website at `http://www.clinicalanswers.com` introduced in Section 2.5.1. This data was used to create two separate XML files containing the questions and the answers.

A simple XML format brackets the major structures of the text. During this process a total of 4,580 unique Q-A pairs of different degrees of difficulty and complexity for 34 medical fields was collected, representing questions asked and answered over a 36 month period. These questions and answers form the corpus of clinical questions which were used during the initial retrieval experiment described in Chapter 5 and for developing the automated retrieval component of RetroRank described in Chapter 6. The corpus can be expanded with new questions from ATTRACT (cf. Section 2.5.1) if more data is necessary. Figure 4.1 and 4.2 illustrate examples of the format of the question and answer files created from the NLH QAS:

```
<questions>
<question id="q6576">Is there any link between Glivec (used for GIST tumours) and vitreous or
retinal haemorrhages?</question>
</questions>
```

Figure 4.1: Example question file.

```
<answers>
<answer id="a6576">The SPC for Glivec [1] has a section on undesirable effects, this includes a
table of adverse reactions in clinical studies. For eye disorders it reports:
Common: Eyelid oedema, lacrimation increased, conjunctival haemorrhage, conjunctivitis, dry eye,
blurred vision
Uncommon: Eye irritation, eye pain, orbital oedema, scleral haemorrhage, retinal haemorrhage,
blepharitis, macular oedema
Rare: Cataract, glaucoma, papilloedema
Reference : 1)http://emc.medicines.org.uk/emc/assets/c/html/displaydoc.asp?documentid=15014
</answer>
</answers>
```

Figure 4.2: Example answer file.

After creating this initial corpus, the research was narrowed down to clinical comparison questions, because they are a common kind of question clinicians have and no methods yet exist that specifically address this kind of clinical question in medical QA. To create a subcorpus containing only comparison questions, the TnT tagger (Brants, 2000) was used to POS-tag the initial corpus with the Penn Treebank tagset. The TnT tagger implements smoothing by interpolation and handles unknown words by using N-Gram models on word suffixes. The tagger was trained on the Wall Street Journal (WSJ) corpus which comes with "gold standard" POS tags, since there was no matching training data available for the domain. Some comparative constructions may have

been missed because of the lack of a suitable training corpus and POS tagging errors. However a manual analysis yielded approximately the same number of comparative questions. A very small number of the POS tags of the comparative sentences had to be manually corrected. The following is an example for a tagged comparison sentence:

(4.1) Which/WDT is/VBZ *better/JJR* for/IN investigation/NN of/IN dementia/NN -/:CT/NNP or/CC MRI/NNP ?/.

To create a corpus of clinical comparison questions, the tagged corpus was searched for the POS tags of the lexical items introduced in Chapter 3.2, i.e., the tags JJR (comparative adjective), JJS (superlative adjective), RBR (comparative adverb), RBS (superlative adverb), and the lexical items from Section 3.2 indicating comparisons, namely the nouns *comparison* and *difference*, the verbs *compared to/with* and *differ from* and the coordinating conjunctions *versus* and *instead of*. *Or* was disregarded because its comparative sense only occurred in sentences already retrieved by looking for comparative and superlative adjectives and adverbs.

POS tag/Lexical item	Occurrences
JJR	195
RBR	124
JJS	207
RBS	68
CC (versus, instead of)	18
VBN (compared to/with, differ from)	45
NN (comparison, difference)	85
Total	**742**

Table 4.1: Number of lexical items indicating comparisons.

Duplicates of questions containing more than one tag which was a comparison indicator were removed. The subcorpus of comparison questions contains 742 out of the total corpus of 4580 Q-A pairs shown in table 4.1. This subset comprises approximately 17% of the original set.

A small number of false positives were found during manual post-processing, as not all words tagged as superlatives signal comparisons. Rather they are part of idiomatic expressions, such as **best practise**, or proportional quantifiers (Huddleston

and Pullum, 2002) such as **Most** NSAIDs'. Scheible (2008) distinguishes eight different classes in which the superlative construction is used in English but only five of the eight classes involve true comparisons.

4.2 Corpus 2

In order to evaluate the retrieval component of RetroRank according to the same criteria used in the *ask*MEDLINE evaluation described in Section 2.5.4, 30 medication comparison questions were collected from the Essential Evidence Plus POEM archive[1] described in Section 2.5.2. Further questions and their associated POEM reference summaries can be collected for additional testing and evaluation.

The POEM corpus was collected by searching for a number of frequent lexical items indicating comparisons (i.e. *better*, *safer*, *compared to*, etc.) that were identified in Corpus 1 and saving the retrieved comparison questions and the associated POEMs in separate files. Example (4.2) shows the first question from the POEM corpus.

(4.2) How safe and effective are aspirin and warfarin therapy in the prevention of stroke in patients with atrial fibrillation?

The full list of questions can be found in Appendix D. The questions were later processed using TTT2 and a Java pipeline which will be described in Chapter 6.

Only POEM questions were used. CATs were not used, as in the evaluation of *ask*Medline by Fontelo et al. (2005), because CATs do not contain a sufficient number of medication comparison questions.

4.3 Corpus 3

A third corpus with " gold standard" references was needed to evaluate the post-retrieval ranking methods introduced in Chapter 7. While Corpus 2 served as a starting point for developing post-retrieval ranking methods, it has the limitation of only listing one "gold standard" reference for each question. This presents a problem for a comparable system evaluation because calculating the Mean Average Precision (MAP), Mean Rank Precision (MRP), Binary Preference (Bpref) and recall requires more than one "gold standard" reference per question in order to see at which positions the "gold

[1] `http://www.essentialevidenceplus.com/content/poems`

standard" references are ranked and how many of the total number of "gold standard" references were retrieved.

To create the post-retrieval ranking evaluation corpus, systematic reviews from Cochrane Library (Section 2.5.3) were used. A manual search of the Cochrane Library identified systematic reviews involving drug comparisons. Because the titles of the systematic reviews are not in question form, a comparison question was created from the title and the description of the aim of the systematic review from the "Objectives" section. The ID of the Cochrane review was also recorded to provide a key for the question. Table 4.2 illustrates a question created from these two sources.

Cochrane ID CD000128
Title Magnesium sulphate versus for eclampsia.
Objectives The objective of this review was to assess the effects of magnesium sulphate compared with diazepam when used for the care of women with eclampsia. Magnesium sulphate is compared with phenytoin and with lytic cocktail in other Cochrane reviews.
Question What are the effects of magnesium sulphate versus diazepam for eclampsia?

Table 4.2: Creating a comparison question from Cochrane Systematic Reviews.

A total of 45 questions were collected. The full list of questions and their IDs can be found in Appendix F.

An API, which is part of the RetroRank post-retrieval ranking module described in Section 8.2.1, was developed to download the systematic reviews and the references from the category of "References included" (c.f. Section 2.5.3) for the comparison questions and recorded IDs. The "References included" form the "gold-standard" for each question. Using third party judgements in form of the references in the Cochrane reviews has the advantage of having access to "gold standard" references carefully chosen by domain experts. However, it has to be noted that the judgements are incomplete and some potentially relevant references might not be included.

Downloading and caching the Cochrane systematic reviews results in a local corpus that can be scanned and updated based on new queries. This allows further questions

and their associated Cochrane abstracts to be collected for additional testing and evaluation.

4.4 Summary

This chapter introduced three different corpora of clinical comparison questions. Corpus 1 created from the NLH QAS served as a starting point for testing different retrieval strategies for clinical comparison questions and for developing the retrieval component of RetroRank, the clinical QA system developed in this research. Corpus 2, collected from the Essential Evidence Plus POEM archive, and Corpus 3, collected from Cochrane Systematic Reviews, served as "gold standard" corpora for further system development and for the evaluation of RetroRank on standard IR metrics described in Chapter 8.

Chapter 5

Initial Experiment in MEDLINE® Abstract Retrieval

In order to develop RetroRank, the automated QA system described in Chapter 8, an initial retrieval experiment was carried out via the OVID® portal to see if MEDLINE® abstracts are a useful resource for answering comparison questions such as "Is drug A better than drug B for treating X?", and to discover the best search strategy for clinical comparison questions. Comparison questions differ from other clinical queries for which systems like *ask*Medline (c.f. Section 2.5.4) were developed and the assumption is that different retrieval strategies are needed for achieving the best possible results.

Intermediate searches, which would have been performed internally by a search engine, are included in this chapter to illustrate the impact of adding the basis of the comparison and the use of a publication type limit on the number of retrieved abstracts. The findings of the initial experiment are implemented in the retrieval component of RetroRank described in Chapter 6.

5.1 Strategies for Retrieving MEDLINE® Abstracts

For the experiment, different strategies to achieve the best possible retrieval of relevant abstracts were tried out with the assistance of Marshall Dozier, a medical librarian from the University of Edinburgh Information Service Department. We also experimented with applying the publication study type limit *comparative study*, which has different effects depending on whether the drugs mentioned in the query are well-studied or not.

For popular, well-studied drugs, looking for the drug names often leads to hundreds of returned abstracts, most of which are not relevant. By including the basis of the

comparison and limiting the study type to comparative studies, the number of returned abstracts for a set of 30 questions drops on average to 15% of the size of the original set of returned abstracts. For Example (5.1) a search for the combination of both drug names retrieved 593 abstracts. Including the basis of the comparison decreased the number to 139 abstracts. After constraining the results to comparative studies, the number of retrieved abstracts dropped to 24, which is a reduction of 83%.

For less-studied drugs, the difference in number of abstracts retrieved by including the basis of the comparison and limiting the search to the *comparative study* publication type is smaller compared to the number retrieved by only looking for the drug names, because fewer abstracts exist for these drugs, but the relevance of the returned abstracts improves as considerably as for the more studied drugs. (Recall was not analysed during the explorations because for answering clinical questions the relevance of the retrieved abstracts is more important than retrieving all possible abstracts.)

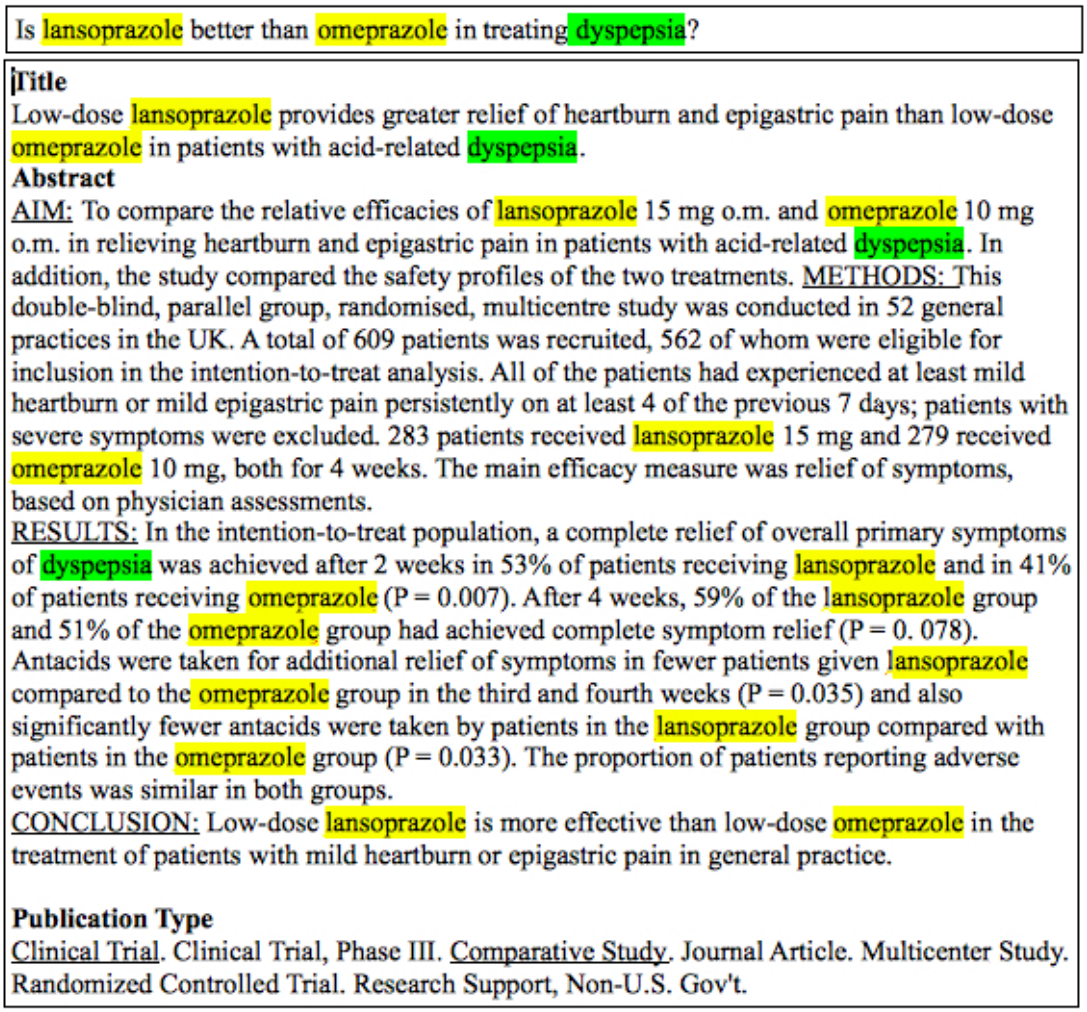
There were a small number of cases where including the basis of the comparison lead to the return of no relevant abstracts. In this case, different strategies from the one discussed above should be considered.

The search strategy for well-studied drugs is described and illustrated with Example question (5.1). The basic search strategy for less-studied drugs is the same with the exception that the comparative study limit is removed when no abstracts are retrieved.

(5.1) Is lansoprazole better than omeprazole in treating dyspepsia?

Titles and abstracts were searched for each of the compared entities (*lansoprazole* and *omeprazole*) and the basis of the comparison (*dyspepsia*). The results were combined to return only abstracts that contained both entities as well as the basis of the comparison, and were *comparative studies*. An example abstract for Question (5.1) is shown in Figure 5.1 The most common sources that were excluded by constraining the search to comparative studies are reviews, clinical trials, evaluation studies, and case reports. These may contain relevant information but the initial focus was on the study type that was most likely to increase precision. (As the evaluation in Section 5.2 shows, the restriction to *comparative studies* is insufficient to guarantee relevance and can prove too narrow.)

Often the basis of the comparison is related to symptoms which are not explicitly mentioned in the question but which are still relevant. For example, in the abstract shown in Figure 5.1 *heartburn* is mentioned as well as *dyspepsia* and the terms are often used interchangeably to refer to the same condition of a burning feeling in the chest



Is lansoprazole better than omeprazole in treating dyspepsia?

Title
Low-dose lansoprazole provides greater relief of heartburn and epigastric pain than low-dose omeprazole in patients with acid-related dyspepsia.

Abstract
AIM: To compare the relative efficacies of lansoprazole 15 mg o.m. and omeprazole 10 mg o.m. in relieving heartburn and epigastric pain in patients with acid-related dyspepsia. In addition, the study compared the safety profiles of the two treatments. METHODS: This double-blind, parallel group, randomised, multicentre study was conducted in 52 general practices in the UK. A total of 609 patients was recruited, 562 of whom were eligible for inclusion in the intention-to-treat analysis. All of the patients had experienced at least mild heartburn or mild epigastric pain persistently on at least 4 of the previous 7 days; patients with severe symptoms were excluded. 283 patients received lansoprazole 15 mg and 279 received omeprazole 10 mg, both for 4 weeks. The main efficacy measure was relief of symptoms, based on physician assessments.
RESULTS: In the intention-to-treat population, a complete relief of overall primary symptoms of dyspepsia was achieved after 2 weeks in 53% of patients receiving lansoprazole and in 41% of patients receiving omeprazole (P = 0.007). After 4 weeks, 59% of the lansoprazole group and 51% of the omeprazole group had achieved complete symptom relief (P = 0. 078). Antacids were taken for additional relief of symptoms in fewer patients given lansoprazole compared to the omeprazole group in the third and fourth weeks (P = 0.035) and also significantly fewer antacids were taken by patients in the lansoprazole group compared with patients in the omeprazole group (P = 0.033). The proportion of patients reporting adverse events was similar in both groups.
CONCLUSION: Low-dose lansoprazole is more effective than low-dose omeprazole in the treatment of patients with mild heartburn or epigastric pain in general practice.

Publication Type
Clinical Trial. Clinical Trial, Phase III. Comparative Study. Journal Article. Multicenter Study. Randomized Controlled Trial. Research Support, Non-U.S. Gov't.

Figure 5.1: Example question (5.1) and MEDLINE® abstract. The compared entities are highlighted in yellow and the basis of the comparison is highlighted in green.

and similar symptoms caused by food digestion problems. In order to recognise that different terms are actually related to the same disease and belong to the same hierarchy, advantage was taken of OVIDs ability to map the entities to their corresponding MeSH (Medical Subject Headings) terms and to "explode" the MeSH terms to include all of the narrower, more specific subheadings during the search. Figure 5.2[1] shows an extract of the MesH hierarchy that the basis of the comparison, *dyspepsia*, from example (5.1) belongs to, and it shows that *heartburn* belongs to the same hierarchy and is related and relevant. Mapping the terms in a query to MeSH terms appears to be a useful step in finding related concepts as shown in Lu et al. (2009). This approach

[1] `http://www.nlm.nih.gov/mesh/MBrowser.html`

corresponds to using WordNet[2], a lexical database for English, which organizes words into sets of synonyms called synsets, for query expansion in information retrieval such as in Voorhees (1994); Gonzalo et al. (1998).

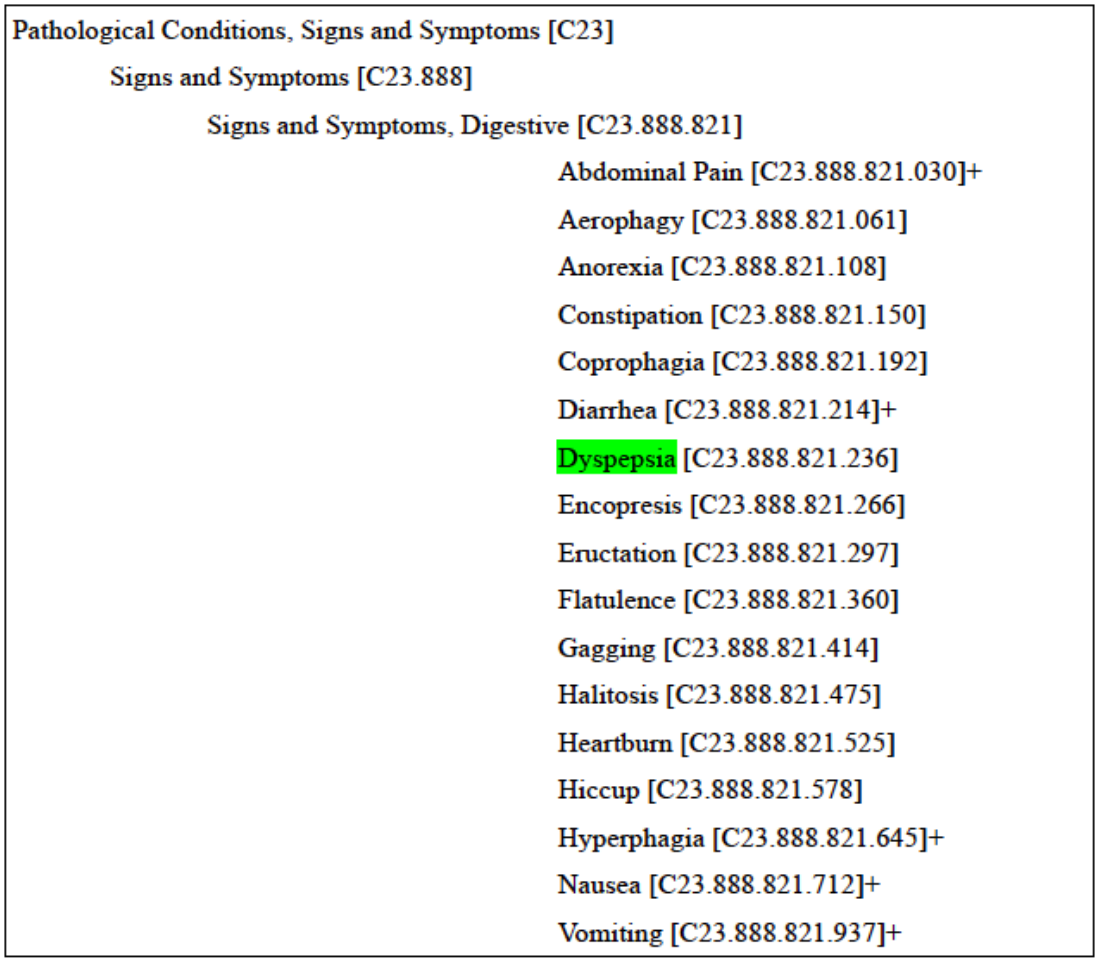


Figure 5.2: Extract from the MesH hierarchy for dyspepsia.

The initial explorations identified the most successful retrieval approach as searching for the compared entities and the basis of the comparison by searching for each element in the title and abstract section of MEDLINE® abstracts and combining the results, so that only abstracts which contain both entities and the basis of the comparison were retrieved.

[2] http://wordnet.princeton.edu/

5.2 Judging the Relevance of MEDLINE® Abstracts

A initial experiment was carried out to evaluate the relevance of the abstracts retrieved from MEDLINE® via Ovid® using the strategies described in the previous section. The evaluation was done in terms of precision. Recall could not be assessed because the number of relevant abstracts that could have been retrieved is unknown. The experimental subjects were eight 4th year medical students, who evaluated the abstracts retrieved for twelve clinical comparison questions in which two drugs were compared to each other with respect to a particular attribute. The questions differ in syntactic structure, but they all contain comparisons of two drugs. Table 5.1 shows the list of questions which are a subset of Corpus 1 described in Section 4.1.

1	Is there any evidence to suggest that torasemide is better than furosemide as a diuretic?
2	Is lansoprazole better than omeprazole in treating dyspepsia?
3	Are there any studies comparing topical diclofenac gel with ibuprofen gel?
4	Effectiveness of Decapeptyl in treatment of prostate cancer in comparison to Zoladex?
5	Which is more effective ibuprofen or diclofenac for arthritis pain for pain relief?
6	Is calcium citrate better absorbed and a more effective treatment for osteoporosis than calcium carbonate?
7	Have any studies directly compared the effects of Pioglitazone and Rosiglitazone on the liver?
8	Is Famvir (famciclovir) better than acyclovir for Herpes zoster?
9	Is it true that men on captopril have a better quality of life than men on enalapril?
10	What is the first choice for Type 2 diabetes patients: sulphonylurea or metformin?
11	Is there any evidence as to which is more effective at preventing malaria: Malarone or Doxycyline?
12	In conjunctivitis which is better chloramphenicol or fucithalmic eye drops?

Table 5.1: Questions used in the experiment.

The material presented to the medical students in the experiment was created as follows: The drug names and the basis of the comparison from the natural language questions were manually mapped to their corresponding MeSH terms and used to retrieve abstracts via OVID® using the final strategy described in Section5.1.

For any question, the maximum number of abstracts given to the student judges was 15, comprising up-to-15 of the most recent abstracts. In total, each judge evaluated 103 abstracts. Each abstract was assigned by each judge into one of three categories, based on the criteria given after the category label:

1. Relevant: Both drugs from the question or their generic names are mentioned in the abstracts, the drugs are directly compared to each other, and the disease or the attribute with respect to which they are being compared is also mentioned and is the

same as stated in the question or synonymous to it (e.g., heartburn and dyspepsia would both count as right because they are closely related).
2. Not Relevant: The drugs or their generic names are not mentioned in the abstract, the drugs are not compared, and/or the disease or the attribute with respect to which they are being compared is wrong (as in different from what is stated in the question, e.g. effect on blood pressure instead of use as a painkiller).
3. Somewhat Relevant: The drugs or their generic names are mentioned but there are no single sentences indicating a comparison between them or the disease is not mentioned. If the wrong disease is mentioned, the abstract should be labeled not relevant.

The judges were also asked to explain the reason for their choice of labels. The inter-annotator agreement between the judges was computed using a variant of the kappa statistic for multiple annotators (Fleiss, 1971). The null hypothesis was rejected and it was ensured that the observed agreement is not accidental.

Overall inter-annotator agreement for all three categories measured by the kappa statistic was moderate at 0.58 for a total of 103 judgements. 47 judgements were in the "somewhat relevant" category. If annotator agreement is only assessed on the remaining 56 judgements from the two categories "relevant" and "not relevant", kappa is 0.97, which represents almost perfect agreement.

5.3 Results and Discussion

Graph 5.3 shows the percentage of abstracts that were judged relevant by the eight judges for each question. The numbers of retrieved abstracts for each question were: 15 abstracts for Question 1, 5, 8 and 10, 9 abstracts for question 7 and 11, 7 abstracts for Question 2, 5 abstracts for Question 9, 4 abstracts for Question 6 and 12, 3 abstracts for Question 3 and 2 abstracts for Question 4.

Question 1, 9 and 12 show a very high percentage of relevant abstracts (a precision of 73%, 80% and 100% respectively), whereas no relevant abstracts were retrieved for questions 4, 5 and 11, and only one relevant abstract (out of 15) for question 10. An abstract was considered relevant when at least five of the eight judges considered it relevant.

The main sources for these disparate results based on both the explanations given by the student judges and discussions with our medical librarian are the following:

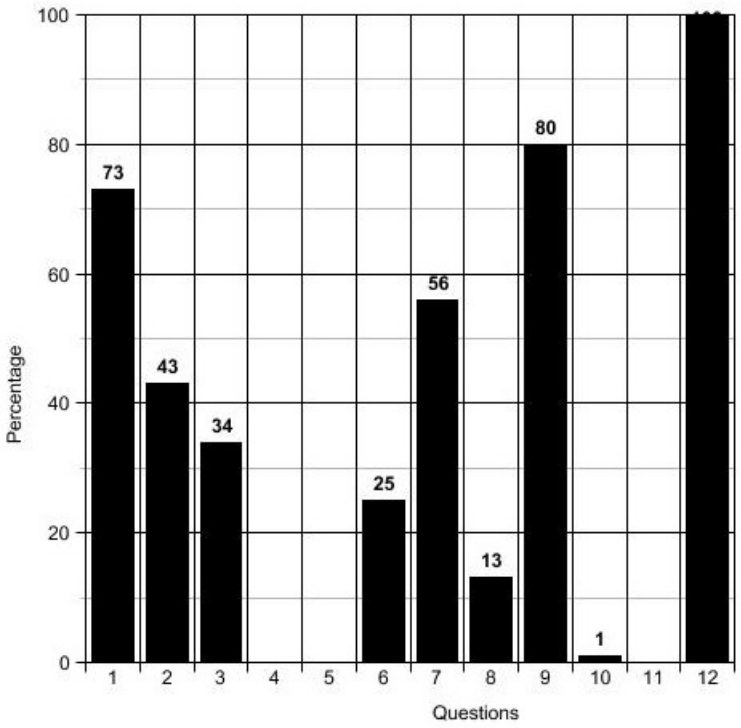


Figure 5.3: Percentage of abstracts judged relevant by the majority of the judges for each of the twelve questions. The label on the top of each bar is the actual percentage.

Approximately 30% (31 of 103) of the abstracts were labeled "not relevant" by the judges because they lacked any direct evidence of a comparison e.g., at least one sentence that explicitly compares the two drugs in question, even though the drugs are mentioned in the abstract and the study is a comparative study (as indicated in its MeSH indices). This is illustrated in Example (5.2) for Question 1, which shows the three sentences from one of the abstracts retrieved for Question 1 that explicitly mention the two drugs:

(5.2) Piretanide and **furosemide** have a constant extrarenal elimination and thus accumulate in renal failure.[...] Elimination of **torasemide** is independent of its renal excretion. Thus in renal failure, **torasemide** is the only loop diuretic in which the plasma concentration is strictly dose dependent.

About 10% (10) of the abstracts were judged to be irrelevant because the drugs were compared as part of a treatment regime in combination with other drugs, as in Abstract 4 for Question 6 in which calcium citrate and calcium carbonate are compared co-administered with different preparations of sodium fluoride. In two cases (2% of the abstracts), doses of a given drug were compared against other dosages instead of the

drugs themselves, e.g., 30 mg lansoprazole versus 20mg omeprazole.

A major factor for "not relevant" judgements was the time frame. This was relevant when retrieving abstracts about well-established drugs that have been in existence for a long time, such as ibuprofen or diclofenac. All but one of the 18 abstracts retrieved for the two questions about these two drugs were irrelevant, even though the two drugs were explicitly mentioned in the abstract. The problem is that they were grouped together as conventional non-steroidal anti-inflammatory drugs (NSAIDs) and compared to newer NSAIDs or different pain medication. Such abstracts could only be excluded by analysing the abstracts themselves.

The final source of "not relevant" judgements was a problem with the judges and not with the abstracts. In Question 2 regarding dyspepsia, two out of seven abstracts were judged irrelevant because the drugs were not explicitly compared regarding dyspepsia but only regarding H. pylori, which is one of the possible causes for dyspepsia. Also abstracts retrieved for Question 7 about the effect on lipid profiles were wrongly categorised by roughly a third of the judges as not being relevant to the liver.

The two main problems discovered during the experiment concern abstracts lacking sentences in which the drugs are directly compared to each other and the retrieval of irrelevant abstracts for well-established drugs, which are used as a reference for comparing newer drugs to, instead of containing direct comparisons of the drugs in question.

5.4 Summary

In this experiment, the focus was on the compared entities and the basis of the comparison. Using a publication type limit decreases the size of the set of retrieved abstracts and potentially increases relevance but can prove problematic for less-studied drugs, for which only a small number of abstracts exists in the first place.

The experiment has shown that searching for the drugs, the basis of the comparison, and studies of the publication type *comparative study* is a first step towards retrieving abstracts that can serve as answer candidates for clinical comparison questions, but it also insufficient to guarantee the relevance of the retrieved abstracts, and the *comparative study* limit can be too restrictive.

These findings have been taken into account for creating the automated system for the process of identifying and extracting the elements of a comparison question as well as the process of retrieving MEDLINE® abstracts described in Chapter 6. Because the

criteria that make a document relevant can be either applied during the abstract retrieval process or during a post-retrieval ranking phase, it was decided to use a less restricted retrieval strategy focusing only on the drug names and the basis of the comparison, which works well for both well and less-studied drugs, to achieve the best possible retrieval and to use post-retrieval ranking, which will be introduced in Chapter 8, to ensure that the most relevant abstracts are displayed in the top of the result set so that a bigger result set does not present a problem.

Chapter 6

Automatic Query Construction and Abstract Retrieval

In this chapter the implementation and evaluation of the query construction and abstract retrieval component of RetroRank is described. The findings of the initial abstract retrieval experiment described in Chapter 5 were taken into account for system development. Section 6.1 describes the question processing with LT-TTT2. Section 6.2 describes the Java pipeline used for keyword extraction and abstract retrieval. Section 6.3 shows an evaluation of the retrieval system. The retrieval component of RetroRank was developed on Corpus 1 described in Section 4.1 and evaluated on Corpus 2 described in Section 4.2.

6.1 LT- TTT2

The first step in creating a system for the automatic retrieval of relevant MEDLINE® abstracts involved processing the plain text questions with LT-TTT2 to automatically mark-up the diseases and drugs in the questions with disease and drug XML tags. LT-TTT2[1] performs shallow linguistic processing. e.g., turning the text into XML format, tokenisation, sentence splitting, part-of-speech tagging, lemmatizing, chunking and rule-based named entity recognition (NER).

LT-TTT2 contains different lexicons to perform rule-based NER for people, locations, dates and common English words. To tailor LT-TTT2 to medical QA, a disease lexicon, a drug lexicon for recognising existing drugs, and a drug stem lexicon for recognising new drugs were added. Also grammars and rules were written for the

[1] http://www.ltg.ed.ac.uk/software/lt-ttt2

lexicons to be able to perform rule-based NER on clinical comparison questions.

The disease lexicon was created from the UMLS Metathesaurus by modifying the UMLS disease file at `ftp://ftp.ebi.ac.uk/pub/software/textmining/corpora/diseases/`. The original file contained 276,155 entries in the format displayed in Figure 6.1.

```
<?xml version='1.0' encoding='UTF-8'?>
<mwt>
<template><z:e sem="disease" ids="%1" disease_type="%2">%0</z:e></template>
<t p1="C1392289" p2="Disease or Syndrome">xanthoma; cerebrotendinous</t>
<t p1="C0013288,C0032763" p2="Disease or Syndrome">Jejunal syndrome</t>
<t p1="C0153028" p2="Disease or Syndrome">iridocyclitis; herpes virus, zoster (etiology)
</t>
<t p1="C0030793" p2="Neoplastic Process">Neoplasm of pelvis</t>
<t p1="C1456118" p2="Disease or Syndrome">Venous embolism and thrombosis of deep
vessels of lower extremity</t>
<t p1="C1691779" p2="Disease or Syndrome">HEARING DISORDER, COCHLEAR</t>
<t p1="C0025292" p2="Disease or Syndrome">Hemophilus; meningitis</t>
<t p1="C1706492" p2="Neoplastic Process">Other neoplastic lymphatic and hematopoietic
tissues</t>
```

Figure 6.1: UMLS disease file.

The first column "p1=" lists the UMLS Metathesaurus concepts related to the disease. Alternate names for the same concept (synonyms, lexical variants, and translations) are linked together in the Metathesaurus, e.g., *Jejunal syndrome* in line 5 is associated with another Metathesaurus entry, which is indicated by its two IDs ("C0013288" and "C0032763") in the first column. The second column "p2=" shows the semantic type of a disease and the disease name. For implementing a disease lexicon for lexical lookup only the disease names are relevant.

After extracting the disease names and removing duplicate entries, 229,698 entries were left. The remaining disease names were put into alphabetical order, converted to lower case and annotated with LT-TTT2 lexicon specific XML tags. An example extract of the final disease lexicon is shown in Table 6.1.

The lexicon was created as an expanded lookup list including all lexical variants rather than normalized terms to ensure maximum matching capacity during lexical lookup. Every entry in the lexicon is treated as a separate name and RetroRank performs phrase matching against the longest possible entry when recognising words in the clinical questions as possible diseases, e.g., "allergic rhinitis" rather than "rhinitis" or "chronic obstructive pulmonary disease" rather than "pulmonary disease" or just "disease". This method has the limitation that only exact matches are found but it still achieves very high accuracy (96.67% diseases in the evaluation questions were correctly recognised in the first round) because of the very large size of the lexicon and

```
<lexicon>
<lex word="abdomen tumor"/>
<lex word="abdomen tumor desmoid type"/>
<lex word="abdomen tumour"/>
<lex word="abdomen tumour desmoid type"/>
<lex word="abdomen wall abscess"/>
<lex word="abdomen wall abscess recurrent"/>
<lex word="abdomen wall tumor myxoid type"/>
<lex word="abdomen wall tumour myxoid type"/>
<lex word="abdomen wound infection"/>
<lex word="abdominal abscess"/>
<lex word="abdominal abscesses"/>
</lexicon>
```

Table 6.1: Extract from the LT-TTT2 disease lexicon.

the fact that it includes different spelling variants as well as singular and plural forms of each disease.

The drug lexicon was created from the Internet Drug Index RxList[2]. A total of 2,945 drug names was collected from the index on the 21st of September 2009. The RxList was founded by pharmacists in 1995 and offers detailed and current information on drugs from reliable sources like the FDA. The list was later supplemented by using the Merck Drug Names: Generic and Trade index[3] to ensure a more complete coverage. Drugs sold by different companies may have several trade names and it is important to have both the generic name and the trade names in a comprehensive drug lexicon. A total of 2,072 generic and brand drug names was collected on the 5th of November 2009 from the Merck index. The two drug lists were merged and after the removal of duplicates, 3,299 entries remained in the final drug lexicon. Each entry represents a drug name and each word in the clinical questions is matched against the entries in the drug lexicon as done in the matching process used for the recognition of diseases. An extract from the drug lexicon is shown in Table 6.2.

A link between a generic drug name and its possible brand names in the drug lexicon was not made in the current research. During the abstract retrieval process, each drug is mapped to its corresponding MeSH term, which is automatically "exploded" analogously to the example described for diseases in Section 5.2. This means that

[2] http://www.rxlist.com/drugs/alpha_z.htm

[3] http://www.merck.com/mmhe/appendixes/ap3/ap3a.html

<lexicon>
<lex word="A - Methapred"/>
<lex word="Abacavir"/>
<lex word="Abarelix"/>
<lex word="Abatacept"/>
</lexicon>

Table 6.2: Extract from the LT-TTT2 drug lexicon.

RetroRank relies on the information included in the MeSH hierarchy for finding relevant related drugs. For example, the MeSH hierarchy for the drug "Sumatriptan"[4] includes the fact that it belongs to the sulfonamide group and lists the other generic drugs that belong to the same group. However, "Sumatriptan" is only associated with one of its brand names "Imigran" in the MeSH data. Its second brand name, "Imitrex", is not associated with "Sumatriptan" in the MeSH data.

Because a drug lexicon can only include drug names that are known at the time the lexicon was created, a solution for recognizing drugs that have been missed because they are not in the lexicon, e.g., new drugs, was implemented using WHO drug word stems[5]. Drug word stems are parts of a drug name such as "-sartan" in the drug "losartan" or "-prazole" in the drug name "omeprazole". Research by Segurabedmar et al. (2008) has shown that the use of word stems based on the nomenclature rules recommended by the WHO International Nonproprietary Names (INNs) Program[6] helps in identifying drugs that have been missed by other methods. INNs are generic, unique names that are globally recognized. These names are built using word stems indicating certain pharmaceutical substances and it is possible to recognize new drug names because the stems are part of their name. A extract of the most recent 2009 list of the words stems and associated substances is published by the WHO and can be found in Appendix E.

There are prefix, postfix (suffix) and infix stems and a grammar was written to recognise each type. For suffix recognition each word in the clinical question is treated as sequence of characters and regular expressions are used to check for a sequence of characters followed by an entry in the suffix lexicon. If a word ends in a suffix from

[4] http://www.nlm.nih.gov/cgi/mesh/2010/MB_cgi?mode=&term=Sumatriptan&field=entry

[5] www.who.int/medicines/services/inn/StemBook2009.pdf

[6] http://www.who.int/medicines/services/inn/en/

the suffix lexicon, it is marked up in the following way indicating that the word is a drug and contains a certain suffix:

(6.1) <drug suffix='gatran'>ximelagatran</drug>

For prefixes and infixes, it is checked whether a word treated as a sequence of characters matches a stem found on the prefix or infix list and if this happens a word gets marked up as a drug and the prefix or infix is shown as can be seen in Example (6.2).

(6.2) <drug prefix='fos'>fosamax</drug>

Table 6.3 shows an extract of the LT-TTT2 WHO suffix stem lexicon created for RetroRank.

<lexicon name="stems">
<lex word="abine"></lex>
<lex word="ac"></lex>
<lex word="acetam"></lex>
<lex word="actide"></lex>
<lex word="adol"></lex>
<lex word="adom"></lex>
<lex word="afenone"></lex>

Table 6.3: Extract from the LT-TTT2 WHO suffix stem lexicon.

A post-process checks whether the marked-up drugs are included in a lexicon of ordinary English words and if so the drug tag gets removed. This is important to filter out false positives for drug names caused by the WHO stem recognition process in cases were the stems are ambiguous and can be part of ordinary English words as well as drug names, e.g., the stem "-al" is very general and occurs in many ordinary English words such as "marginal" or "temporal". Another example are words such as "suggest" or "restriction". The word "suggest" contains the infix "gest" and the word "restriction" contains the infix "estr".

The accuracy of the rule-based name entity disease recogniser is 96.67% on Corpus 2 and 98.5% on Corpus 3. The accuracy of the rule-based named entity drug recogniser using solely the drug lexicon is 95% for Corpus 2 and 81% for Corpus 3. Using WHO stem recognition in addition to the drug lexicon increases the accuracy for drug recognition to 98.75% for Corpus 2 and to 98.33% for Corpus 3. A total of 87 drugs

were identified using the WHO rules of which 13 were not included in the drug lexicon.

A system diagram highlighting the main steps for question processing with LT-TTT2 is in shown in Figure 6.2.

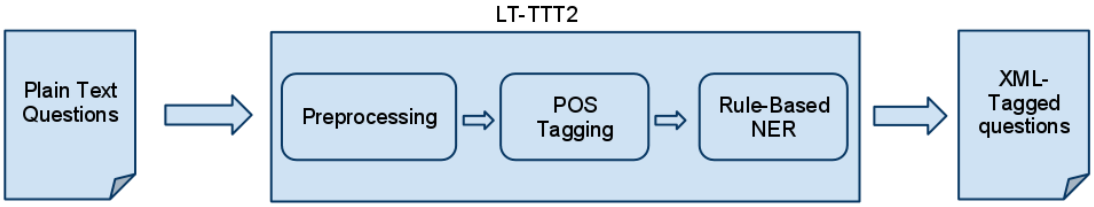


Figure 6.2: System diagram of the main processing steps of LT-TTT2.

A list of plain text questions is loaded into LT-TTT2. The questions are then turned into XML format, tokenised, lemmatised, POS-tagged and NER-tagged. During the NER step drugs and diseases are bracketed with drug and disease tags. The LT-TTT2 output before further post-processing for the question (6.3) is shown in Example(6.4).

(6.3) Is eletriptan more effective and at least as safe as sumatriptan for the treatment of acute migraine headache?

(6.4)

<s id="s7"><w l="be" pws="yes" c="w" id="w784" p="VBZ">Is</w>< w l="eletriptan" pws="yes" c="w" id="w787" p="NN" **drug="true"**>**eletriptan**</w><w pws="yes" c="w" id="w798" p="RBR">more</w> <w pws="yes" c="w" id="w803" p="JJ">effective</w> <w pws="yes" c="w" id="w813" p="CC">and </w><w pws="yes" c="w" id="w817" p="IN">at</w<w pws="yes" c="w" id= "w820"p="JJS">least</w><w pws="yes" c="w" id="w826" p="RB">as</w> <w pws="yes" c="w" id="w829" p="JJ">safe</w><w pws="yes" c="w" id="w834" p="IN">as</w><w l="sumatriptan" pws="yes" c="w" id= "w837" p="NN" **drug="true"**>**sumatriptan**</w> <w pws="yes" c="w" id="w849" p="IN">for</w><w pws="yes" c="w" id="w853" p="DT">the</w><w vstem="treat" l="treat-ment" pws="yes" c="w" id="w857" p="NN">treatment</w><w pws="yes" c="w" id="w867" p="IN">of</w><w pws="yes" c="w" id="w870"p="JJ">acute</w><w **disease="true" l="migraine"** pws="yes" c="w" id="w876" p="NN">migraine</w><w **disease="true" l="headache"** pws="yes" c="w" id="w885" p="NN">headache</w><wc=".” sb="true" pws="no" id="w893" p=".">?</w></s>

Example (6.4) contains the following fields:

** ... **: Each sentence is wrapped with paragraph elements which are required for the tokenisation step.

<**s**>**...**</**s**> and <**w....**>**...**</**w**>: The tokeniser segments the character data content in <p> elements into <w> (word) elements, and identifies sentences which are wrapped into <s> elements.

"id": "id" is an attribute of <w> or <s> elements. For <w> elements it shows a unique id for each word. For <s> elements it shows a sequential number of the sentences that are processed.

"l": The "l" attribute shows the lemma of a word. Only word types that can be lemmatised have this element.

"c": The "c" attribute is used to encode the word type which is only relevant for internal purposes. *c="abbr"* would indicate the word is an abbreviation, for example.

"pws": The "pws" attribute encodes white space. *pws="yes"* means that there is a white space between a word and the word before.

"sb": The "sb" attribute indicates a final full stop at the end of a sentence to differentiate it from possible internal full stops. The "pws" and "sb" attributes are relevant for the NER component.

"p": "p" encodes the POS tag of a word, e.g., NN for a proper noun, JJ for an adjective, IN for a preposition, etc. The POS tags are Penn Treebank POS tags.

"drug" and **"disease"**: These attributes encode drug and disease names within the NER component. They get the value "true" when a drug or disease is recognised and are not displayed when their value is "false" because a word is not recognized as a drug or disease. In Example (6.4) the drugs *eletriptan* and *sumatriptan* are encoded with drug tags and the disease migraine headache is marked up with disease tags.

The experiment in Chapter 5 showed that the drug names (which are the compared entities) and the disease name (the basis of the comparison) are a good starting point for retrieving potentially relevant abstracts. Also, the performance of Entrez PubMed (cf. Section 2.3.3) shows that sending additional terms, such as "safe" or "effective" in the case of Example (6.3), has detrimental effects on accuracy during abstract retrieval.

Therefore all additional markup except for the sentence ID, the drug and the disease tags is removed during a second post-processing step with XLST stylesheets. The final output for question (6.3) is shown in Example (6.5). If a drug is recognised by both the initial lexical lookup in the drug lexicon and by checking for the WHO word stems, the drug tags get nested as can be seen in Example (6.5) for the drugs *eletriptan* and *sumatriptan*. This only serves an informative purpose and does not make a difference for later keyword extraction. Each part of a disease is wrapped in its own disease tag for the reason that keywords in the initial E-Utilities ESearch query are treated as single keywords and not as compound keywords e.g., *migraine headache* needs to be split up into *migraine* + *headache* instead of "*migraine headache*".

(6.5)

<s id="s7">Is <drug><drug suffix="triptan">eletriptan</drug> </drug> more effective and at least as safe as <drug><drug suffix="triptan">sumatriptan</drug></drug> for the treatment of acute <disease>migraine</disease><disease> headache</disease></s>

The XML-tagged question file is then passed on to the Java pipeline for keyword extraction and abstract retrieval.

6.2 Java Pipeline

The second program component is a Java pipeline that processes the tagged questions generated by LT-TTT2 and retrieves MEDLINE® abstracts. Figure 6.3 shows the system diagram for the Java pipeline.

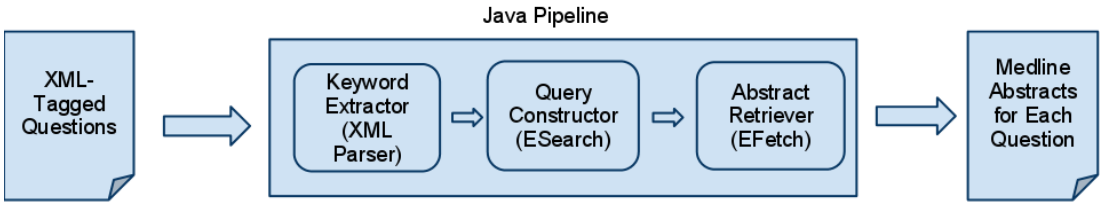


Figure 6.3: System diagram of the Java pipeline.

At the start the XML-tagged questions generated with LT-TTT2 are loaded into the Java pipeline. From Example (6.5) the Keyword Extractor extracts the drug and disease entities (*eletriptan*, *sumatriptan*, *migraine* and *headache*). The Query Constructor translates these keywords into an E-Utilies ESearch query URL (see Section 2.3.3) to

search for associated PubMed IDs (PMIDs) in PubMed. Example (6.6) shows an example query URL for the question keywords. The example query URL in (6.6) does not include any additional search limits such as restricting the search to humans or a certain date range. The ESearch query will retrieve a list of PMIDs of MEDLINE® abstracts that are relevant to the query.

(6.6)

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed &term=eletriptan +AND+sumatriptan+AND+migraine+AND+headache

The Abstract Retriever constructs an E-Utilities EFetch query that retrieves the MEDLINE® abstracts from PubMed for the PMIDs retrieved by the ESearch query. The MEDLINE® abstracts are saved as text files. An example EFetch query URL is shown in Example (6.7). The first three PMIDs retrieved by the ESearch query from Example (6.6) are shown for illustration purposes in the example.

(6.7)

```
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=
pubmed&id=22272067, 21593190, 21375444,...
```

The next section shows the performance of RetroRank on 30 clinical comparison questions from Corpus 2 and compares it to two existing abstract retrieval systems.

6.3 System Evaluation using Corpus 2

The performance of the retrieval component of RetroRank was assessed in terms of accuracy using 30 medication comparison questions taken from Corpus 2, which was collected from the Essential Evidence Plus POEM archive introduced in Section 3.2.2 (see Appendix D for the full question list). This enabled RetroRank to be evaluated using the same criteria used in the *ask*MEDLINE evaluation described in Section 2.5.4. The 30 POEM questions were first processed with LT-TTT2 and the output file was then used for retrieving relevant MEDLINE® abstracts for each question with the Java Pipeline.

RetroRank produces two different abstract result sets. This is done to evaluate the impact of a publication type limit filter and to assess whether an additional filter

is a good idea. One set ("All Studies") is retrieved without applying any publication type limits. The other is limited to the publication type *Comparative Studies* ("Comp Studies"). Table 6.4 shows the total number of MEDLINE® abstracts retrieved for the two result sets. Later analysis will show that the study type limit is too restrictive to allow for optimal performance.

	Number of abstracts retrieved
"All Studies"	3425
"Comp Studies"	883

Table 6.4: Number of retrieved abstracts for the two result sets.

	Number of abstracts retrieved
*ask*MEDLINE	2001
Entrez PubMed	35778

Table 6.5: Number of retrieved abstracts for *ask*MEDLINE and Entrez PubMed

*ask*MEDLINE (Section 2.5.4) and Entrez PubMed (Section 2.3.3) are both assessed on the same set of questions to determine what is gained by the approach to specialized retrieval for comparison questions in this research. The total number of retrieved abstracts for each of the two systems in shown in Table 6.5. Even though Entrez PubMed retrieved about ten times as many abstracts as RetroRank for "All studies", the accuracy is very low as will be shown in the analysis below.

The accuracy of all three retrieval systems (RetroRank, *ask*MEDLINE and Entrez PubMED) was determined using the "gold standard" reference from Corpus 2 for the 30 comparison questions. If the retrieved MEDLINE® abstract set for a question contains the "gold standard" POEM reference, the abstract retrieval process for that question is considered successful. Table 6.6 shows the results for all three systems.

RetroRank ("All Studies") achieved an accuracy of 70% for the condition without a publication type limit, which corresponds to the condition for the other systems which do not include a publication type limit. This means for 21 out of the 30 questions the "gold standard" reference was among the set of retrieved MEDLINE® abstracts. *ask*MEDLINE found 53.3% of the cited articles while the accuracy of Entrez PubMed is just 30%. The statistical significance of the results was determined using

	"Gold-standard" retrieved	"Gold-standard" not retrieved	No reference retrieved
RetroRank ("All Studies")	21/30 (70%)	7/30 (23.3%)	2/30 (6.7%)
RetroRank ("Comp Studies")	16/30 (53.3%)	10/30 (33.3%)	4/30 (13.3%)
***ask*MEDLINE**	16/30 (53.3%)	13/30 (43.3%)	1/30 (3.3%)
Entrez PubMed	9/30 (30%)	12/30 (40%)	9/30 (30%)

Table 6.6: Accuracy for the comparison questions in Corpus 2.

the Wilcoxon signed-rank test. The performance of RetroRank was significantly better than the performance of Entrez PubMed ($p < 0.005$). While the difference in performance between RetroRank and *ask*MEDLINE was not found statistically significant ($p > 0.1$), it still represents a marked improvement.

The "gold standard" reference was not among the retrieved abstracts for 23.3% (7/30) of the questions for RetroRank ("All Studies"), 43.3% (13/30) of the questions submitted to *ask*MEDLINE and 40% (12/30) questions submitted to Entrez PubMed. The overall retrieval failure for RetroRank ("All Studies") was 6.7% compared to 3.3% for *ask*MEDLINE and 30% for Entrez PubMed.

The accuracy of 53.3% of *ask*MEDLINE on the set of 30 clinical comparison questions was lower than the accuracy of 62.1% reported by Fontelo et al. (2005) for their set of POEM questions, which might be due to the fact that the *ask*MEDLINE algorithm was not developed for comparison questions. On the other hand, Entrez PubMed performed better here than in Fontelo's study where it only achieved 13.7% accuracy, although it still achieves the lowest accuracy of all systems.

The accuracy of RetroRank for "Comp studies" is 53.3% which is 16.7% lower than the accuracy of "All Studies". This indicates that the inclusion of the publication type "comparative study" during retrieval has detrimental effects. The retrieval failure for RetroRank ("Comp Studies") is 13.3% which is another indicator that this publication type limit during retrieval is too restrictive.

6.4 Error Analysis

In three cases, RetroRank fails to retrieve the "gold standard" citation whereas *ask*MEDLINE does. In one case the "gold standard" reference was retrieved by both *ask*MEDLINE and Entrez PubMed. Example (6.8) is a question for which both of the other systems retrieved the correct reference.

(6.8) In adults or children with moderate to severe atopic dermatitis, is either tacrolimus (Protopic) or pimecrolimus (Elidel) more effective than topical corticosteroids? (Question 2 in Corpus 2)

The query terms RetroRank used were: *tacrolimus*, *Protopic*, *pimecrolimus*, *Elidel*, *corticosteroids*, *atopic*, *dermatitis*, and the PubMed limit *Humans*, restricting the results to studies about humans. *ask*MEDLINE retrieved the correct abstract during the 1st retrieval round. This can be determined because the retrieval process terminates when the retrieval count is between 1 and 50,000 and a total of 45 abstracts was retrieved during this questions. The *ask*MEDLINE algorithm is detailed in Section 2.5.4. However, too many elements of the *ask*MEDLINE algorithm are not specified to make an informed guess about what element lead to successful abstract retrieval, e.g., the content of the personal stop word list and the content of the MeSH backup vocabulary is not published.

Entrez PubMed successfully retrieved the "gold standard" abstract by mainly using the same query terms used in RetroRank. However, the query included the additional terms: *moderate*, *severe*, *effective* and *topical* as well as the limits *adult(s)* and *child(ren)*, because all words that are not PubMed stop words get included in the search. Using *adult(s)* and *child(ren)* instead of the limit *humans* does not lead to retrieving the "gold standard" citation. The terms *moderate*, *severe*, *effective* and *topical* are not MeSH terms but occur in the "gold standard" abstract and one or more of them might have led to the successful retrieval. The overall performance of Entrez PubMed with many query terms is rather low, however, and there are several cases where too many query terms led to retrieval failure. Therefore this approach has not been adopted for RetroRank.

Example (6.9) is a question for which *ask*MEDLINE retrieved the "gold standard" reference but neither RetroRank nor Entrez PubMed did.

(6.9) Is ximelagatran as effective as warfarin in preventing stroke in patients with nonvalvular atrial fibrillation? (Question 15 in Corpus 2)

RetroRank used the following query terms: *ximelagatran*, *warfarin*, *stroke*, *nonvalvular*, *atrial*, *fibrillation* which all occur in the reference citation shown in Example (6.10)

(6.10) Executive Steering Committee on behalf of the SPORTIF III Investigators.

> Stroke prevention with the oral direct thrombin inhibitor ximelagatran compared with warfarin in patients with non-valvular atrial fibrillation (SPORTIF III): randomised controlled trial. Lancet 2003; 362: 1691-98.

The only real difference between the query terms contained in the "gold standard" reference and in the reference about the same study retrieved by RetroRank (see Figure 6.4) is the spelling of the term *nonvalvular* (*non-valvular* in the reference citation). The reference retrieved by RetroRank refers to the same trial but is not the "gold standard" reference. Although the associated abstract actually answers the question even though it is not the reference citation, a trial with a different spelling of *nonvalvular* makes the crucial difference in retrieving the "gold standard" abstract, and shows that it would be beneficial to search for different spellings of the query terms. This will be further discussed in Chapter 9. Entrez PubMED used the same query terms RetroRank used including the same spelling of *nonvalvular* and also failed to retrieve the correct reference.

```
PMID: 12974871 [PubMed - indexed for MEDLINE]

22: Am Heart J. 2003 Sep;146(3):431-8.

Ximelagatran compared with warfarin for prevention of thromboembolism in
patients with nonvalvular atrial fibrillation: Rationale, objectives, and design
of a pair of clinical studies and baseline patient characteristics (SPORTIF III
and V).

Halperin JL; Executive Steering Committee, SPORTIF III and V Study
Investigators.
```

Figure 6.4: Gold standard reference for Question 15.

Example (6.11) is the third question RetroRank failed to retrieve the "gold-standard" reference for. Entrez PubMED also failed to retrieve the correct reference and retrieved no references at all, whereas RetroRank retrieved ten references of which at least one seems relevant to the question.

(6.11) Is losartan comparable to captopril in CHF? (Question 22 in Corpus 2)

The problem in this case is that *CHF* was not associated with the disease *heart failure* which occurs in the "gold standard" reference because an error occurred during the query translation process of the E-Utilities ESearch module. The query translation shown in Example (6.12) shows that *CHF* was wrongly mapped to "Congest Heart Fail" [Journal] instead of "Congestive Heart Failure" [All Fields].

(6.12)

<QueryTranslation>
("losartan"[MeSH Terms] OR "losartan" [All Fields]) AND ("captopril"[MeSH Terms] OR "captopril" [All Fields]) AND ("Congest Heart Fail"[Journal] OR "chf" [All Fields]) AND "humans"[MeSH Terms] AND 1950/01/01 [EDAT] : 2010/03/31[EDAT]
</QueryTranslation>

The same error occurred during the Entrez PubMED search where the query translation is shown in Example (6.13). Entrez PubMED retrieved no references at all which seems to be due to the inclusion of the term *comparable* which restricts the search too much during retrieval.

(6.13)

("losartan"[MeSH Terms] OR "losartan"[All Fields]) AND comparable [All Fields] AND ("captopril"[MeSH Terms] OR "captopril"[All Fields]) AND ("Congest Heart Fail"[Journal] OR "chf"[All Fields])

6.5 Summary

The evaluation and error analysis of RetroRank has shown that searching for only drugs and diseases achieves an accuracy of 70% on the test corpus, which is a significantly better result than the two other retrieval systems achieved. Additional query terms may prove successful for some queries but lead to worse performance overall as can be see for the accuracy of PubMed. Therefore this approach will not be adopted for abstract retrieval for RetroRank. Limiting the search to the publication type *Comparative Study* during retrieval was found to be too restrictive to allow for good performance and has therefore been abandoned.

In the next chapter different post-retrieval ranking strategies to rerank the retrieved abstracts according to different criteria will be tested and evaluated. The goal is to improve on the recency ordering retrieved by MEDLINE®.

Chapter 7

Post-retrieval Ranking Strategies

This chapter introduces different post-retrieval ranking strategies which will be used to re-rank the MEDLINE® abstracts retrieved for each question in Corpus 2 and Corpus 3 in order to display the most relevant abstracts as high in the result set as possible. During retrieval from MEDLINE®, retrieved abstracts are ordered by recency with the most recent abstract displayed first. This order forms the baseline for the evaluation of the different ranking strategies described in Chapter 8. Because recency does not equate to relevance, post-retrieval ranking is an important feature in a system that is geared towards providing the most relevant abstracts at the top of the result set to enable clinicians to find the most relevant information quickly and reliably.

7.1 Ranking Strategies

7.1.1 Rank by Reverse Chronological Time Order ("recency")

The baseline for the post-retrieval ranking module of RetroRank is the simple default ordering by date as returned by PubMed. The PubMed IDs (PMIDs) are sorted in descending chronological order with the last indexed abstract at the top of the result set. PMIDs are consistent with Entrez Dates (EDAT) but not with Publication Dates (PDAT). That is, they are indexed by the date the citation was added to MEDLINE®, which does not have to be the date the article was published. However, this approach is consistent with the default order displayed in PubMed and is the order encountered by clinicians performing a MEDLINE® search. Therefore it was chosen as a baseline for RetroRank.

7.1.2 Term Frequency ("tf")

Term count, or raw term frequency, refers to the number of times a query term occurs in each retrieved MEDLINE® abstract. The term frequency $(tf_{t,d})$ of term t in document d is defined as the number of times that t occurs in d (Manning and Schütze, 1999). The underlying assumption is that documents with a higher percentage of query terms are more relevant to the query than documents with a lower percentage of query terms. This has also been found during a manual analysis of the "gold standard" abstracts in Corpus 2. However, it is hard to determine how much more relevant a document is when the percentage of query terms goes up because there is no linear relationship between the number of query terms and relevancy. To overcome this problem log-weighted term frequency is used instead of raw term frequency. The log frequency weighting of term t is:

$$w_{t,d} = 1 + \log 10\ \mathrm{tf}_{t,d}$$

The score for a document-query pair equals the sum over all terms t in both the question q and the document d:

$$\text{Score} = \sum_{t \in q \cap d} (1 + \log 10\ \mathrm{tf}_{t,d})$$

If none of the query terms is present in the document the score is 0.

Term frequency only considers how frequent query terms are in a document but not how frequent they are over the whole corpus. To capture this information the next strategy is needed.

7.1.3 Term Frequency/Inverse Document Frequency ("tf/idf")

Rare terms contain more salient information than frequent terms, which is why it is important to remove words from the corpus that are frequent but low in information content such as stop words (Appendix C). A document containing rare terms such as the query terms is likely to be relevant to the query and rare terms should therefore receive a higher weight. To statistically determine the importance of a word to a document in a corpus the tf/idf (term frequency/inverse document frequency) is used. tf/idf

offsets the frequency of a word in a document by the frequency of the word in the corpus.

As mentioned above, term frequency simply refers to the number of times a term appears in a document, which is usually normalized to prevent a bias towards longer documents. The inverse document frequency (idf) is a measure of the general importance of the term (obtained by dividing the total number of documents (N) by the number of documents containing the term (df_t), and then taking the logarithm of that quotient). Idf for a term t is defined in the following equation (Manning and Schütze, 1999)

$$\mathrm{idf}_t = \log 10\, N/\mathrm{df}_t$$

The tf-idf weight of a term is the product of its term frequency (tf) weight and its inverse document frequency (idf) weight (Manning and Schütze, 1999).

$$\mathrm{w}_{t,d} = (1 + \log 10\ \mathrm{tf}_{t,d})\ \mathrm{x}\ \log 10\, N/\mathrm{df}_t$$

The tf-idf weight increases with the number of occurrences of a term within a document and with the rarity of the term in the collection. This means that during post-retrieval ranking documents with a high number of occurences of query terms are ranked higher than documents with a low number of occurences of query terms. While this seems reasonable, it will need to be determined in Chapter 8 how well the strategy performs in terms of ranking the relevant documents higher in the result set.

7.1.4 PercentLast ("Last20%")

A manual analysis of the "gold standard" abstracts from Corpus 2 introduced in Section 4.2 has shown that the last section of the abstract often contains the most relevant information in form of outcome statements for answering drug comparison questions. The last section is called "Conclusion(s)" in structured abstracts but also often exists in the final four or five sentences in unstructured abstracts despite not being officially designated. This finding is consistent with the finding in Demner-Fushman and Lin (2007) that "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 evaluation in Chapter 8 shows how well the strategy performs on the test and development data and in comparison to the results of Demner-Fushman and Lin (2007).

7.1.5 Citation Indices

The following two strategies re-rank abstracts based on an external, human-evaluated source namely citation indices. They set out to prove the assumption that articles that have been frequently cited provide better quality and therefore more relevant clinical evidence. According to Bernstam et al. (2006), citation-based algorithms can identify important articles more effectively than vector-based algorithms such as tf/idf or Boolean queries. This will be shown in the evaluation in Chapter 8.

7.1.5.1 Google Scholar Citation Index ("google")

The first source for citation count information is Google Scholar[1]. Google Scholar is freely available without a subscription and provides a fast and easy-to-use entry point. It covers a wide range of sources including articles, theses, books, and abstracts. Google Scholar ranks documents by weighing the full text of each document, where it was published, who it was written by, as well as how often and how recently it has been cited in other scholarly literature. It automatically extracts citation information from reference lists of scientific documents. The automated approach causes a number of problems ranging from problems with typographical errors in the source documents to parsing errors of the reference due to non-standard formats (Harzing and van der Wal, 2008). It also includes grouping errors for identical citations, which result in duplicates, inflated citation counts and problems with correctly identifying authors (Jasco, 2006, 2008).

To obtain the Google Scholar citation count, RetroRank passes the abstract title and author names to Google Scholar and screenscrapes the citation count. The full implementation is described in Section 8.1.1. Google Scholar does not support searching by PMID, which would have been a more reliable way of identifying the right article and related citation count, because searching for the title and author names sometimes leads to more than one result if an article has been published in different languages or is a reprint for example. RetroRank choses the first result, but it is possible that another result could be the right one. This problem was addressed by using the ISI ci-

[1] http://scholar.google.co.uk/

tation count described in Section 7.1.5.2, which allows searching for articles by PMID and thereby removes any possible ambiguities that arise from searching by title and authors only.

7.1.5.2 ISI citation index ("isi")

The second source for citation count information is the Thomson Reuters ISI Web of Science (WoS)[2], which is the most recognized and trusted database for peer-reviewed journal content. It can only be accessed with a paid license available through an institutional subscription for example. The journal selection process is based on publication standards, quality of the citation data and expert judgements (Garfield, 1990). It only contains citations to articles published in ISI-indexed journals (Roediger, 2006). Unlike Google Scholar it does not include citations to books, book chapters, theses, conference papers or papers in non-ISI journals (Harzing and van der Wal, 2008). While Google Scholar often has better citation coverage than ISI in disciplines like social sciences or humanities, its coverage in the natural and health sciences is less comprehensive and ISI often provides higher citations counts for these disciplines (Harzing and van der Wal, 2008). This becomes evident in the evaluation of the two citation count strategies in the system evaluation in Chapter 8.

The ISI WoS was queried using an OpenURL query searching for each article's PMID as part of Thomson Reuters Links Article Match Retrieval Service[3]. If a match is found, the citation count is returned as well as a link to the full record. The implementation of the query process is described in greater detail in Section 8.1.1.

7.1.6 Cosine Minimum Span Weighting ("msw")

One retrieval method that is widely used in question answering systems is passage-based retrieval, because experience has shown that most questions can be answered with short text segments that span only a sentence or two. This method takes the proximity between query terms into account to rank documents higher where the terms are closer together and lower where the terms are further apart. However, passage-based retrieval has been shown to have poorer performance compared to full-text retrieval in most cases, most likely because a range of parameters, such as passage size or the degree of overlap between passages, have to be adjusted correctly (Monz, 2004).

[2] http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/

[3] http://wokinfo.com/products_tools/products/related/amr/

To make use of the intuition that term proximity is a useful indicator for relevance which dates back to the work of Luhn (1958), while remedying the problems associated with passage-based retrieval, Monz (2004) developed a new proximity-based approach to document retrieval that combined full-document retrieval with proximity information. This approach performed significantly better than full-document retrieval alone. A proximity-based approach does not need to take the parameters regarding passage length into account because it does not check whether terms occur in the same passage but rather looks at the distance between terms regardless of the position in the document.

In this research a new strategy of proximity-based ranking was developed which consists of minimum span weighting (MSW) in combination with cosine document normalization stemming from the original approach by Monz (2004). His approach is based on "the minimal size of a text excerpt that covers all terms that are common between the document and the query, the number of common terms vs. the number of query terms, and the global similarity between the document and the query" (Monz, 2004). All of these components are parametrized to produce a final similarity score.

A minimal matching span is the smallest section of a document that contains all the query terms. The minimal span is used to calculate the similarity between the query terms and a MEDLINE® abstract.

MSW has three determining factors:

1. Document similarity: Document similarity is computed using cosine similarity.

$$similarity = \cos\theta = \frac{A \cdot B}{||A||||B||}$$

   Cosine similarity emphasises the relationship between a query and a document and uses negative and positive information about how related a document is to a query in equal measures. The closer two documents are related to each other the smaller the angle between the vectors. Completely unrelated documents are orthogonal to each other.

   The approach in this research using the cosine similarity differs from that in Monz (2004), who uses Lnu weighting introduced in Buckley et al. (1995), which is more optimal for full text documents with subtopics. Buckley et al.

(1995) showed that cosine similarity would penalise full text documents containing multiple subtopics of which only some could be relevant to the query. The irrelevant subtopics might dominate using the cosine weighting reducing the overall ranking of the document. As MEDLINE® abstracts are short and about a single topic, cosine similarity is a more optimal approach than Lnu.

2. Span size ratio: The number of unique matching query terms in a span of text over the total number of tokens in that text span.

3. Matching term ratio: The number of unique matching terms over the number of unique terms in the query after the removal of stop words.

The MSW score is the sum of the normalized original retrieval status value (RSV) measuring global similarity and the spanning factor which measures local similarity. For a query q, RSV is normalized with respect to the highest retrieval status value for the query as illustrated below (Monz, 2004).

$$RSV_n(q,d) = \frac{RSV(q,d)}{max_d RSV(q,d)}$$

The spanning factor is determined by the span size ratio weighted by α, the matching term ration weighted by β and overall similarity weighted by λ. Monz (2004) determined the values of the three variables empirically using the TREC-9 data collection[4]. The values were set at $\alpha = 1/8$, $\beta = 1$ and $\lambda = 0.4$ (Monz, 2004).

$$\begin{aligned} RSV'(q,d) &= \lambda RSV_n(q,d) \\ +(1-\lambda) &\left(\frac{|q \cap d|}{1 + max(mms) - min(mms)}\right)^{\alpha} \\ &\cdot \left(\frac{|q \cap d|}{|q|}\right)^{\beta} \end{aligned}$$

If $|q \cap d| = 1$ then $RSV(q,d) = RSVn(q,d)$.

Example (7.1) illustrates the MSW algorithm described above on two query terms for a drug comparison question.

[4] http://trec.nist.gov/data/t9_filtering.html

(7.1) Is Ibuprofen better than aspirin?

The query q is $\{ibuprofen, aspirin\}$. Assume there is a MEDLINE® abstract d which matches the query terms at the following positions: pos_d(ibuprofen)=$\{20,31,70\}$ and pos_d(aspirin)=$\{34,80\}$. Then the minimal matching span (mms) = $\{31,34\}$, the span size ratio is $2/(1+34-31) = 0.5$ and the matching term ratio is $2/2$. The spanning factor consists of the last two values and their normalisation factors α and β and is $0.5^{1/8} \cdot 2/2 = 0.917$. If the normalized similarity between the query q and the document d is $n(0 < n \leq 1)$, for example n=0.8 and λ=0.4, the final msw-score for q and d ($RSV(q,d)$) is $0.4 \cdot 0.8 + 0.6 \cdot 0.611 = 0.6866$.

To scale up to more than two query terms, i.e. $\{ibuprofen, aspirin, headache\}$, it is required to find the minimum distance between each term pair ($\{A,B\}$ $\{A,C\}$ and $\{B,C\}$) and select the term position in each pair that is closest as shown in Example (7.2). The numbers refer to the positions at which each term occurs.

(7.2)

Ibuprofen $\{20, \mathbf{31}, 70\}$
Aspirin $\{\mathbf{34}, 80\}$
Headache $\{24, \mathbf{40}\}$
Closest term position for all pairs: $\{31,34,40\} \rightarrow \{31,40\} \rightarrow \{40-31\} = 9$

Monz (2004)'s MSW approach results in identifying minimal spans containing a correct answer in 64.1% to 71.8% of the cases where the document is known to contain a correct answer. The evaluation in this thesis can only determine if a "gold standard" was retrieved and at which position it is ranked in the set of MEDLINE® abstracts.

The cosine MSW strategy implemented in RetroRank only considers the proximity of terms in the final 20% of each abstract because it has been empirically shown that this approach yields the best results on the corpus. The approach was also restricted in the sense of requiring all query terms to occur in a document rather than allowing for a partial match similar to the approach of (Kwok et al., 2000). The results are shown in Chapter 8.

7.1.7 Backing Off ("tf/idf - isi back off")

The seventh re-ranking strategy is tf/idf - isi backing off. If a "gold standard" abstract does not have an ISI citation count and its rank therefore cannot be determinded, the tf/idf - isi strategy backs off to the rank given by the term frequency/inverse document frequency (tf/idf) strategy which was empirically proven to provide the second best overall ranking strategy as far as individual strategies are concerned. Table 7.1 illustrates the substitution of "n/a" produced by the lack of an ISI citation with rank "3" produced by the tf strategy.

"isi" rank	"tf/idf - isi" rank
n/a	**3**

Table 7.1: tf/idf - isi back-off.

This strategy maximises the potential of using the best available individual strategy by defaulting back to the second best strategy if the best strategy is not available.

7.1.8 Expert Voting ("voting")

The last strategy, expert voting, serves as an expert system looking at each of the following six ranking strategies to find the optimal ranking of the "gold standard" documents by majority voting:

1. tf
2. tf/idf
3. Last 20%
4. Google citation count
5. ISI citation count
6. Cosine Minimum Span Weighting

The voting algorithm looks at each strategy in turn and adds the position assigned to each document by each strategy to a list. The ISI citation count, which has proven to be the best individual strategy from the evaluation of the development set, is weighted

at $\lambda = 3$. λ was determined empirically. Each number is zero padded to be in the range between [0, 10,000].

The individual metric rankings for each document are reordered in ascending order as shown in Example (7.3)

(7.3) $[3,10,50,1,2,1] -> [1,1,2,3,10,50]$

The rank positions are padded out to biggest number to the 10^n, in this case 10,000 because the maximum number of retrieved abstracts per document is 5,000. The padding is important because otherwise a combination with lower rank positions would be ranked higher than a combination of higher rank positions as demonstrated in Example (7.4)

(7.4) $[1,1,50,1] -> [11150]$
$[8,8,8,4] -> [4888]$

If the numbers are padded out, the lower ranking is reflected in a higher number as shown in Example (7.5).

(7.5) $[1,1,50,1] -> [1|00001|00001|00050]$
$[8,8,8,4] -> [4|00008|00008|00008]$

The first number which reflects the better ranking is now smaller than the second number and gets precedence.

Example (7.6) illustrates how the documents are reranked with the voting algorithm. For clarity, no padding is used in the example because the numbers are > 10. The numbers in front of the square brackets refers to the document number.

(7.6)

$1[1,1,2,4,5]$
$2[2,3,1,3,1]$
$3[4,5,3,2,2]$
$4[5,4,4,1,3]$
$5[3,2,5,5,4]$

Reordering the rankings for each document in turn leads to Example (7.7).

(7.7)

$1[1,1,2,4,5]$
$2[1,1,2,3,3]$
$3[2,2,3,4,5]$
$4[1,3,4,4,5]$
$5[2,3,4,5,5]$

Reordering the list of documents by their new combined ranking strategy produces Example (7.8).

(7.8)

$2[1,1,2,3,3]$
$1[1,1,2,4,5]$
$4[1,3,4,4,5]$
$3[2,2,3,4,5]$
$5[2,3,4,5,5]$

The new document ranking is determined by the ascending order of the new combined ranking strategy. Example (7.9) shows the final ranking resulting from applying the voting algorithm.

(7.9)

$[2,1,4,3,5]$

Document number 2 is now first, followed by document 1, 4, 3 and 5.

One of the advantages of using the expert voting algorithm is that it is less sensitive to outliers such as shown in Example (7.4), where the fourth strategy ranked the "gold standard" 50th, whereas the other strategies ranked it 1st. Because of the reordering, the strategies that produce lower ranks will get a lower weight and thereby lower priority since they will be given a less significant position in the final document reranking

number.

7.2 Summary

This chapter described different post retrieval ranking strategies that will be tested and evaluated in Chapter 8. The baseline strategy, “recency”, was introduced, followed by the term frequency strategies “tf”, “tf/idf” and “Last20%”. Then the two strategies using different citation indices, “google” and “isi”, were described before introducing the vector-based minimum span approach “msw”. Lastly, two strategies that combine other strategies were described. These are “tf/idf - isi voting” and “voting”. The next chapter shows an evaluation of those strategies and illustrates their strength and weaknesses.

Chapter 8

Post-Retrieval Ranking with RetroRank

In this chapter the implementation and evaluation of the post-retrieval ranking module of RetroRank is described. The post-retrieval ranking component was developed on Corpus 2 (Section 4.2), which was used in Chapter 6 to assess the retrieval accuracy of RetroRank and to compare the retrieval component's performance to the results achieved by *ask*Medline and Entrez PubMed. While Corpus 2 was a valuable resource for the initial system comparison, it has the limitation of only providing one "gold standard" reference per question. Therefore, Corpus 3, which provides multiple "gold standard" references per question was used for the full evaluation of RetroRank's post-retrieval ranking module. Having multiple "gold standards" available allows an evaluation with standard IR metrics and a comparison with the post-retrieval ranking system developed in Demner-Fushman and Lin (2007). It is shown that the automatic *ISI-citation count* based strategies and the *Expert Voting* strategy are a significant improvement over the PubMed recency baseline and are a strong contender to the strategies based on query frames and annotated data developed by Demner-Fushman and Lin (2007).

8.1 Post-retrieval Ranking on Corpus 2

8.1.1 System Implementation

To evaluate the post-retrieval ranking strategies introduced in Chapter 7, the retrieval component of RetroRank was augmented with a Python post-retrieval ranking module. In the augmented version of RetroRank, the retrieval module was extended to store the

MeSH terms in addition to the text and title of the abstracts. The post-retrieval ranking module reranks the MEDLINE® abstracts retrieved for each question in Corpus 2 (c.f. Section 4.2) to obtain an optimal ranking. Figure 8.1 shows the system diagram for RetroRank.

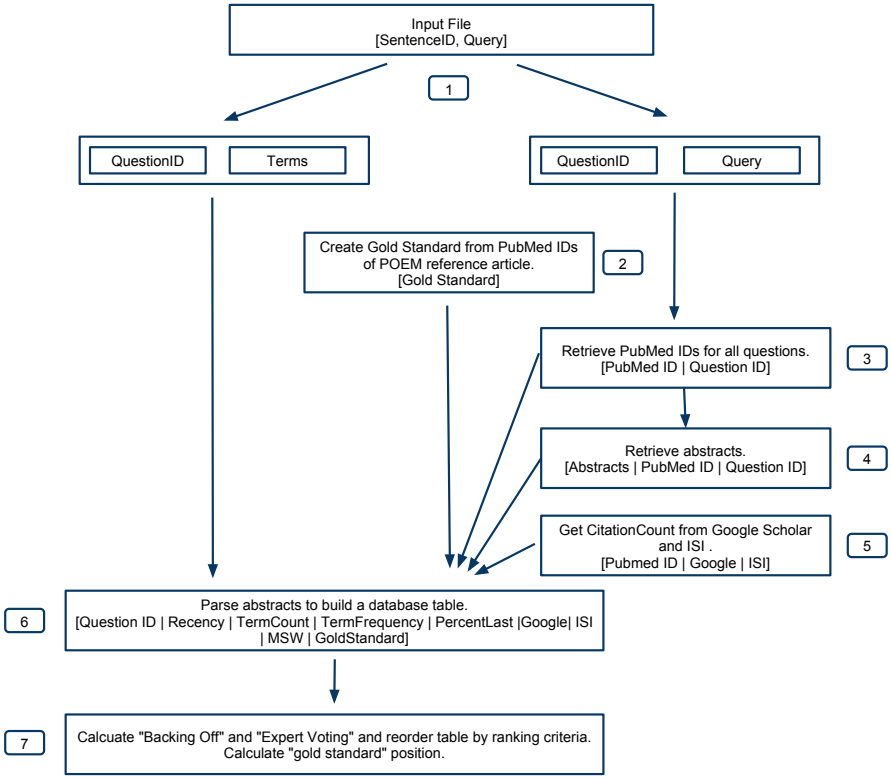


Figure 8.1: System diagram of the RetroRank system.

In the first step, the post-retrieval ranking module parses a .csv file consisting of the questions in Corpus 2 marked-up by LT-TTT2 (c.f. Section 6.1) and their question IDs to create two maps. Maps are lists of key - value pairs. Table 8.1 shows Question 1 from Corpus 2. The rest of the questions can be found in Appendix D.

ID	Question
1	How safe and effective are aspirin and warfarin therapy in the prevention of stroke in patients with atrial fibrillation?

Table 8.1: Extract from the input file for the RetroRank post-retrieval module.

The first map consists of the question ID and the list of query terms (drugs and

diseases). The map will be used for calculating the post-retrieval ranking strategies in a later step. The second map consists of the question ID and the E-Utilities query for the questions, which will be used for retrieving MEDLINE® abstracts and their associated MeSH terms for each question.

The second step is to create a map with the PubMed IDs (PMIDs) of the "gold standard" articles by parsing a .csv file that contains the question ID and a "gold standard" PubMed ID for each question. The "gold standard" articles in Corpus 2 are the articles cited as the reference for each POEM. The references are a good "gold standard" because they have been carefully chosen by experts as the articles which best answer the question.

The third step is to use the map of E-Utilities queries created in the first step to retrieve all PubMed abstracts for the query terms in each question and to create a map of the question IDs and the PMIDs retrieved per question. In the fourth step a map consisting of the PMIDs and their associated abstracts is created.

The fifth step involves downloading the citation information for each abstract from Google Scholar and the ISI Web of Knowledge and storing it in a map containing the PMID and the citation count.

The six step involves parsing the abstracts. During this stage the following processes take place:

- Stop words are removed using the PubMed stop word list (see Appendix C).
- The number of words and keywords is counted.
- The term frequency of the query terms is determined.
- The number of query terms in the last 20% of the abstracts is calculated.
- Cosine Minimum Span Weighting ("msw") is calculated.
- The position of the abstracts ordered by recency as retrieved by PubMed is recorded as a baseline.
- It is checked whether an abstract is a "gold standard".

In addition a database table is created that contains the fields shown in Table 8.2. There are as many rows for each question as there are PubMed abstracts and the abstracts are ordered by recency.

The fields contain the following information:

Question ID	Recency	tf	tf/idf	PercentLast	Google CitationCount	ISI Citation Count	MSW	GoldStandard

Table 8.2: Database table generated by parsing the abstracts in Corpus 2 and calculating the ranking strategies.

- "Question ID": The question ID for each question.
- "Recency": Contains a PubMed ID
- "tf" : Contains the absolute number of query terms per abstract.
- "tf/idf": Contains the number of query terms divided by the number of words in the abstracts over all words in the corpus.
- "PercentLast": Contains the number of query terms in the last 20% of the abstract.
- "Google CitationCount": Contains the number of citations for each abstract as given by Google Scholar.
- "ISI Citation Count": Contains the number of citations for each abstract as given by ISI.
- "MSW": Contains the cosine adjusted minimal span weighting information.
- "GoldStandard": Contains a "0" or "1", where a "0" indicates the abstract is no "gold standard" article for the question, and a "1" indicates that the abstract is on the list of "gold standard" articles. The position of the "gold standard" articles is recorded for the recency baseline. This is required to evaluate the post-retrieval ranking strategies.

During the last step the "tf/idf - isi voting" and "voting" strategies are calculated based on the results from step six and the table containing the abstract IDs is reordered applying each of the different strategies. The position of the "gold standard" articles after reranking is recorded to determine where the "gold standard" reference is after ranking in comparison to its rank position in the baseline.

8.1.2 Rank-ordered List of MEDLINE® Abstracts for Corpus 2

This section shows an example of the rank-ordered database tables of MEDLINE® abstract PMIDs, which determine the order in which the abstracts would be displayed to a clinician to serve as answer candidates for a clinical comparison question. The abstracts for each PMID are saved in the database.

Table 8.3 shows the database table for the first ten PMIDs for the individual ranking strategies[1] for Question 26 from Corpus 2. The results are ordered by goldstandard="1", meaning the "gold standard" PMID is displayed in the first result row. The number in the recency column refers to a PMID's original ranking position in the PubMed recency baseline.

Question	PMID	goldstandard	recency	tf	tf/idf	Last20%	google	isi	msw
s26	9691103	1	45	2.946	0.167	0	0	166	0
s26	19827443	0	2	0.000	0.000	0	0	0	0
s26	19821389	0	3	0.000	0.000	0	0	0	0
s26	19160212	0	4	2.099	0.071	1	0	0	0
s26	18646064	0	5	1.693	0.071	0	0	0	0
s26	18402168	0	6	1.000	0.004	0	4	0	0
s26	18271184	0	7	0.000	0.000	0	0	0	0
s26	18239407	0	8	1.000	0.010	1	0	12	0

Table 8.3: Database table for the first 10 PMIDs for Question 26 ordered by goldstandard="1".

Before reordering the table by the different post-retrieval strategies, the "gold standard" (Position 45 in the PubMed baseline) would not be displayed to the clinician at the top of the result set and she would need to look through a large number of possibly irrelevant abstracts until she would get to the "gold standard" reference. Using the post-retrieval strategies developed in this thesis, a new rank-ordered table is created that reorders the retrieved abstracts to display the "gold standard" abstract as high in the result set as possible. Table 8.4[2] shows the database table for the first ten PMIDs after the table was reordered using the citation count information from the best individual post-retrieval ranking "isi". This step brings the "gold standard" reference to the first position because it has the highest ISI citation count.

[1] "tf/idf - isi voting" and "voting" are calculated based on the results of the initial database table shown in the example.

[2] The column headings for "recency" and "isi" have been changed to "Baseline Rank" and "ISI citation count" respectively for clarity.

Question	PMID	goldstandard	Baseline Rank	ISI citation count
s26	9691103	1	45	166
s26	8586030	0	51	66
s26	12182751	0	29	60
s26	12895211	0	22	40
s26	12381231	0	28	21
s26	11815767	0	34	18
s26	17877506	0	11	16
s26	9691110	0	44	14
s26	12786611	0	23	13
s26	11922560	0	33	13

Table 8.4: Database table for the top 10 PMIDs for Question 26 after applying the "isi" post-retrieval ranking strategy.

Because only one "gold standard" is known for each question in Corpus 2, it cannot be determined for certain without expert judges how relevant the abstracts in the following positions are, but it can be assumed that a high citation count correlates with high quality research and abstracts that are potentially relevant. The assumption that more than one relevant abstract is displayed high in the result set by applying the post-retrieval ranking strategies can be empirically proven on Corpus 3, which has multiple "gold standard" references per question.

8.1.3 Results & Evaluation

The system retrieved a total of 3,425 MEDLINE® for the 30 questions in Corpus 2. The "gold standard" reference was retrieved for 21 out of the 30 questions, which results in a retrieval accuracy of 70% as shown in the previous analysis in Section 6.3. Table 8.5 shows the ranking of the "gold standard" abstract for each ranking strategy. The first column shows the question ID, the second column the number of abstracts retrieved per question, the third column shows the "recency" baseline followed by the columns for "tf", "tf/idf", "Last 20%", "google", "isi", "msw", "tf/idf - isi voting' and "voting". Results are shown both in terms of absolute improvement in rank over the rank of "gold standard" in the baseline and in percentage of improvement over the baseline. In the case of "n/a" none of the retrieved abstracts were the "gold standard" reference.

The average rank of the "gold standard" abstract in the baseline is rank 95. The

Question	# of Abstracts	recency	tf	tf/idf	Last 20%	google	isi	msw	tf-isi voting	voting
s1	524	344	345	74	97	404	2	48	2	8
s5	59	45	10	3	11	14	11	7	11	8
s6	79	77	62	55	59	1	1	41	1	1
s7	61	54	36	4	24	n/a	4	22	4	10
s9	34	29	4	8	4	2	2	3	2	2
s10	321	161	117	59	259	51	6	212	6	26
s11	452	109	36	33	8	60	49	13	49	33
s12	150	73	68	15	69	1	1	56	1	0
s13	3	1	2	1	2	n/a	1	1	1	0
s14	153	136	57	17	39	9	6	25	6	21
s16	41	30	4	19	4	22	11	1	11	3
s17	213	183	34	20	72	25	11	31	11	44
s18	48	16	13	5	40	2	3	36	3	7
s19	79	77	62	55	59	1	1	41	1	1
s20	60	55	29	6	15	4	1	5	1	3
s23	10	3	1	1	3	4	3	9	3	1
s24	286	222	140	20	188	79	68	90	68	66
s25	52	41	4	2	19	n/a	1	9	1	1
s27	153	121	87	19	63	27	9	78	9	35
s29	125	48	31	1	60	5	2	87	2	1
s30	163	163	93	2	122	1	1	66	1	0
Average Rank		94.667	58.810	19.952	57.952	39.556	9.238	41.952	9.238	12.905
Improvement										
# of ranks			35.857	74.714	36.714	55.111	85.429	52.714	85.429	81.762
Improvement (%)			0.379	0.789	0.388	0.582	0.902	0.557	0.902	0.864

Table 8.5: Rank of 1st relevant abstract for each strategy for Corpus 2.

best individual ranking strategy, "isi"[3], ranks the "gold standard" on rank 9 on average, which is a 90.2% improvement over the baseline. Google Scholar's citation count information is far less helpful than ISI's as is evident by the inferior performance of the "google" strategy compared to "isi". "google" ranks the "gold standard" reference 30 positions below "isi" on average and provides only a 58.2% improvement over the baseline. While Google Scholar performs well in fields such as humanities or social sciences, its coverage in the natural and health sciences is less comprehensive than that of the ISI Web of Science (Harzing and van der Wal, 2008), and there is a big gap in performance between the two strategies. This trend can also be observed on Corpus 3 (c.f. Section 8.2.3).

The second best strategy "voting", which is based on a weighted combination of all the individual ranking strategies, ranks the "gold standard" on position 13 on average, which is a 86.4% improvement over the baseline.

[3]Because there is only one "gold standard" per question and the "gold standard" always received an ISI citation count, "isi" and "tf-isi voting" are identical since it was not necessary to back off to "tf/idf" to make up for a lack of citation count information.

For Corpus 2, "tf/idf", which favours abstracts with a high amount of rare terms in regards to the whole corpus, performs better than "tf" or "Last20%" providing a 78.9% improvement over the baseline. "tf" and "Last20%", which are strategies that only consider how frequent query terms are in a document but not how frequent they are over the whole corpus, only provide a 38% improvement over the baseline. "msw", which uses a vector-based approach using term proximity yields a 55.7% improvement over the recency baseline and performs better than the pure term frequency based approaches "tf" and "Last20%", but not as well as the term frequency/inverse document frequency approach "tf/idf".

Table 8.6 shows the Mean Reciprocal Rank (MRR) for each question which measures how far down on a list the first relevant abstract is. The higher the MRR, the better the ranking strategy. The MRR for the baseline is very low at 7.9%, the MRR for the most successful strategy "isi" is 48.4%. It can be seen that "google" performs 19.6% worse at 28.8% than the other citation-based strategy "isi", which is consistent with the results discussed at the beginning of this section and also holds true for Corpus 3 as will be shown in Section 8.2.

The MRR for the post-retrieval strategies based on term frequency is significantly lower than the MRR for the strategies using citation count information. The results vary between 8.9% for the worst strategy "Last 20%" and 26.2% for the best term frequency based strategy "tf/idf". Because all MEDLINE® abstracts retrieved for a query are short texts, which contain all the query terms and share the same general topic, it is harder to discriminate between abstracts using strategies that only rely on the abstract content rather than strategies using external information like the citation count. This is evident in the difference in the performance of the strategies relying on information intrinsic to an abstract versus external information.

Ranking Strategy	MRR
recency	0.079
tf	0.128
tf/idf	0.262
Last 20%	0.089
google	0.288
isi	0.484
msw	0.155
tf/idf - isi voting	0.484
voting	0.330

Table 8.6: Mean Reciprocal Rank per ranking strategy for Corpus 2

Because each question in Corpus 2 only has one "gold standard" reference, it is not possible to evaluate the results for Corpus 2 according to other standard IR metrics such as Average Precision (AP), Mean Average Precision (MAP) or Mean Rank Position (MRP). A full evaluation of the post-retrieval ranking strategies will be given for Corpus 3 in Section 8.2.

8.1.4 System Comparison

Table 8.7 compares the best retrieval strategy of RetroRank to the ranking of the "gold standard" article achieved by *ask*Medline and Entrez PubMed.

Question	RetroRank Best Strategy "isi"	*ask*Medline	Entrez
s1	2	n/a	n/a
s5	11	2	6
s6	1	n/a	n/a
s7	4	71	n/a
s9	2	7	n/a
s10	6	23	11
s11	49	1	1
s12	1	60	n/a
s13	1	1	n/a
s14	6	35	43
s16	11	6	n/a
s17	11	n/a	610
s18	3	10	1288
s19	1	n/a	n/a
s20	1	n/a	n/a
s23	3	1	n/a
s24	68	n/a	1194
s25	1	41	n/a
s27	9	n/a	n/a
s29	2	n/a	1314
s30	1	139	n/a
Average Rank	9.238	30.538	558.375

Table 8.7: System comparison for the best ranking strategy "isi".

*ask*Medline and Entrez PubMed do not perform post-retrieval ranking but use different retrieval strategies which affect the ranking position of the "gold standard". *ask*Medline (c.f. Section 2.5.4) uses a more restricted retrieval than RetroRank and generally retrieves less articles, which means the "gold standard" automatically ap-

pears in a higher position than in the baseline for RetroRank. EntrezPubMed (c.f. Section 2.3.3) on the other hand uses a very unrestricted retrieval approach and retrieves a very large set of articles, which means the "gold standard" can appear in a very low ranking position. Table 8.8 shows the number of abstracts retrieved for each of the three compared systems.

	# of abstracts retrieved
RetroRank	3425
*ask*Medline	2001
Entrez PubMed	35778

Table 8.8: Number of retrieved abstracts per system for Corpus 2.

Table 8.7 shows that the average ranking position of the "gold standard" reference is rank 9 in RetroRank, rank 31 in *ask*Medline and rank 558 in Entrez PubMed[4]. It can be seen that the best post-retrieval ranking strategy, "isi", in RetroRank outperforms both systems in displaying the "gold standard" in at the top of the result set in addition to achieving a better retrieval accuracy than the other two systems (RetroRank 70% accuracy, *ask*Medline 53.3% accuracy and Entrez PubMed 30% accuracy (c.f. Chapter 6).

8.1.5 Error Analysis for Corpus 2

The following section shows an analysis of the questions on which RetroRank performed worse than average or worse than *ask*Medline or Entrez PubMed as far as the ranking position of the "gold standard" reference is concerned (see Table 8.7). There are two components that determine the performance of RetroRank. The first component is the retrieval module, which deliberately employs a far less restrictive retrieval strategy than the retrieval strategy used in *ask*Medline or Entrez PubMed (c.f. Section 2.5.4 and Section 2.3.3) to maximise retrieval accuracy and recall. The second component is the post-retrieval module. The post-retrieval ranking strategies used in RetroRank are very successful but still do not produce the best result for all questions.

The error analysis for Corpus 2 has the limitation that only one "gold standard" reference is know for each question from the Essential Evidence Plus POEM archive and no human evaluation by clinicians of the relevance of the retrieved abstracts is

[4]In the case of "n/a" none of the retrieved abstracts were the "gold standard" reference.

available. Under real circumstances usually more than one MEDLINE® abstract provides a relevant answer to a drug comparison question and the abstracts ranked higher than the "gold standard" may well provide relevant answers despite not being selected as the "gold standard" abstract.

The three cases in which RetroRank fails to perform as well as *ask*Medline regarding the ranking position of the "gold standard" abstract are Question 5, Question 11 and Question 16 (c.f. Table 8.7). Question 11 is also noteworthy because it shows the second worst performance for the best overall post-retrieval ranking strategy "isi", the worst being its performance on Question 24 (c.f. Table 8.5). In the examples the drugs and diseases are highlighted in bold and additional vital information is highlighted in italics for easier readability.

Example (8.1) shows Question 5 in which the intranasal corticosteroids **budesonide** and **fluticasone propionate** are compared to determine which has a higher efficacy for treating **allergic rhinitis**.

(8.1) "Which nasal spray, **budesonide (Rhinocort)** or **fluticasone propionate (Flonase)**, is superior for once daily treatment of **allergic rhinitis**?" (Question 5 in Corpus 2)

The "gold standard" abstract is shown in Example (8.2). It is an abstract of a randomized controlled trial published in the *Journal of Allergy and Clinical Immunology* in 1998 and has an ISI citation count of 32. It contains an explicit comparison between the two drugs and provides a good and concise answer to the question. RetroRank ranks the "gold standard" on position 11 out of 59, while *ask*Medline ranks it on position 2 out of 36.

(8.2)

PMID: 9847429

Title: Comparison of the efficacy of **budesonide** and **fluticasone propionate** aqueous nasal spray for *once daily* treatment of perennial **allergic rhinitis**.

Abstract:
BACKGROUND: Intranasal corticosteroids, such as **budesonide** and **fluticasone propionate**, are widely prescribed in the treatment of perennial **allergic rhinitis**. *Once daily* **budesonide** dry powder and **fluticasone propionate** aqueous suspension have been found to provide similar efficacy in controlling symptoms of perennial **allergic rhinitis**.
OBJECTIVE: The purpose of this study was to assess the efficacy and safety of

treatment with *once daily* **budesonide** aqueous nasal spray.
METHODS: This study involved a multicenter, blinded, randomized, parallel-group, placebo-controlled trial of adults with perrenial **allergic rhinitis**. Patients (n = 273) recorded daily nasal symptoms for 8 to 14 days (baseline) and 6 weeks (treatment).
RESULTS: **Budesonide** decreased combined symptoms to a significantly greater extent than did **fluticasone** (P =.03); both treatments significantly decreased mean combined nasal symptoms scores compared with placebo. Of the 3 nasal symptoms assessed (ie, nasal blockage, runny nose, and sneezing), nasal blockage was significantly (P =. 009) more decreased with **budesonide** compared with **fluticasone**. Both treatments also significantly improved runny nose and sneezing compared with placebo. Improvement in combined nasal symptom scores of the **budesonide**-treated group reached statistical significance within 36 hours compared with placebo (P =.01); in those patients treated with **fluticasone**, significant improvement compared with placebo was first observed within 60 hours. Adverse events were mild and transient.
CONCLUSIONS: *Once daily* **budesonide** aqueous nasal spray, 256 microgram, was significantly better in controlling the symptoms of perrenial **allergic rhinitis** than *once daily* **fluticasone propionate**, 200 microgram, especially nasal blockage. Both treatments were superior to placebo. **Budesonide** may have a faster onset of action than **fluticasone**.

RetroRank's best strategy "isi" ranks the abstract in Example (8.3) first, which is the abstract for a review of inhaled and intranasal corticosteroids like the two drugs in the question published in *Drug Safety* in 2000. The drugs are explicitly mentioned but not directly compared to each other which makes the abstract less relevant than the "gold standard" abstract which contains a direct comparison of **budesonide** and **fluticasone**. While it presents a good overview of corticosteroids, it fails to address the question in a concise manner.

(8.3)
PMID: 10915030

Title: Safety of inhaled and intranasal corticosteroids: lessons for the new millennium.

Abstract
Although inhaled and intranasal corticosteroids are first-line therapy for asthma and **allergic rhinitis**, there has recently been an increasing awareness of their propensity to produce systemic adverse effects. The availability of more potent and lipophilic corticosteroids and new chlorofluorocarbon (CFC)-free formulations has focused attention on these safety issues. The main determinant of systemic bioavailability of these drugs is direct absorption from the lung or nose, where there is no first-pass inactivation. Consequently, the systemic bioavailability of inhaled corticosteroids is greatly influenced by the efficiency of the inhaler

> device. Thus, when comparing different inhaled corticosteroids it is imperative to consider the unique drug/device interaction. The pharmacokinetic profile is important in determining the systemic bioactivity of inhaled and intranasal corticosteroids. For highly lipophilic drugs, such as **fluticasone propionate** or mometasone furoate, there is preferential partitioning into the systemic tissue compartment, and consequently a large volume of distribution at steady state. In contrast, drugs with lower lipophilicity, such as triamcinolone acetonide or **budesonide**, have a smaller volume of distribution. The systemic tissue compartment may act as a slow release reservoir, resulting in a long elimination half-life for the lipophilic drugs. For intranasal corticosteroids, a high degree of lipophilicity diminishes water solubility in mucosa and therefore increases the amount of drug swept away by mucociliary clearance before it can gain access to tissue receptor sites. This may reduce the anti-inflammatory efficacy in the nose, but might also reduce the propensity for direct systemic absorption from the nasal cavity. The hydrofluoroalkane (HFA) formulations of beclomethasone dipropionate are solutions and exhibit a much higher respirable fine particle dose than do the CFC formulations. Dose-response studies with one of the HFA formulations have shown therapeutic equivalence at half the dosage, with little evidence of adrenal suppression at dosages up to 800 microg/day. A lack of similar studies for another of the available HFA formulations has led to a discrepancy in the recommendations for equivalence. Although in vitro studies have pointed to a similar fine particle distribution for the HFA and CFC formulations of **fluticasone propionate**, this is not supported by in vivo data for lung bioavailability, suggesting that care will be required when switching these formulations. Prescribers of inhaled and intranasal corticosteroids should be aware of the potential for long term systemic effects. The safest way to use these drugs is to 'step-down' to achieve the lowest possible effective maintenance dosage.

The abstract ranked second has 69 ISI citations and is for a randomized controlled trial comparing intranasal corticosteroids to topical antihistamines in the treatment of allergic rhinitis. The drugs from the question appear in the MeSH terms as intranasal corticosteroids, however, like the first abstract, this abstract fails to supply a direct comparison between the two drugs.

The abstract ranked third (Example (8.4)) is for a review published in *Respiratory Medicine* in 1995 and has 66 ISI citations, which is twice as many as the "gold standard" abstract. It does provide a direct comparison between the drugs like the "gold standard" reference, however it discusses the clinical potency of the drugs rather than their use as a once daily treatment.

(8.4)

PMID: 7569173

Title: Fluticasone propionate–an update on preclinical and clinical experience.

> **Outcome sentence:** Results show that **FP** [**Fluticasone propionate**] has at least twice the clinical potency of beclomethasone dipropionate and **budesonide**. This appears to be achieved without an accompanying increase in systemic effects, suggesting a therapeutic index which may be higher than other currently available inhaled corticosteroids.

*ask*Medline displays the abstract in Example (8.5) on position 1, which provides a better answer to Question 5 than the abstract ranked 1st by RetroRank. Because *ask*Medline uses a more restricted retrieval algorithm that includes query terms such as "once", "daily" and "treatment", it retrieves abstracts that match the question very closely.

(8.5)
PMID: 11422149

> **Title**: Comparison of once daily **fluticasone** propionate aqueous nasal spray with *once daily* **budesonide** reservoir powder device in patients with perennial **rhinitis**.

In conclusion it can be stated that the ISI citation count alone fails to achieve the desired result of displaying the "gold standard" within the Top 10 of the result set for Example (8.1) and that the abstracts ranked on the first three positions fail to provide a good answer. While different post-retrieval ranking strategies such as "tf/idf" achieve a result similar to *ask*Medline by ranking the "gold standard" reference 3rd, *ask*Medline's performance suggests that drug dosage information plays an important role in a query and could be a valuable addition to RetroRank's retrieval module.

The second question for which RetroRank performs worse than *ask*Medline is Question 11 shown in Example (8.6). In Question 11 **enoxaparin** (**ENOX**) is compared to unfractioned **heparin** (**UFH**) for patients with **myocardial infarction**.

(8.6) In patients with acute **ST-elevation myocardial infarction** who receive fibrinolysis and subsequent percutaneous intervention, is **enoxaparin** superior to unfractionated **heparin**? (Question 11 in Corpus 2).

RetroRank retrieves 452 abstracts for this question, while *ask*Medline retrieves 25 and Entrez PubMed only 1. RetroRank uses the least restricted retrieval query out of the three systems, which leads to a large result set in which the best post-retrieval strategy "isi" ranks the "'gold standard" reference 49th, while *ask*Medline and Entrez PubMed

display it as the 1st search result. The main cause of the poorer performance of RetroRank seems to lie with the large result set retrieved by the search query which is not specific enough to exclude results that do not match the question exactly. Therefore only one out of the eight post-retrieval ranking strategies brings the "gold standard" reference in the Top 10 by ranking it 8th. The overall best strategy "isi" fails to bring a "gold standard" to the top 10, because there a large number of abstracts retrieved that have a very high citation count, while not being a "gold standard". Many of the retrieved abstracts provide good clinical evidence but are not specific enough to the exact question.

Example (8.7) shows the title of the "gold standard" reference along with its conclusion section containing the desired drug comparison and matches the question very well. It is a randomized controlled trial published in the *Journal of the American College of Cardiology* in 2007 and has 45 citations in the ISI Web of Science.

(8.7)

PMID: 17560287

Title: Percutaneous coronary intervention in patients receiving **enoxaparin** or unfractionated **heparin** after fibrinolytic therapy for **ST-segment elevation myocardial infarction** in the ExTRACT-TIMI 25 trial.
[...]
CONCLUSION:
Among patients treated with fibrinolytic therapy for **STEMI** who underwent subsequent PCI, **ENOX** administration was associated with a reduced risk of death or recurrent MI without difference in the risk of major bleeding. The strategy of **ENOX** support for fibrinolytic therapy followed by PCI is superior to **UFH** and provides a seamless transition from the medical management to the interventional management phase of **STEMI** without the need for introducing a second anticoagulant in the cardiac catheterization laboratory.

Example (8.8) shows the reference ranked 1st by RetroRank, which has 975 citations in the ISI Web of Science and is a randomized controlled trial published in the *New England Journal of Medicine* in 1997.

(8.8)

PMID: 9250846

Title: A comparison of low-molecular-weight **heparin** with unfractionated **heparin** for unstable coronary artery disease. Efficacy and Safety of Subcutaneous **Enoxaparin** in Non-Q-Wave Coronary Events Study Group.

Abstract
BACKGROUND:
Antithrombotic therapy with **heparin** plus aspirin reduces the rate of ischemic

> events in patients with unstable coronary artery disease. Low-molecular-weight **heparin** has a more predictable anticoagulant effect than standard unfractionated **heparin**, is easier to administer, and does not require monitoring.
> **METHODS**:
> [...]
> **RESULTS**:
> At 14 days the risk of death, **myocardial infarction**, or recurrent angina was significantly lower in the patients assigned to **enoxaparin** than in those assigned to unfractionated **heparin** (16.6 percent vs. 19.8 percent, P=0.019). At 30 days, the risk of this composite end point remained significantly lower in the **enoxaparin** group (19.8 percent vs. 23.3 percent, P=0.016). The need for revascularization procedures at 30 days was also significantly less frequent in the patients assigned to **enoxaparin** (27.1 percent vs. 32.2 percent, P=0.001). The 30-day incidence of major bleeding complications was 6.5 percent in the **enoxaparin** group and 7.0 percent in the unfractionated-**heparin** group, but the incidence of bleeding overall was significantly higher in the **enoxaparin** group (18.4 percent vs. 14.2 percent, P=0.001), primarily because of ecchymoses at injection sites.
> **CONCLUSIONS**:
> Antithrombotic therapy with **enoxaparin** *plus aspirin* was more effective than unfractionated **heparin** *plus aspirin* in reducing the incidence of ischemic events in patients with unstable angina or non-Q-wave **myocardial infarction** in the early phase. This benefit of **enoxaparin** was achieved with an increase in minor but not in major bleeding.

The abstract ranked first by RetroRank provides high quality clinical evidence as is evident in the high citation count and the high impact factor of the journal the trial was published in, but unlike the "gold standard" reference the drugs are compared to each other in combination with *aspirin*. Also *fibrinolytic therapy* and *ST-segment elevation* are not mentioned in the main text of the abstract and are only visible in the MeSH terms. To a layman it is hard to judge the relevance of this abstract because clinical knowledge is needed to determine the connection between *ST-segment elevation myocardial infarction* and *non-Q-wave myocardial infarction*. Given the quality of the study it might still be a useful resource for a clinician, who has the expert knowledge to determine its suitability.

Example (8.9) shows Question 16, in which **ivermectin** is compared to **lindane** for treating **scabies** to find out if **invermectin** is more effective.

(8.9) Is **ivermectin** more effective than **lindane** for treating **scabies**? (Question 16 in Corpus 2).

RetroRank's best strategy "isi" ranks the "gold standard" reference on position 11, while *ask*Medline displays it as the 6th result and Entrez PubMed does not retrieve it at all. An extract of the "gold standard" reference is shown in Example (8.10). The "gold

standard" is a randomized controlled trial published in the *Journal of Dermatology* in 2000.

(8.10)

PMID: 11603388

Title: Oral **ivermectin** in **scabies** patients: a comparison with 1% topical **lindane** lotion.

Abstract
Scabies, which constitutes a significant proportion of the outpatient attendance in tropical dermatology clinics, has so far been treated with **lindane**, crotamiton, sulphur, permethrin, etc. **Ivermectin**, an orally administered drug, was tried in **scabies** patients and compared with 1% topical **lindane** lotion to evaluate its effects and toxicity profile.
[...]
Oral **ivermectin** is an easy drug to administer. It is given as a single oral dose, unlike **lindane**, which has to be applied topically. The compliance is accordingly increased. Moreover, **ivermectin** induces an early and effective improvement in signs and symptoms. Thus, it may be a better option for **scabies** than the traditional topical **lindane** lotion.

Example (8.11) shows an extract of the top ranked abstract in RetroRank, which has 50 citations in the ISI Web of Knowledge, whereas the "gold standard" reference only has 18 citations. The top ranked abstract in RetroRank is a randomized controlled trial published in the *Archives of Dermatology* in 1999.

(8.11)

PMID: 10376691

Title: Equivalent therapeutic efficacy and safety of **ivermectin** and **lindane** in the treatment of human **scabies**.

Abstract
OBJECTIVE:
To compare the therapeutic efficacy and safety of **ivermectin** and **lindane** for the treatment of human **scabies**.
[...]
INTERVENTION: Patients received either a single oral dose of **ivermectin** (150-200 microg/kg of body weight) or a topical application of 1% **lindane** solution. Treatment was repeated after 15 days if clinical cure had not occurred. [...]
CONCLUSIONS:
Ivermectin is as effective as **lindane** for the treatment of **scabies**. **Ivermectin** is simpler to use and, therefore, is a promising tool to improve compliance and to control infestations.

As can be seen from the excerpt, the abstract ranked first by RetroRank is relevant and provides a good and concise answer to the question consistent with the "gold

standard" reference. While it was not chosen as the "gold standard" for Question (8.9), its high citation count suggests it provides good quality clinical evidence and could be a good "gold standard" candidate. Question (8.9) is an example for a question for which the "isi" strategy ranked the "gold standard" outside of the top 10 but where it still provided a very good abstract at the top of the result set. While "isi" performed worse than average for this question, the "gold standard" was ranked first by the "msw" strategy in RetroRank.

The question with the poorest performance of "isi" in Corpus 2 regarding the ranking postition of the "gold standard" is Question 24 shown in Example (8.12). RetroRank does however outperform both *ask*Medline and Entrez PubMed on this question. *ask*Medline did not retrieve any abstracts for the question. Entrez PubMed displayed the "gold standard" on position 1194, while the "isi" strategy in RetroRank ranks it at position 69.

(8.12) Which inhaled corticosteroid is most effective in the treatment of persistent **asthma**: **fluticasone** (Flovent) or **beclomethasone** (Beclovent, Vanceril)? (Question 24 in Corpus 2).

Example (8.13) shows an excerpt from the "gold standard" reference, which is a randomized controlled trial published in the *Journal of Allergy and Clinical Immunolgy* in 1999 and has an ISI citation count of 27. The abstract ranked first by RetroRank is from the same year but has a much higher citation count of 409.

(8.13)

PMID: 10329812

Title: A comparison of multiple doses of **fluticasone** propionate and **beclomethasone** dipropionate in subjects with persistent **asthma**.

Abstract
BACKGROUND:
Inhaled corticosteroids are recommended for the treatment of persistent **asthma**. Comparative clinical studies evaluating 2 or more doses of these agents are few.
OBJECTIVE:
We sought to compare the efficacy and safety of 2 doses of **fluticasone** propionate (88 micrograms twice daily and 220 micrograms twice daily) with 2 doses of **beclomethasone** dipropionate (168 micrograms twice daily and 336 micrograms twice daily) in subjects with persistent **asthma**.
[...]
CONCLUSION:
Fluticasone propionate provides greater **asthma** control at roughly half the dose of **beclomethasone** dipropionate, with a comparable adverse event profile.

Example (8.14) shows the abstract ranked first by RetroRank.

(8.14)

PMID: 10326936

Title: Systemic adverse effects of inhaled corticosteroid therapy: A systematic review and meta-analysis.

Abstract
OBJECTIVE:
To appraise the data on systemic adverse effects of inhaled corticosteroids.
METHODS:
A computerized database search from January 1, 1966, through July 31, 1998, using MEDLINE, EMBASE, and BIDS and using appropriate indexed terms. Reports dealing with the systemic effects of inhaled corticosteroids on adrenal gland, growth, bone, skin, and eye, and reports on pharmacology and pharmacokinetics were reviewed where appropriate. Studies were included that contained evaluable data on systemic effects in healthy volunteers as well as in *asthmatic* children and adults. A statistical meta-analysis using regression was performed for parameters of adrenal suppression in 27 studies.
RESULTS:
Marked adrenal suppression occurs with high doses of inhaled corticosteroid above 1.5 mg/d (0.75 mg/d for **fluticasone** propionate), although there is a considerable degree of interindividual susceptibility. *Meta-analysis showed significantly greater potency for dose-related adrenal suppression with **fluticasone** compared with **beclomethasone** dipropionate, budesonide, or triamcinolone acetonide, whereas prednisolone and fluticasone propionate were approximately equivalent on a 10:1-mg basis.* Inhaled corticosteroids in doses above 1.5 mg/d (0.75 mg/d for **fluticasone** propionate) may be associated with a significant reduction in bone density, although the risk for osteoporosis may be obviated by postmenopausal estrogen replacement therapy. Although medium-term growth studies showed suppressive effects with 400-microg/d **beclomethasone** dipropionate, there was no evidence to support any significant effects on final adult height. Long-term, high-dose inhaled corticosteroid exposure increases the risk for posterior subcapsular cataracts, and, to a much lesser degree, the risk for ocular hypertension and glaucoma. Skin bruising is most likely to occur with high-dose exposure, which correlates with the degree of adrenal suppression.
CONCLUSIONS:
All inhaled corticosteroids exhibit dose-related systemic adverse effects, although these are less than with a comparable dose of oral corticosteroids. *Metaanalysis shows that **fluticasone** propionate exhibits greater dose-related systemic bioactivity compared with other available inhaled corticosteroids, particularly at doses above 0.8 mg/d.* The long-term systemic burden will be minimized by always trying to achieve the lowest possible maintenance dose that is associated with optimal *asthmatic* control and quality of life.

It is a systematic review published in the *Archives of Internal Medicine* in 1999. Systematic reviews are recognized as authoritative, high quality clinical evidence.

While not being the "gold standard" reference, this systematic review represents a relevant and very good answer option for Question (8.12). The most relevant sentences in Example (8.14) are highlighted in italics. Like for Question (8.9), RetroRank failed to rank the "gold standard" reference in the top 10 but instead ranked another abstract first, which provides an authoritative answer.

This error analysis shows that RetroRank overall provides relevant answers to clinical comparison questions even though it does not always achieve the optimal result. The overall best strategy "isi" does not always perform best for all questions. However, Question (8.9) and especially Question (8.12) illustrate examples, where "isi" ranks relevant, good quality abstracts at the top of the result set, which could be potential "gold standard" reference candidates if a human evaluation was performed on the results. Question (8.1) and Question (8.6) illustrate examples where the post-retrieval ranking module of RetroRank suffers from problems associated with the less restricted retrieval strategy employed in the retrieval module of RetroRank. However, a certain trade-off between recall and precision is to be expected for any automated system.

8.2 Post-retrieval Ranking on Corpus 3

8.2.1 System Implementation

In order to fully evaluate the post-retrieval ranking strategies described in Chapter 7, the RetroRank system was extended to deal with the different data structure of Corpus 3 (c.f. Section 4.3). Figure 8.2 shows the system diagram of the extended system.

The first step in the evaluation system involves parsing a .csv file consisting of questions marked-up by LT-TTT2 as described Section 6.1 and their associated Cochrane review IDs to create two maps. Table 8.9 shows an example from the input file. The rest of the questions in Corpus 3 can be found in Appendix F.

Cochrane ID	Question
CD008418	How efficient and safe is <drug> Formoterol</drug> versus short-acting <drug>beta-agonists</drug> as relief medication for adults and children with <disease>asthma</disease> ?

Table 8.9: Extract from the input file for ranking system for Corpus 3.

The first map consists of the Cochrane ID and the list of query terms (drugs and

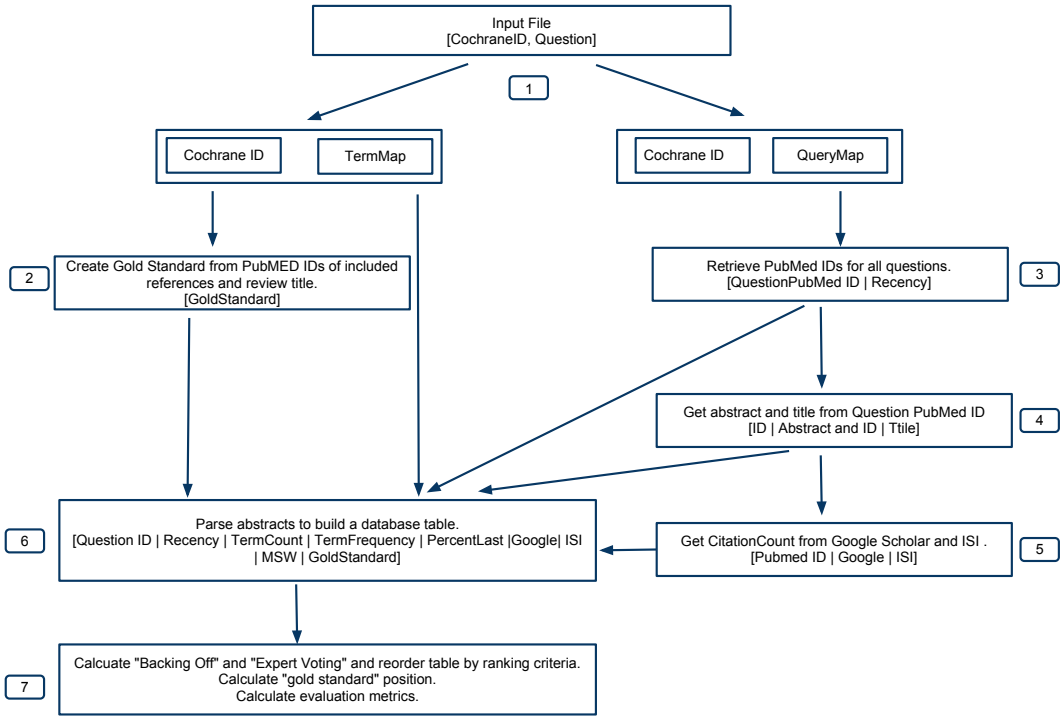


Figure 8.2: System diagram of the ranking system for Corpus 3.

diseases). The second map consists of the Cochrane ID and the E-Utilities query corresponding to the question. The query terms are used later on to calculate some of the ranking strategies. The E-Utilities query is used to retrieve PubMed abstracts for each question and their associated MeSH terms.

The second step is to create a list of PMIDs of "gold standard" articles. "Gold standard" articles are the Cochrane review itself and the references in the "references included" section of the Cochrane review. Because Cochrane reviews consist of controlled, high-quality clinical evidence, the references included in them present a good "gold standard" for a full evaluation of the different ranking strategies.

Creating the gold standard file involves two stages. In the first stage, the Cochrane review abstract is retrieved for each Cochrane ID from the input file. The abstract includes the title of the review which is looked up in PubMed. If the review is indexed in PubMed, its PMID is retrieved and stored. By knowing the PMID of the Cochrane review it can later be checked if it was retrieved by the retrieval component of RetroRank.

During the second stage all references that are included in the Cochrane review

are retrieved. These can be found under “references” in the bibliography section. For each retrieved reference it is checked, whether it is indexed in PubMed. If it has a PMID, the ID is retrieved and stored. References from other biomedical databases such as Embase, which do not have a PMID are discarded, because the evaluation of RetroRank is only performed for PubMed articles. At the end of both stages a list of gold standard PMIDs exists for each question, which will be used to evaluate the retrieved articles.

During the third step the retrieval system introduced in Chapter 6 is used to get a list of all the PMIDs for the PubMed abstracts retrieved for the query terms in each question. In a fourth step, the abstracts and abstract titles are associated with their corresponding PMIDs.

During the fifth step the citation count for all retrieved articles is obtained by searching for the abstract title in Google Scholar and for the PMID of each abstract in the ISI Web of Knowledge. The PMID and the corresponding Google Scholar and ISI citation count numbers are stored in a map.

Step number six involves parsing the abstracts. During this stage a number of processes take place:

- Stop words are removed using the PubMed stop word list (see Appendix C).
- The number of words and keywords is counted.
- The term frequency of the query terms is determined.
- The number of query terms in the last 20% of the abstracts is calculated.
- Cosine Minimum Span Weighting (“msw”) is calculated.
- The citation count is applied.
- The position of the abstracts ordered by recency as retrieved by PubMed is recorded as the baseline (”recency”).
- It is checked whether an abstract is a “gold standard”.

In addition, the table in Example 8.10 is created which contains the following fields:

The fields contain the following information:

- “Question ID”: The question ID for each question.

Question ID	Recency	tf	tf/idf	PercentLast	Google CitationCount	ISI Citation Count	MSW	GoldStandard

Table 8.10: Table generated by parsing the abstracts in Corpus 3.

- "Recency": Contains a PMID.
- "tf" : Contains the absolute number of query terms per abstract.
- "tf/idf": Contains the number of query terms divided by the number of words in the abstracts over all words in the corpus.
- "PercentLast": Contains the number of query terms in the last 20% of the abstract.
- "Google CitationCount": Contains the number of citations for each abstract as given by Google Scholar.
- "ISI Citation Count": Contains the number of citations for each abstract as given by ISI.
- "MSW": Contains the cosine adjusted minimal span weighting information.
- "GoldStandard": Contains a "0" or "1", where a "0" indicates the abstract is no "gold standard" and a "1" indicates an abstract is on the list of "gold standard" articles for the question. The position of the "gold standard" articles is recorded for the recency baseline. This is required to evaluate the post-retrieval ranking strategies.

During the last step the "tf/idf - isi voting" and "voting" strategies are calculated based on the results from step six and the table containing the abstract IDs is reordered applying each of the different strategies. The position of the "gold standard" articles after reranking is recorded to determine where the "gold standard" references are after ranking in comparison to the baseline position. Using this data, the following evaluation metrics are calculated.

- **Average Precision (AP)**: The average precision value.
- **Mean Average Precision (MAP)**: The average of the precision values after each relevant document is retrieved (Baeza-Yates, Ricardo A. and Ribeiro-Neto, B., 1999).

- **Bpref**: A preference-based measure that depends on the number of judged non-relevant documents retrieved before the judged relevant ones ("gold standard" documents).
- **Mean Rank Precision (MRP)**: The mean value of the rank positions calculated over all questions at a certain rank.
- **P@10**: The fraction of relevant documents in the top ten results.
- **Mean Reciprocal Rank (MRR)**: Measures how far down on a list the first relevant abstract is.
- **Total Document Reciprocal Rank (TDRR)**: The sum of the reciprocal ranks of all relevant documents. Unlike MRR it captures the ranks of all relevant documents.

8.2.2 Rank-ordered List of MEDLINE® Abstracts for Corpus 3

This section shows an example of the rank-ordered database tables of MEDLINE® abstracts for Corpus 3. These tables determine the order in which the abstracts would be displayed to a clinician to serve as answer candidates for her clinical comparison question in a front end for RetroRank.

Table 8.11 shows the database table for the "gold standard" PMIDs for all the ranking strategies for Question CD006015. The number in the recency column refers to a "gold standard" PMID's original ranking position in the PubMed recency baseline. The numbers in the other colums refer to a "gold standard" PMID's ranking position after applying the different post-retrieval ranking strategies.

The first "gold standard" reference would be displayed on position 23 in the example in Table 8.11 using the recency baseline before reordering the table with the post-retrieval ranking strategies developed for RetroRank. After applying the "isi" strategy, the same reference would be displayed on position 2.

Table 8.12 and Table 8.13[5] show the reordered tables for the first 20 PMIDs after applying the "isi" and "google" post-retrieval ranking strategies respectively.

Both citation count-based strategies bring a "gold standard" reference to the first position in the result set. "isi", which is the best individual post-retrieval strategy overall, outperforms "google" and brings eight "gold standard" references into the top

[5]The column headings for "recency" and "isi" have been changed to "Baseline Rank" and "ISI citation count" respectively in both tables for clarity.

PMID	recency	tf	tf/idf	Last20%	google	isi	msw	tf-isi voting	voting
16406915	23	4	7	9	5	2	4	2	6
1383816	174	153	103	160	n/a	n/a	150	144	87
7678871	167	117	73	124	115	78	103	78	67
11377309	73	43	19	32	24	16	29	16	16
14681504	40	5	11	10	12	3	13	3	7
12639651	51	19	16	16	16	11	21	11	10
1283370	179	165	133	167	n/a	n/a	161	146	118
7684524	170	135	89	131	125	98	109	95	75
12559281	53	29	18	23	21	13	23	13	12
9475762	122	60	32	65	42	35	58	35	42
9152564	128	87	48	75	67	50	70	52	54
9751354	109	57	28	44	35	27	47	26	36
8911291	136	88	58	80	68	51	73	53	55
8595647	151	97	66	114	109	63	89	64	61
7495111	148	93	64	89	80	61	88	61	60
10737482	84	55	22	35	27	22	33	22	20
10210385	100	56	27	42	34	25	42	25	26
9610579	115	59	31	51	41	28	55	28	41
7542109	156	113	71	116	112	67	93	66	66
17869295	13	2	3	3	1	1	2	1	1
9146609	125	63	40	72	58	43	69	42	50
7678931	171	141	102	139	n/a	n/a	121	133	85
8684407	139	92	63	86	78	59	77	59	58
11018616	82	44	21	33	26	18	32	18	17
12796667	47	12	14	14	14	10	19	10	9

Table 8.11: Database table showing the rank position of the “gold standard” PMIDs for Question CD006015.

20 as opposed to five “gold standard” references for “google”. The “gold standard” reference brought to the first position is not the same for the two metrics. As stated above, it is not relevant which “gold standard” reference is displayed first because all “gold standard” references provide high quality, relevant clinical evidence. However, a post-retrieval ranking strategy is more successful if it can bring more “gold standard” references to the top of the result set, which is the case for “isi”.

Table 8.14 shows the database table after reordering it using the “tf” strategy. While this strategy also brings five “gold standard” references into the top 20 and displays the first “gold standard” ranked highly on position 2, it is interesting to see that it has far less discriminative power than the citation-based strategies. This is evident in several

Question	PMID	goldstandard	Baseline Rank	ISI citation count
CD006015	1383816	1	174	707
CD006015	9475762	1	122	516
CD006015	14681504	1	40	511
CD006015	8804493	0	138	257
CD006015	1371291	0	177	217
CD006015	9820264	0	108	183
CD006015	7510911	0	161	152
CD006015	10096369	0	103	152
CD006015	1373779	0	176	137
CD006015	12559281	1	53	123
CD006015	8911291	1	136	113
CD006015	15126539	0	36	111
CD006015	7495111	1	148	105
CD006015	10510925	0	89	96
CD006015	9187688	0	124	92
CD006015	9610579	1	115	91
CD006015	9697781	0	112	86
CD006015	9751354	1	109	84
CD006015	8922564	0	135	75
CD006015	8636309	0	144	71

Table 8.12: Database table for the top 20 PMIDs for Question CD006015 after applying the “isi” post-retrieval ranking strategy.

PMIDs having the same term frequency result, which is not surprising given the nature of the retrieved abstracts, which are all short texts about the same topic containing the same query terms, and are therefore much more similar to each other than abstracts retrieved with partial matches or full-text documents. This problem does not exist for the citation-based strategies which reorder the abstracts using external sources.

8.2.3 Results & Evaluation

RetroRank retrieved a total of 22,916 MEDLINE® abstracts for the 45 questions in Corpus 3. “Gold standard” references were retieved for 38 out of 45 test questions from Corpus 3. Seven questions did not have PubMed articles listed as references, which means they were not considered because retrieval was only done for PubMed articles. RetroRank retrieved “gold-standard” articles for 31 from the 38 questions with “gold standard” references. The retrieval accuracy on Corpus 3 is 81.6%, which is 11.8% higher than on Corpus 2. The seven questions that did not retrieve the “gold

Question	PMID	goldstandard	Baseline Rank	Google citation count
CD006015	14681504	1	40	728
CD006015	20927745	0	1	653
CD006015	8804493	0	138	332
CD006015	1371291	0	177	202
CD006015	12559281	1	53	175
CD006015	8659328	0	143	150
CD006015	10796790	0	83	132
CD006015	19370565	0	3	132
CD006015	12137626	0	61	132
CD006015	8922564	0	135	121
CD006015	9187688	0	124	121
CD006015	8911291	1	136	116
CD006015	7510911	0	161	95
CD006015	9475762	1	122	90
CD006015	8636309	0	144	85
CD006015	8684407	1	139	75
CD006015	15142149	0	35	68
CD006015	12670567	0	49	66
CD006015	16516013	0	20	65
CD006015	11276294	0	75	63

Table 8.13: Database table for the top 20 PMIDs for Question CD006015 after applying the “google” post-retrieval ranking strategy.

standard” references were mostly “other” or “placebo” questions, meaning that a drug was compared to another drug that was not named or to a placebo. This type of question is more difficult for an automated system than an explicit comparison question naming all compared entities.

Table 8.15 shows the total number of abstracts retrieved and the position of the first ”gold-standard” abstract for each ranking strategy. The first column shows the question ID, the second column the number of abstracts retrieved per question, the third column shows the recency baseline followed by the columns for “tf”, “tf/idf”, “Last 20%”, “google”, “isi”, “msw”, “tf/idf - isi voting” and “voting”. Results are shown in terms of absolute improvement in rank over the rank of the “gold standard” in the baseline, and in percentage of improvement over the baseline. In the case of “n/a” none of the retrieved abstracts were the “gold standard” reference.

The average rank of the “gold standard” abstract in Table 8.15 for the recency base-

Question	pubmedid	goldstandard	recency	tf
CD006015	1371291	0	177	4.295837
CD006015	7684524	1	170	4.218876
CD006015	10235693	0	99	4.178054
CD006015	8595647	1	151	4.135494
CD006015	10025034	0	105	4.044522
CD006015	7542109	1	156	4.044522
CD006015	7545721	0	155	3.995732
CD006015	9824790	0	107	3.995732
CD006015	11750255	0	69	3.995732
CD006015	1373779	0	176	3.944439
CD006015	9530536	0	120	3.890372
CD006015	7692110	0	165	3.890372
CD006015	9152564	1	128	3.890372
CD006015	10165827	0	131	3.833213
CD006015	8928243	0	137	3.833213
CD006015	7513437	0	166	3.833213
CD006015	7512661	0	160	3.833213
CD006015	7678871	1	167	3.772589
CD006015	9011975	0	134	3.772589

Table 8.14: Database table for the top 20 PMIDs for Question CD006015 after applying the "tf" post-retrieval ranking strategy.

line is rank 179. The best ranking strategy, "isi"[6], ranks the "gold standard" on position 7 on average, which is a 96.1% improvement over the baseline. The Google Scholar citation-based strategy "google" performs 7.4% worse than "isi" over the baseline on Corpus 3 giving an 88.7% improvement. While the performance gap between the two citation-based strategies is smaller than on Corpus 2 (c.f. Section 8.1.3), the ISI Web of Science still is a more valuable resource for citation count information than Google Scholar for the medical domain. The second best strategy, "voting", which is based on a weighted combination of all the individual ranking strategies, ranks the "gold standard" on rank 9 on average, giving a 94.8% improvement over the baseline.

The purely term frequency based strategies "tf" and "Last20%", as well as the vector-based term proximity approach "msw", achieve very similar results with an improvement between 86.8% and 89.5% over the baseline. "tf/idf", which looks at term frequency in regards to the whole corpus performs slightly worse, achieving an

[6]Because there is only one "gold standard" per question and the "gold standard" always received an ISI citation count, "isi" and "tf/idf - isi voting" are identical since it was not necessary to back off to "tf/idf" to make up for a lack of citation count information.

Question	# of Abstracts	recency	tf	tf/idf	Last20%	google	isi	msw	tf-isi voting	voting
CD006626	985	753	45	112	73	15	6	30	6	22
CD003462	97	93	14	71	9	48	36	17	36	32
CD004707	52	49	18	6	22	1	1	7	1	1
CD006654	2335	655	34	90	36	3	1	85	1	1
CD006015	180	13	2	3	3	1	1	2	1	1
CD002258	580	180	53	41	30	153	4	48	4	14
CD006918	53	37	21	13	14	2	1	22	1	1
CD003082	3077	930	82	13	50	11	4	39	4	15
CD002960	13	13	3	13	2	9	1	3	1	1
CD000128	30	16	2	6	1	11	1	1	1	1
CD008418	427	170	11	33	19	7	10	136	10	42
CD001439	15	7	2	1	1	4	1	1	1	1
CD000127	33	18	2	2	2	10	1	1	1	1
CD006117	1343	308	8	76	1	45	11	1	11	2
CD000284	227	40	1	4	4	5	2	2	2	3
CD003135	48	34	2	7	1	6	5	4	5	3
CD000067	985	372	10	30	4	n/a	1	17	1	1
CD006352	106	24	6	1	4	18	4	6	4	3
CD003492	89	76	61	2	26	23	14	22	14	8
CD007022	539	203	1	18	1	28	16	4	16	2
CD001387	43	34	1	27	2	n/a	15	7	15	3
CD003615	146	127	23	53	83	n/a	3	32	3	9
CD001281	235	174	70	163	159	41	13	69	13	49
CD006628	112	77	3	40	2	11	9	4	9	5
CD005967	164	4	9	10	11	6	11	3	11	12
CD001031	587	421	4	42	6	86	1	5	1	1
CD001211	29	12	5	9	2	2	1	5	1	1
CD006453	35	4	17	16	33	4	6	32	6	17
CD004278	352	275	23	9	37	62	36	45	36	37
CD001100	608	421	40	33	43	4	1	77	1	1
CD002310	28	21	9	10	9	17	1	8	1	1
Average Rank		179.387	18.774	30.774	22.258	20.323	7.032	23.710	7.032	9.387
Improvement										
# of ranks			160.613	148.613	157.129	159.065	172.355	155.677	172.355	170.000
Improvement (%)			0.895	0.828	0.876	0.887	0.961	0.868	0.961	0.948

Table 8.15: Rank of 1st relevant abstract for each strategy for Corpus 3.

82.8% improvement over the recency baseline.

While the ISI citation-based strategies are consistently the most successful strategies for both corpora, the strategies relying purely on information intrinsic to the abstracts perform better on Corpus 3 than on Corpus 2. There is also a general increase in performance on Corpus 3, which is most likely due to the higher number of "gold standard" references that are retrieved and ranked per question. Also the average rank of the first relevant abstract in the baseline is lower than for Corpus 2, which leaves more room for improvement.

Table 8.16 shows the average precision (AP) and the mean average precision (MAP) for each ranking strategy on Corpus 3. A higher percentage indicates better retrieval of the "gold standard" references, i.e., a percentage of 50% means half the "gold stan-

dard" references were retrieved or one in two documents is a "gold standard" reference, whereas a low percentage of 10% indicates that only 10% of the "gold standard" references were retrieved or that one in ten documents is a "gold standard" reference. The PubMed recency baseline has a MAP of only 5.6%.

	recency	tf	tf/idf	Last20%	google	isi	msw	tf-isi voting	voting
CD003492	0.013	0.016	0.500	0.038	0.043	0.071	0.045	0.071	0.125
CD003462	0.011	0.071	0.014	0.111	0.021	0.028	0.059	0.028	0.031
CD004707	0.020	0.056	0.167	0.045	1.000	1.000	0.143	1.000	1.000
CD006654	0.019	0.029	0.027	0.026	0.073	0.121	0.024	0.121	0.059
CD006015	0.110	0.234	0.311	0.220	0.261	0.393	0.241	0.414	0.371
CD002258	0.016	0.025	0.032	0.026	0.001	0.064	0.018	0.066	0.032
CD006918	0.027	0.048	0.077	0.071	0.500	1.000	0.045	1.000	1.000
CD003082	0.006	0.015	0.019	0.016	0.046	0.075	0.014	0.075	0.035
CD002960	0.077	0.333	0.077	0.500	0.111	1.000	0.333	1.000	1.000
CD000128	0.102	0.482	0.155	0.608	0.156	0.250	0.813	0.347	0.649
CD008418	0.009	0.043	0.027	0.029	0.090	0.059	0.009	0.059	0.025
CD001439	0.348	0.628	0.515	0.675	0.324	0.657	0.764	0.748	0.740
CD000127	0.097	0.515	0.247	0.546	0.131	0.313	0.833	0.386	0.618
CD006117	0.023	0.121	0.048	0.140	0.063	0.077	0.111	0.080	0.107
CD000284	0.142	0.215	0.227	0.222	0.220	0.242	0.225	0.282	0.254
CD003135	0.081	0.307	0.174	0.381	0.193	0.323	0.307	0.323	0.212
CD000067	0.004	0.067	0.019	0.090	0.000	0.395	0.043	0.402	0.237
CD006352	0.047	0.103	0.337	0.099	0.064	0.085	0.140	0.097	0.166
CD006626	0.013	0.028	0.025	0.024	0.051	0.084	0.020	0.084	0.037
CD007022	0.019	0.425	0.086	0.331	0.055	0.078	0.138	0.083	0.197
CD001387	0.029	1.000	0.037	0.500	0.000	0.067	0.143	0.067	0.333
CD003615	0.012	0.063	0.026	0.018	0.000	0.417	0.046	0.417	0.146
CD001281	0.006	0.014	0.006	0.006	0.024	0.077	0.014	0.077	0.020
CD006628	0.013	0.333	0.025	0.500	0.091	0.111	0.250	0.111	0.200
CD005967	0.114	0.109	0.132	0.091	0.106	0.042	0.250	0.057	0.117
CD001031	0.003	0.142	0.016	0.090	0.012	0.512	0.105	0.512	0.505
CD001211	0.130	0.234	0.186	0.281	0.444	0.583	0.187	0.631	0.440
CD006453	0.250	0.059	0.063	0.030	0.250	0.167	0.031	0.167	0.059
CD004278	0.010	0.038	0.052	0.030	0.018	0.021	0.026	0.023	0.025
CD001100	0.005	0.012	0.024	0.019	0.185	0.621	0.017	0.621	0.383
CD002310	0.060	0.146	0.106	0.118	0.029	0.500	0.146	0.548	0.543
MAP	0.056	0.191	0.121	0.190	0.147	0.304	0.179	0.319	0.312

Table 8.16: Average precision (%) per metric and MAP (%) for Corpus 3

Consistent with the results shown in Table 8.15, the ISI citation-based strategies have a good MAP. The best combined strategy "tf/idf - isi voting" has a MAP of 31.9%, while the best individual ranking strategy "isi" has a MAP of 30.4%. This means that

the MAP of the best post-retrieval ranking strategy in RetroRank is 26.3% higher than for the PubMed baseline.

The MAP of the "google" strategy is only half as good as the MAP for the "isi" strategy. This finding is consistent with the other results in which "isi" achieves much better performance, and reinforces that the ISI Web of Knowledge is a valuable resource for the biomedical domain. The MAP of the strategies relying on information intrinsic to the abstracts varies between 12.1% and 19%, which is much lower than the MAP that can be achieved by using external information for post-retrieval ranking.

	recency	tf	tf/idf	Last20%	google	isi	msw	tf-isi voting	voting
CD006626	0.000	0.000	0.000	0.000	0.030	0.085	0.007	0.085	0.011
CD003462	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CD004707	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CD006654	0.000	0.007	0.000	0.006	0.061	0.107	0.000	0.107	0.031
CD006015	0.024	0.150	0.246	0.144	0.184	0.307	0.136	0.309	0.274
CD002258	0.000	0.000	0.000	0.000	0.000	0.063	0.000	0.063	0.000
CD006918	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CD003082	0.000	0.000	0.036	0.000	0.051	0.107	0.000	0.107	0.023
CD002960	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CD000128	0.000	0.250	0.000	0.313	0.000	0.188	0.563	0.188	0.375
CD008418	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CD001439	0.000	0.490	0.204	0.510	0.143	0.571	0.612	0.612	0.592
CD000127	0.000	0.375	0.125	0.375	0.000	0.188	0.563	0.188	0.188
CD006117	0.000	0.139	0.000	0.091	0.000	0.042	0.104	0.042	0.062
CD000284	0.003	0.094	0.174	0.151	0.147	0.195	0.106	0.196	0.184
CD003135	0.000	0.194	0.000	0.139	0.000	0.083	0.056	0.083	0.083
CD000067	0.000	0.000	0.000	0.056	0.000	0.139	0.000	0.167	0.139
CD006352	0.000	0.000	0.160	0.040	0.000	0.040	0.000	0.040	0.080
CD003492	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CD007022	0.000	0.313	0.000	0.258	0.000	0.000	0.129	0.000	0.133
CD001387	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CD003615	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CD001281	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CD006628	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CD005967	0.056	0.000	0.000	0.000	0.000	0.000	0.194	0.000	0.000
CD001031	0.000	0.000	0.000	0.000	0.000	0.250	0.000	0.250	0.250
CD001211	0.000	0.000	0.000	0.125	0.250	0.375	0.000	0.375	0.188
CD006453	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CD004278	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CD001100	0.000	0.000	0.000	0.000	0.000	0.375	0.000	0.375	0.188
CD002310	0.000	0.000	0.000	0.000	0.000	0.250	0.000	0.250	0.250
Average Bref	0.003	0.065	0.030	0.071	0.028	0.109	0.080	0.111	0.098

Table 8.17: Bpref per Question and Average Bpref for Corpus 3

Table 8.17 shows the bpref values per question and the average bref value for the whole corpus. Although bpref is highly correlated with average precision when the judgments are effectively complete, the value of bpref deviates from average precision and from its own value as the judgment set becomes smaller (Buckley and Voorhees, 2004), especially at very low levels of assessment. Corpus 3 contains a number of questions, which have very few ”gold standards” relative to the number of abstracts retrieved for a question, which makes bpref an unreliable metric for the task in this research illustrated in the results for bpref which range from 57.1% for “isi” for Question CD001439 with many ”gold standard” references to 0% for the questions with very few ”gold standard” references.

Table 8.18 shows the MAP and MRP at rank 5, 10 and 20. The MRP at a certain rank is the mean value of the rank positions calculated over all questions. The cut-off was chosen at position 20 because it is assumed a clinician would likely like to find a relevant document within at most 20 documents. The MRP therefore expresses how useful a ranking strategy is in practice. The results again suggest that the ISI citation-based strategies are the most successful considering the number of retrieved relevant documents at the given ranks. The difference between the best ranking strategies and the baseline strategy is again considerable with the baseline performance being considerably worse across all cut-off points.

Rank Strategy	MAP	MRP Top 5	MRP Top 10	MRP Top 20
Recency	0.058	0.013	0.026	0.039
tf	0.19	0.167	0.162	0.156
tf/idf	0.121	0.070	0.089	0.101
Last 20%	0.189	0.167	0.155	0.154
google	0.147	0.116	0.130	0.142
isi	0.304	0.317	0.323	0.311
msw	0.178	0.137	0.127	0.130
tf-isi voting	0.319	0.269	0.260	0.243
voting	0.312	0.265	0.236	0.236

Table 8.18: Mean Rank Precisions at Rank 5, 10 and 20 for Corpus 3.

The last section of the evaluation compares mean recall and precision. Graph 8.3 shows the interpolated recall/precision graph for all questions in Corpus 3. When the curve is in the upper-right portion of the graph a strategy performs well. A curve in the lower-left portion of the graph indicates that the strategies' overall performance is poor. The “x” axis shows recall and the “y” axis shows precision. Recall describes the

ratio of relevant abstracts retrieved to the total number of relevant abstracts available. Precision is the ratio of the number of relevant abstracts retrieved to the total number of retrieved abstracts. Recall and precision are inversely related and recall goes up while precision goes down and vice versa. To increase recall the search needs to be widened, which often leads to the retrieval of material that is not relevant and thereby negatively affects precision. In this analysis, the total number of relevant abstracts equals the number of "gold standards" for all questions. Recall equals "1" means all "gold standard" abstracts were retrieved.

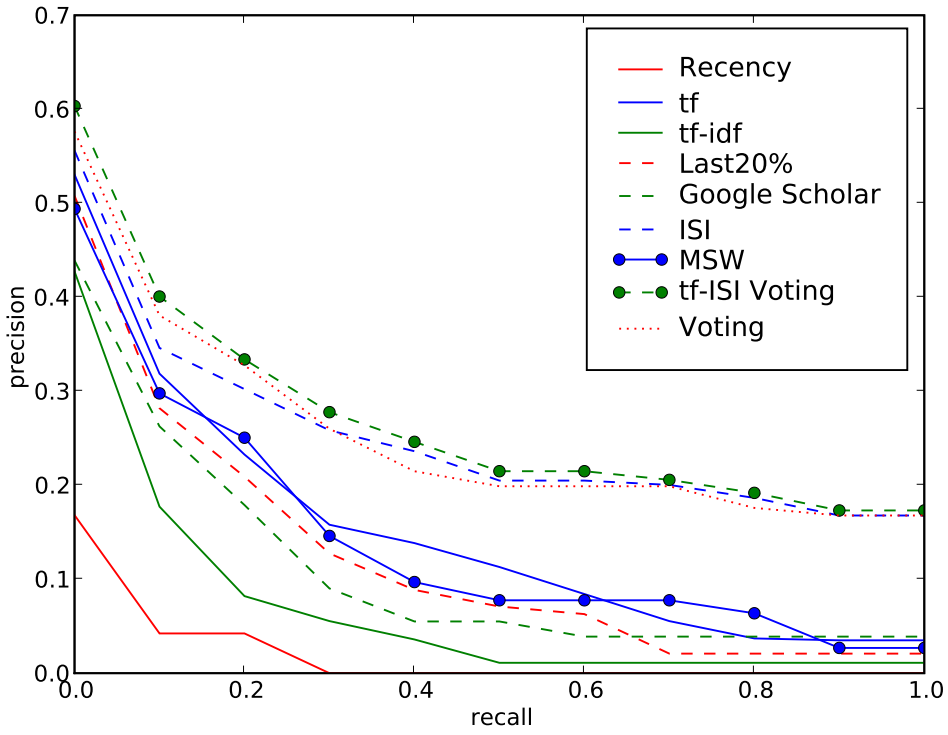


Figure 8.3: Interpolated recall/precision graph over all questions.

The "isi" strategy, as well as the two ISI citation-based voting strategies perform significantly better than the other strategies relying purely on information from the abstracts. Their initial precision is approximately 60% going down to around 18% when recall equals 100%. The baseline only has approximately 18% precision in the beginning and precision goes down to almost 0% when recall equals 30%. The graph illustrates that the ISI citation-based strategies are highly successful at identifying important articles compared to non-citation based strategies and have a good recall/precision ratio.

The evaluation in this chapter could only be performed in terms of retrieval of "gold standard" abstracts because no human judges were involved in the system evaluation. The results suggest that other abstracts that would be relevant to a clinician are brought to the top of the result set given the very good performance of the post-retrieval ranking strategies. A second, human, evaluation under a lenient condition that also considers abstracts that are not the official "gold standard" but have been judged as relevant by clinicians would be a useful avenue in future work.

8.2.4 System Comparison

The final two tables (Table 8.19 and Table 8.20) in this evaluation compare the results of RetroRank to the performance of the post-retrieval ranking system developed by Demner-Fushman and Lin (2007). The comparison is made for the therapy task described in Section 2.4.4, which is the task that best corresponds to the task of RetroRank. Demner-Fushman and Lin (2007) used PICO query frames and assessed the relevance of the retrieved abstracts by a medical doctor, whereas RetroRank works on natural language queries and does not require a doctor to evaluate the relevance of the retrieved abstracts because the corpus has a set of "gold standard" references for each question. In order to compare the two systems, the results for RetroRank were put into the same format used by Demner-Fushman and Lin (2007), namely Precision at Rank 10 (P@10), Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Total Document Reciprocal Rank (TDRR), which is the sum of the reciprocal ranks of all relevant documents.

Ranking Strategy	P@10	MAP	MRR	TDRR
PubMed	0.180	0.061	0.282	0.495
Term	0.192	0.082	0.368	0.700
EBM	0.233	0.109	0.397	0.807
Combo	0.258	0.120	0.556	0.969

Table 8.19: Performance of all systems on the test set for the therapy task (Demner-Fushman and Lin, 2007).

The results show that Demner-Fushman and Lin (2007)'s PubMed baseline is higher than the PubMed baseline for RetroRank. This is likely due to the more restricted way the query was manually constructed. The best ranking system in Demner-Fushman and Lin (2007) is the "Combo" system (A combination of the term-based reranker and the EBM scorer normalised using weighted linear interpolation). For P@10 it achieves

Ranking Strategy	P@10	MAP	MRR	TDRR
recency	0.056	0.058	0.042	0.113
tf	0.261	0.19	0.250	0.495
tf/idf	0.163	0.121	0.165	0.313
Last 20%	0.289	0.189	0.305	0.522
google	0.222	0.147	0.200	0.345
isi	0.390	0.304	0.510	0.788
msw	0.305	0.178	0.252	0.485
tf - isi voting	0.390	0.319	0.510	0.805
voting	0.390	0.312	0.516	0.765

Table 8.20: P@10, MAP, MRR and TDRR for each strategy for Corpus 3.

25.8%, whereas "isi", the best overall ranking strategy in RetroRank, achieves 39%. The MAP value for "Combo" is 12% compared to a MAP of 30.4% for "isi". The MRR of 55.6% for "Combo" is comparable to the MRR of 51% for "isi". The only metric in which "Combo" cleary performs better than "isi" is TDRR, where "Combo" achieves a score of 96.9% and "isi" only 78.8%. This metric is not as telling as the other metrics, though, because the main task for a good post-retrieval ranking system is to present the first relevant abstracts as high in the result set as possible - a task in which RetroRank outperforms the system in Demner-Fushman and Lin (2007). While the results of Demner-Fushman and Lin (2007) are a significant improvement over the PubMed baseline, the performance of RetroRank shows that it is possible to successfully use natural language input and a fully automated approach to achieve very good results.

8.2.5 Error Analysis for Corpus 3

Although the overall performance of RetroRank on Corpus 3 is very good, some questions prove more challenging than others. The following section shows the two questions for which the overall best strategy "isi" failed to bring a "gold standard" reference into the top 20 of the result set. The query terms are highlighted in bold and additional important information in the abstracts is highlighted in italics.

The first question is Question CD003462 shown in Example (8.15), in which the drug **heparin** is compared to a **placebo**.

(8.15) What are the effects of **heparin** versus **placebo** for acute **coronary syndromes**? (Question CD003462 in Corpus 3)

For this question only one "gold standard" reference out of the existing eight "gold standard" references was retrieved on position 36 after reranking the result set using the isi" strategy. The abstract for the "gold standard" reference is shown in Example (8.16). It is the abstract for a randomized controlled trial published in the *American Journal of Cardiology* in 1997 and has 16 ISI citations.

(8.16)

PMID: 9296466

Title: Low-molecular-weight **heparin** (Fragmin) during instability in *coronary artery disease* (FRISC). FRISC Study Group.

Abstract

This study evaluated whether the low-molecular-weight (LMW) **heparin** dalteparin sodium (Fragmin) had protective effects against cardiac events in aspirin-treated patients with unstable **coronary artery** syndromes. Patients (n = 1,506) with unstable angina or non-Q-wave myocardial infarction were randomized to double-blind, **placebo**-controlled treatment with LMW heparin. The treatment was given as subcutaneous injections: 120 U/kg body weight/12 hours during the first 5-7 days and 7,500 U once daily during the following 35-45 days. The primary endpoint, death or myocardial infarction after 6 days, showed a 3% (4.7%-1.7%) absolute and a 65% relative reduction in the LMW **heparin** group. There was a 6.8% (15.5%-8.7%) absolute and a 47% relative reduction of urgent revascularization or need for heparin or nitroglycerin infusions in combination with the primary endpoint. After 40 days there was an absolute reduction of death or myocardial infarction of 2.8% (10.7%-7.9%) and its combination with incapacitating angina was reduced by 5.9% (30.7%-24.8%). The survival analysis indicated a reactivation of the instability soon after lowering the dose at 5-7 days. With long-term follow-up, 3-4 months after termination of LMW **heparin**, the differences between groups were no longer statistically significant. However, the cumulative reduction in death, myocardial infarction, and revascularization because of incapacitating angina of 5.1% (25.3%-20.4%) was maintained. No cerebral and few major bleeds occurred. Compliance was adequate. *Thus, subcutaneous LMW **heparin** protects against cardiac events in the acute phase of unstable coronary artery disease*. The subcutaneous regimen also allows prolongation of treatment in the outpatient setting, which might maintain the initial benefits over a longer period.

The abstract chosen as a "gold standard" by the Cochrane Group discusses heparin in a placebo-controlled trial as specified in the question, whereas the abstract ranked 1st by RetroRank shown in Example (8.17) does address the question but the placebo-controlled trial also includes *aspirin* and a combination of **heparin** and *aspirin* treatment. It is the abstract for a randomized controlled trial published in the *New England Journal of Medicine* in 1992.

(8.17)

PMID: 1608405

Title: Reactivation of unstable angina after the discontinuation of **heparin**.

Abstract
BACKGROUND:
Heparin is an effective, widely used treatment for unstable angina. Among patients enrolled in a double-blind, randomized, **placebo**-controlled trial comparing intravenous **heparin**, *aspirin*, *both treatments*, and *neither* during the acute phase of unstable angina, we encountered patients in whom unstable angina was reactivated after **heparin** was discontinued.
METHODS:
The study population included 403 of the original 479 patients in the trial who had completed six days of blinded therapy without refractory angina or myocardial infarction. After the discontinuation of therapy, clinical events, including reactivation of unstable angina and myocardial infarction occurring within 96 hours after hospitalization, were closely monitored.
RESULTS:
Early reactivation occurred in 14 of the 107 patients who received **heparin** alone, as compared with only 5 patients in each of the other three study groups (P less than 0.01). These reactivations required urgent intervention (thrombolysis, angioplasty, or coronary-bypass surgery) in 11 patients treated with **heparin** alone, but in only 2 patients in the other groups combined (P less than 0.01). Four of the six patients who had a myocardial infarction during a reactivation of their disease were in the **heparin** group. Reactivations in this group occurred in a cluster a mean (+/- SD) of 9.5 +/- 5 hours after the discontinuation of the study drug but were randomly distributed over the initial 96 hours in the other three groups.
CONCLUSIONS:
*Although **heparin** is beneficial in treating unstable angina, the disease process may be reactivated within hours of the discontinuation of this drug.* Concomitant therapy with aspirin may prevent this withdrawal phenomenon.

The study ranked first by RetroRank has been cited 380 times as opposed to 16 times for the "gold standard" reference. While the remit of the study is wider than in the "gold standard" study, it does provide a valuable answer in the "Conclusion" section and its high citation count suggests that it is a high impact study. Nevertheless, it fails to address the exact question. RetroRank also only retrieved one out of eight possible "gold standards" references for this question.

The second question for which the "isi" strategy failed to bring a "gold standard" reference into the top 20 is Question CD004278 shown in Example (8.18), which is a question in which **haloperidol** is compared to **chlorpromazine** without specifying in what respect the two drugs should be compared.

(8.18) How does **haloperidol** compare to **chlorpromazine** for people with **schizophrenia**? (Question CD004278 in Corpus 3)

For this question RetroRank retrieved 50% of the known "gold standards" but failed to rank any "gold standard" higher than position 36. The "gold standard" reference shown in Example (8.19) addresses the comparison in a clear and succinct manner. It is a randomized clinical trial published in *Diseases of the Nervous System* in 1974.

(8.19)

PMID: 17894080

Title: Parenteral **haloperidol** for rapid control of severe, disruptive symptoms of acute **schizophrenia**.

Abstract
Intramuscular **haloperidol**, at three dose levels, (5 mg, 2 mg, and 1 mg) **chlorpromazine** (25 mg), and placebo were compared for efficacy, rapidity of therapeutic onset, and safety in 50 acute psychotic patients requiring rapid control. The drugs were administered parenterally under double-blind conditions at half-hour intervals until successful control of moderate to very severe symptomatology was achieved or a maximum of four injections had been given. Global evaluation, BPRS, and target symptom ratings were performed. *The overall results indicated that the 5 mg and 2 mg **haloperidol** doses were significantly superior to the 1 mg **haloperidol** and 25 mg **chlorpromazine** doses and to placebo.* Transfer of patients to oral **haloperidol** was satisfactorily accomplished. Side effects for all medications were minimal and included slight to moderate EPS and drowsiness. The use of anti-parkinson drugs completely controlled the extrapyramidal symptoms.

While the "gold standard" abstract clearly addressed the question, the reference ranked first by RetroRank's "isi" strategy fails to do so. The focus of the study in Example (8.20), a randomized clinical trial published in the *Archives of General Psychiatry* in 1988, is on assessing the efficacy of *clozapine* in comparison to **chlorpromazine**. The drug *Haloperidol* is mentioned in the abstract but is not part of the comparison. While the abstract contains all keywords from the query and is very highly cited with 2486 ISI citations, it does not answer the question.

(8.20)

PMID: 3046553

Title: Clozapine for the treatment-resistant schizophrenic. A double-blind comparison with **chlorpromazine**.

Abstract
The treatment of schizophrenic patients who fail to respond to adequate trials of neuroleptics is a major challenge. Clozapine, an atypical antipsychotic drug, has

> long been of scientific interest, but its clinical development has been delayed because of an associated risk of agranulocytosis. This report describes a multicenter clinical trial to assess clozapine's efficacy in the treatment of patients who are refractory to neuroleptics. DSM-III schizophrenics who had failed to respond to at least three different neuroleptics underwent a prospective, single-blind trial of **haloperidol** (mean dosage, 61 +/- 14 mg/d) for six weeks. Patients whose condition remained unimproved were then randomly assigned, in a double-blind manner, to clozapine (up to 900 mg/d) or **chlorpromazine** (up to 1800 mg/d) for six weeks. Two hundred sixty-eight patients were entered in the double-blind comparison. When a priori criteria were used, 30% of the clozapine-treated patients were categorized as responders compared with 4% of **chlorpromazine**-treated patients. Clozapine produced significantly greater improvement on the Brief Psychiatric Rating Scale, Clinical Global Impression Scale, and Nurses' Observation Scale for Inpatient Evaluation; this improvement included "negative" as well as positive symptom areas. Although no cases of agranulocytosis occurred during this relatively brief study, in our view, the apparently increased comparative risk requires that the use of clozapine be limited to selected treatment-resistant patients.

Example (8.18) illustrates the shortcomings of an automated system, which can only retrieve abstracts based on the query terms in the question but lacks the human insight necessary to determine the actual relevance of an abstract that contains all the query terms but not in the right combination. In such cases even a human-based, external strategy such as the ISI Web of Science citation count fails to rank the most relevant abstracts first, because it is not enough to know a reference provides high quality evidence if it does not deal with the exact problem in the question.

8.3 Review of the Post-Retrieval Ranking Strategies

RetroRank utilizes different post-retrieval ranking strategies. The evaluation on Corpus 2 and Corpus 3 has shown that there is a marked difference between the performance of strategies that use information purely inherent to a MEDLINE® abstract such as term frequency, term proximity or the appearance of the query terms in the "Outcome" or "Conclusion" section of the abstracts and strategies that use external information such as the citation count.

While ranking strategies that are based purely on the text of an abstract and the MeSH terms associated with that abstract provide a very good improvement over the PubMed recency baseline, they still perform less well than the best overall post-retrieval ranking strategy "isi", which reranks abstracts based on their citation count as indexed by the ISI Web of Science, which is the most trusted database for peer-reviewed content.

The MEDLINE® abstracts retrieved for a query are all short texts containing the query terms and share the same general topic. This makes it harder to discriminate between abstracts using strategies that rely only on the abstract content. While it holds true that abstracts with a higher density of query terms or a high density of query terms in the "Outcome" or "Conclusion" section tend to be more relevant, the evaluation shows that there is a limit to what can be done relying on information that can be determined from an abstract.

This problem can be solved using external sources to rerank the abstract set retrieved for a query. External sources, such as the ISI citation count, benefit from making use of expert knowledge and human judgement, i.e., the knowledge of researchers in the clinical domain, who show the impact and value of studies by citing them frequently. By using the citation count from the ISI Web of Knowledge, one can also make sure that only articles from journals that have a good reputation and are peer-reviewed are used for post-retrieval ranking. The "isi" strategy outperforms the "google" strategy, which uses the citation count information from Google Scholar, which includes a wider range of publications and has a less complete coverage for the biomedical domain.

It can be concluded that post-retrieval ranking based on the ISI citation count is a very successful means of displaying relevant and high quality research in the top 10 of the result set. The "isi" strategy achieves a 96.1% improvement over the PubMed baseline for Corpus 3 regarding the rank of the first "gold standard" abstract in the result set. The MAP of "isi" is a respectable 30.4% compared to a MAP of only 12% achieved by the best ranking system in the research by Demner-Fushman and Lin (2007).

The value of the ISI citation count is also evident in the "voting" strategy which reranks the abstracts by majority voting using all individual strategies but placing a higher weight on the results produced by "isi". The "voting" strategy achieves the second best overall improvement with 94.8% for Corpus 3 regarding the rank of the first "gold standard" abstract and has a MAP of 31.2% for Corpus 3.

8.4 Summary

In this chapter the implementation and evaluation of the post-retrieval ranking module of RetroRank was described. Firstly, the implementation and testing of the post-retrieval ranking strategies described in Chapter 7 was shown for Corpus 2. Secondly,

the implementation of the ranking strategies was described for Corpus 3, which allowed a full evaluation with standard IR metrics and a comparison with the post-retrieval ranking system developed in Demner-Fushman and Lin (2007). It was shown that the automatic ISI-citation based strategies "isi" and "voting" are a significant improvement over the PubMed recency baseline and are a strong contender to the strategies based on query frames and annotated data developed by Demner-Fushman and Lin (2007). The performance of RetroRank shows that it is possible to successfully use natural language input and a fully automated approach to retrieve answer candidates for clinical drug comparison questions.

Chapter 9

Summary and Conclusion

This chapter summarise the work presented in this thesis, assesses its contributions, points out some limitations and concludes with ideas for further research.

9.1 Summary

This thesis proposed a new QA system for clinical comparison questions called RetroRank that provides clinicians with a rank-ordered list of MEDLINE® abstracts targeted to clinical questions framed in natural language. RetroRank takes the clinician's plain text question as input, processes it and outputs a rank-ordered list of potential answer candidates, i.e., MEDLINE® abstracts. The rank-ordered list is reordered using several post-retrieval ranking strategies to ensure the most topically-relevant abstracts are displayed as high in the ranking as possible.

Chapter 1 introduced the research question of automatically answering clinicial comparison questions and outlined the scope and contributions of the thesis.

Chapter 2 provided a foundation and motivation for the research undertaken in this thesis by giving background knowledge about the information needs of clinicians and the domain of Evidence Based Medicine (EBM) as well as an overview of Information Retrieval (IR) and Question Answering (QA) techniques in the biomedical domain and existing clinical QA services and applications, which served as a basis for the evaluation of the performance of the RetroRank.

Chapter 3 introduced the characteristics of clinical comparison questions and showed the types of different comparative constructions that appear in clinical questions. Also the lexical items indicative of comparison questions were introduced, which were used for the creation of Corpus 1.

Chapter 4 described the creation and preprocessing of the three corpora of clinical questions that were used to develop and evaluate the performance of RetroRank. Corpus 1 (NLH QAS) was used for the manual exploration of the best strategy for abstract retrieval in Chapter 5 and for the development of the automated abstract retrieval component of RetroRank in Chapter 6. Corpus 2 (Essential Evidence Plus POEMs) was used for the evaluation of the retrieval component in Chapter 6 according to the same criteria used for the evaluation of *ask*Medline by Fontelo et al. (2005), and for developing the post-retrieval ranking strategies described in Chapter 7. Corpus 3 (Cochrane Systematic Reviews), which provides multiple "gold standard" references per question, was used for a full evaluation of RetroRank according to standard IR metrics.

Chapter 5 presented the initial experiment on manual query construction, abstract retrieval, and evaluation to determine the best search strategy and to evaluate it using human judges. The findings in this chapter were used to develop the automatic RetroRank QA system introduced in Chapter 6.

Chapter 6 described the implementation and evaluation of the query construction and abstract retrieval component of RetroRank. The retrieval component of RetroRank was developed on Corpus 1 (Section 4.1) and evaluated in terms of retrieval accuracy using Corpus 2 (Section 4.2).

Chapter 7 introduced the different post-retrieval ranking strategies which were used to rerank the MEDLINE® abstracts retrieved for each question in Corpus 2 and Corpus 3 in order to display the most relevant abstracts as high in the result set as possible to improve over the PubMed recency baseline, which does not display results by relevance. Post-retrieval ranking is an important feature in a system that is geared towards providing the most relevant abstracts at the top of the result set to enable clinicians to find the most relevant information in a quick and reliable way.

Chapter 8 described the implementation and evaluation of the post-retrieval ranking module of RetroRank. The post-retrieval ranking component was developed and tested on Corpus 2 and fully evaluated on Corpus 3. The evaluation showed how the different post-retrieval ranking strategies perform on the different corpora and in comparison to the post-retrieval ranking system developed in Demner-Fushman and Lin (2007). It was shown that the automatic *ISI-citation* based strategies and the *Expert Voting* strategy are a significant improvement over the PubMed recency baseline and are a strong contender to the strategies based on query frames and annotated data developed by Demner-Fushman and Lin (2007).

9.2 Contributions

The main contribution of this thesis are new automated QA methods for clinical comparison questions posed in natural language, which are implemented in the RetroRank system. Enabling natural language input removes the burden of having to translate clincial questions into PICO query frames and is a more natural way of asking questions. RetroRank implements new post-retrieval ranking strategies and achieves a significant improvement over the PubMed baseline and performs better or equally well than previous approaches to post-retrieval ranking relying on query frames and annotated MEDLINE® data such as the approach by Demner-Fushman and Lin (2007). The performance of RetroRank shows that it is possible to successfully use natural language input and a fully automated approach to obtain answers to clinical drug comparison questions.

The RetroRank prototype addresses the problem time-pressed clinicians have when requiring answers to clinical comparison questions in a timely and concise manner. Clinicians need to keep up with a vast amount of ever changing research to be able to use the current best evidence in individual patient care (Sackett et al., 1996). While clinical search engines or electronic clinical decision support systems can be used to facilitate the retrieval and presentation of clinical evidence, there are limits concerning their usability and accessibility when timely guidance is of the essence and clinicans are reluctant to use electronic resources, because these present multiple problems. Converting a clinical question into a searchable strategy can be challenging and often leads to the retrieval of incomplete or non-useful information (Verhoeven et al., 1995; Davies, 2007). RetroRank facilitates the search process by using natural language questions to return relevant MEDLINE® abstracts that address the question.

The second contribution are two new evaluation corpora of clinical comparison questions with "gold standard" references, which can be used as test collections in future research in medical QA. Previously no "gold standard" corpora for clinical comparison question existed, which allowed an automated evaluation. The two new corpora are available on request.

9.3 Future Work

This section outlines future work which addresses the limitations of this research and presents further development of the ideas described in this thesis.

The main focus of this work is on developing new methods for answering clinical comparison questions and implementing a prototype QA system. This work only describes the back-end of RetroRank, while the front-end in form of a graphical user interface (GUI) was not addressed. Adding a user interface will be important to implement RetroRank as a publicly used system. The system architecture of RetroRank rank allows it to be used as part of a client-server application, where a central back-end server will be able to cache all MEDLINE® articles, perform the question processing and answer generation and the front-end GUI will present the results.

Another interesting avenue for future work concerns the display of the answer candidates retrieved by RetroRank using different forms of visualisation techniques such as multi-dimensional scaling or spring models (Morrison et al., 2003), that allow the user to group similar results and visualise the links between the abstracts.

As far as the performance of RetroRank is concerned, there are two components that can be augmented to improve the system. The first component is the retrieval module. The retrieval module in RetroRank employs a fairly unrestricted retrieval strategy to maximise recall and achieve a high retrieval accuracy. While a more restricted query has been found to negatively affect retrieval accuracy, it might still be worthwhile to consider expanding the query with additional information such as the dosage of a drug or methods of administering a drug, e.g., oral or topical to make the result set more relevant. In addition, a link between a generic drug name and its possible brand names in the lexicon has not been made in the current research. It might be useful to add such a link and introduce a retrieval step in which the alternative names for the drugs in the question are added to the query by using the associations between generic and brand names from. To implement an augmented drug recogniser, RxNorm[1], a normalized naming system for generic and branded drugs, which also includes drug strength and dose form, would be a valuable resource. However, with any change to the retrieval strategy, the trade-off between recall and precision will need to be considered. In addition the use of a relevance ranking engine such as Lucence[2], a search engine that implements TF-IDF weighting, could be explored as a stronger baseline for abstract retrieval instead of using PubMed's reverse chronological order as a baseline.

The second component of RetroRank that can be augmented is the post-retrieval module. The post-retrieval ranking strategies used in RetroRank are very successful but still do not produce the best result for all questions. While the "isi" strategy is

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

[2] `http://lucene.apache.org/core/`

a very strong strategy, there is potential to enhance it. One strategy that could be explored is to weigh the citation count with the publication date to favour more recent articles similar to the strategy used by Demner-Fushman and Lin (2007). It might also be worth investigating if adding a weight for the journal impact factor leads to a performance gain. Another strategy could be to investigate if a newer article with a lower citation count references a top ranked article, which is older and has a higher citation count and to boost its position by factor "x" if it does as shown in Example (9.1). Factor "x" would need to be determined in future experiments.

(9.1)

"isi" citation count article "A" $>$ "isi" citation count Article "B"
"A" is newer and references "B"
"B" is a top-ranked article
"A" is in the top 20 but ranked lower than "B"
The rank of "A" should be improved by factor "x"

Another way to augment the "isi" stategy could be to look at the articles related to the top ranked articles, which are published in high-impact journals and have a high "isi" citation count. Potentially this approach could be extended to other external sources, because it was shown that the strategy that used external sources performed significantly better than other strategies.

In conclusion, this thesis addressed a new research question concerning an automated natural language approach for answering clinical comparison questions. The results show that the prototype of an end-to-end QA system implemented in RetroRank achieves a significant improvement over the PubMed recency baseline and performs better or equally well than previous approaches to post-retrieval ranking using query frames and annotated data. The performance of RetroRank shows that it is possible to successfully use natural language input and a fully automated approach to obtain answers to clinical drug comparison questions. It was demonstrated that expert domain knowlege can be successfully implemented in an automated system by using external sources such as the ISI Web of Science, which provides high quality clinical evidence. The work in this thesis opens new avenues for future research on automated retrieval and post-retrieval ranking strategies for therapy questions in the clinical domain. There is also scope to extend the approach developed in this thesis to answer questions about other clinical tasks such as treatment questions.

Appendix A

Comparative Questions used by Demner-Fushman (2006)

Questions in the FPIN collection

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

What is the best treatment for analgesic rebound headaches?

First- or second-generation antihistamines: which are more eective at controlling pruritus?

What is the most eective nicotine replacement therapy?

(P) What are the best medications for panic disorder?

What is the most eective treatment for ADHD in children?

Other than anticoagulation, what is the best therapy for those with atrial fibrillation?

Do acetaminophen and an NSAID combined relieve osteoarthritis pain better than either alone?

(P) Whats the best treatment for epididymitis?

Is the ThinPrep better than conventional Pap smear at detecting cervical cancer?

Appendix B

Patterns Used by Fiszman et al. (2007) to Augment SemRep

comp1: Compared terms

C1: Term1 BE compare with/to Term2

C2: compare Term1 with/to Term2

C3: compare Term1 and/versus Term2

C4a: Term1 comparison with/to Term2

C4b: comparison of Term1 with/to Term2

C4c: comparison of Term1 and/versus Term2

C5 Term1 versus Term2

comp2: Scalar patterns

S1: Term1 BE as ADJ as BE Term2

S2a: Term1 BE more ADJ than BE Term2

S2b: Term1 BE ADJer than BETerm2

S2c: Term1 BE less ADJ than BE Term2

S4: Term1 BE superior to Term2

S5: Term1 BE inferior to Term2

{BE} means that some form of be is optional. The slash indicates disjunction.

Appendix C

List of PubMed stop words

	Stopwords
A	a, about, again, all, almost, also, although, always, among, an, and, another, any, are, as, at
B	be, because, been, before, being, between, both, but, by
C	can, could
D	did, do, does, done, due, during
E	each, either, enough, especially, etc
F	for, found, from, further
H	had, has, have, having, here, how, however
I	i, if, in, into, is, it, its, itself
J	just
K	kg, km
M	made, mainly, make, may, mg, might, ml, mm, most, mostly, must
N	nearly, neither, no, nor
O	obtained, of, often, on, our, overall
P	perhaps, pmid
Q	quite
R	rather, really, regarding
S	seem, seen, several, should, show, showed, shown, shows, significantly, since, so, some, such
T	than, that, the, their, theirs, them, then, there, therefore, these, they, this, those, through, thus, to
U	upon, use, used, using
V	various, very
W	was, we, were, what, when, which, while, with, within, without, would

Figure C.1: List of PubMed stopwords[a].

[a] http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helppubmed&part=pubmedhelp&rendertype=table&id=pubmedhelp.T43

Appendix D

Corpus 2

s1 How safe and effective are aspirin and warfarin therapy in the prevention of stroke in patients with atrial fibrillation?

s2 In adults or children with moderate to severe atopic dermatitis, is either tacrolimus (Protopic) or pimecrolimus (Elidel) more effective than topical corticosteroids?

s3 In patients with chronic obstructive pulmonary disease, do anticholinergics provide better benefit than beta-2 agonists?

s4 Do intranasal steroids control symptoms of allergic rhinitis better than antihistamines?

s5 Which nasal spray, budesonide (Rhinocort) or fluticasone propionate (Flonase), is superior for once daily treatment of allergic rhinitis?

s6 Is the long-acting anticholinergic drug tiotropium more effective than salmeterol in patients with chronic obstructive pulmonary disease?

s7 Is eletriptan more effective and at least as safe as sumatriptan for the treatment of acute migraine headache?

s8 Is the analgesic dextropropoxyphene-acetaminophen more effective than acetaminophen alone?

s9 Which is better for the treatment of pneumonia in hospitalized patients: levofloxacin (Levaquin) or ceftriaxone (Rocephin)?

s10 Is enoxaparin as effective as unfractionated heparin in patients with non-ST-segment elevation acute coronary syndromes?

s11 In patients with acute ST-elevation myocardial infarction who receive fibrinolysis and subsequent percutaneous intervention, is enoxaparin superior to unfractionated heparin?

s12 Is carvedilol better than metoprolol in the treatment of chronic heart failure?

s13 Is oral ketoprofen an effective treatment for acute migraine, and how does it compare with zolmitriptan?

s14 Is oxybutynin (Ditropan XL) or tolterodine (Detrol) more effective in the treatment of overactive bladder?

s15 Is ximelagatran as effective as warfarin in preventing stroke in patients with nonvalvular atrial fibrillation?

s16 Is ivermectin more effective than lindane for treating scabies?

s17 Which is better tolerated: Tolterodine (Detrol) or oxybutynin (Ditropan)?

s18 Can stress ulcers in critical patients be prevented with the use of sucralfate or ranitidine?

s19 Is the long-acting anticholinergic drug tiotropium more effective than salmeterol in patients with chronic obstructive pulmonary disease?

s20 Is cilostazol (Pletal) more effective than pentoxifylline (Trental) in the treatment of symptoms of intermittent claudication?

s21 Which intranasal formulation is most effective in the treatment of acute mi-

graine: sumatriptan (Imitrex) or dihydroergotamine (DHE)?

s22 Is losartan comparable to captopril in CHF?

s23 Is either omeprazole or cisapride effective in the control of heartburn symptoms?

s24 Which inhaled corticosteroid is most effective in the treatment of persistent asthma: fluticasone (Flovent) or beclomethasone (Beclovent, Vanceril)?

s25 Which is more effective for the treatment of Crohn's disease, a controlled release form of budesonide or a slow release form of mesalamine

s26 Is tramadol (Ultram) more effective than hydrocodone-Acetaminophen (Vicodin) in the treatment of acute musculoskeletal pain?

s27 Is losartan (Cozaar) more effective at preventing bad outcomes than atenolol (Tenormin) in patients with isolated systolic hypertension and left ventricular hypertrophy?

s28 Does fondaparinux improve outcomes better than enoxaparin in patients with acute coronary syndrome?

s29 Is hydroxyurea or anagrelide more effective for the treatment of essential thrombocythemia?

s30 Is caspofungin a safe and effective alternative to amphotericin B for invasive candida infections?

Appendix E

Extract from List of WHO INN Stems 2009

ALPHABETICAL LIST OF COMMON STEMS AND THEIR DEFINITION

A

-abine (see -arabine and -citabine)	arabinofuranosyl derivatives; nucleoside antiviral or antineoplastic agents, cytarabine or azactidine derivatives
-ac	anti-inflammatory agents, ibufenac derivatives
-acetam (see -racetam)	amide type nootrope agents, piracetam derivatives
-actide	synthetic polypeptide with a corticotropin-like action
-adol/-adol-	analgesics
-adom	analgesics, tifluadom derivatives
-afenone	antiarrhythmics, propafenone derivatives
-afil	inhibitors of phosphodiesterase PDE5 with vasodilator action
-aj-	antiarrhythmics, ajmaline derivatives
-al	aldehydes
-aldrate	antacids, aluminium salts
-alol (see -olol)	aromatic ring related to -olols
-alox (see -ox)	antacids, aluminium derivatives
-amivir (see vir)	neuraminidase inhibitors
-ampanel	antagonists of the ionotropic non-NMDA (*N*-methyl-D-aspartate) glutamate receptors (Namely the AMPA (amino-hydroxymethyl-isoxazole-propionic acid) and/or KA (kainite antagonist)

Figure E.1: Extract from List of WHO INN Stems 2009[a].

[a] www.who.int/medicines/services/inn/StemBook2009.pdf

Appendix F

Corpus 3

CD000067 What is the efficacy of azathioprine compared to 6-mercaptopurine for maintenance of remission in quiescent Crohn's disease?

CD000127 What are the effects of magnesium sulphate compared with diazepam when used for the care of women with eclampsia?

CD000128 What are the effects of magnesium sulphate versus phenytoin for eclampsia?

CD000284 What are the effects of chlorpromazine for schizophreniain comparison with placebo?

CD001031 How does lamotrigine compare to carbamazepine monotherapy for epilepsy?

CD001100 What is the effect of fixed dose subcutaneous low molecular weight heparins versus adjusted dose unfractionated heparin for venous thromboembolism?

CD001211 What are the benefits and harms of antibiotics versus placebo for acute bacterial conjunctivitis?

CD001281 What are the comparative efficacy safety and side-effects of long-acting beta-2 agonists and theophylline in the maintenance treatment of adults and adolescents with asthma?

CD001387 What is the efficacy and safety of ipratropium bromide versus short acting beta-2 agonists for stable chronic obstructive pulmonary disease?

CD001439 Are there benefits of using antibiotics instead of placebo for the prevention of postoperative infection after an appendicectomy?

CD001895 What is the effectiveness and acceptability of progestogens alone and oestrogens and progestogens in combination in the management of irregular bleeding associated with anovulation?

CD002258 What is the efficacy and safety of bromocriptine monotherapy for delaying the onset of motor complications associated with levodopa therapy in patients with Parkinson's disease?

CD002310 What is the the efficacy and safety of fluticasone versus beclomethasone or budesonide for chronic asthma in adults and children?

CD002314 What is the safety and efficacy of anti-leukotriene agents compared to inhaled glucocorticoids?

CD002738 What is the efficacy of beclomethasone compared to placebo for chronic asthma?

CD002960 What are the effects of magnesium sulphate compared with lytic cocktail when used for the care of women with eclampsia?

CD003082 What are the clinical effects of haloperidol for the management of schizophreniaand other similar serious mental illnesses compared to placebo?

CD003135 How does the efficacy and safety of Fluticasone compare to the use of a placebo for chronic asthma in adults and children?

CD003462 What are the effects of heparin versus placebo for acute coronary syndromes?

CD003492 What are the effects of lithium versus antidepressants for the long-term treatment of unipolar affective disorder?

CD003530 How does beclomethasone compare to budesonide for chronic asthma?

CD003615 What is the best evidence for oxcarbazepine versus phenytoin monotherapy for epilepsy?

CD004278 How does haloperidol compare to chlorpromazine for people with schizophrenia?

CD004707 What are the benefits and harms of voriconazole compared to amphotericin B when used for prevention or treatment of invasive fungal infectionsin cancerpatients with neutropenia?

CD005154 What are the effects of adenosine versus intravenous calcium channel antagonists for the treatment of supraventricular tachycardia in adults?

CD005309 How does fluticasone compare to extrafine HFA beclomethasone dipropionate for chronicasthmain adults and children?

CD005535 What are the effects of addition of long-acting beta2-agonists to inhaled corticosteroids versus same dose inhaled corticosteroids for chronic asthma in adults and children?

CD005967 How does artesunate compare to quinine for treating severe malaria?

CD006015 How do the clinical effectiveness and harms of finasteride compare to placebo and active controls in the treatment of benign prostatic hyperplasia?

CD006114 How does fluvoxamine compare to other anti-depressive agents for depressionin terms of effectiveness, tolerability and side effects?

CD006117 What is the the efficacy acceptability and tolerability of sertraline in

comparison with other antidepressive agents for depression?

CD006217 What is the efficacy of ciclesonide versus placebo for chronicasthmain adults and children?

CD006300 What is the possible effectiveness of cyclophosphamide compared with that of ifosfamide for paediatric and young adult patients with sarcoma?

CD006352 What are the effects of oral fluphenazine for schizophreniain comparison with placebo?

CD006453 How do the efficacy and tolerability of carbamazepine and oxcarbazepine monotherapy compare for partial onset seizures?

CD006626 What are the effects of risperidone versus other atypical antipsychotics for schizophrenia?

CD006628 What are the effects of zotepine compared with other second generation antipsychotic drugs for people suffering from schizophrenia?

CD006654 What are the effects of olanzapine versus other atypical antipsychotics for schizophrenia?

CD006918 What are the clinical effects of oral risperidone for people with schizophrenia and schizophrenia-like psychosesin comparison with placebo?

CD007022 What is the efficacy and safety of vancomycin versus teicoplanin in patients with proven or suspected infection?

CD007570 Is Lactulose or Polyethylene Glycol more effective at treating chronicconstipationand faecal impaction?

CD007695 What are the serious adverse effects for regular treatment with formoterol versus regular treatment with salmeterol for chronic asthma?

CD007811 What are the effects of sulpiride for schizophreniaand other similar serious mental illnesses in comparison with placebo?

CD007891 What are the effects of combination inhaled steroid and long-acting beta2-agonist therapy versus tiotropium for chronic obstructive pulmonary disease?

CD008418 How efficient and safe is Formoterol versus short-acting beta-agonists as relief medication for adults and children with asthma?

Appendix G

POEM System Comparison

ID	Abstracts	recency	tf	tf/idf	Last 20%	google	isi	msw	tf-isi voting	voting	AskMedline	Entrez
s1	524	344	346	75	98	405	3	49	3	9	n/a	n/a
s5	59	45	11	4	12	15	12	8	12	9	2.00	6
s6	79	77	63	56	60	2	2	42	2	2	n/a	n/a
s7	61	54	37	5	25	n/a	5	23	5	11	71.00	n/a
s9	34	29	5	9	5	3	3	4	3	3	7.00	n/a
s10	321	161	118	60	260	52	7	213	7	27	23.00	11
s11	452	109	37	34	9	61	50	14	50	34	1.00	1
s12	150	73	69	16	70	2	2	57	2	1	60.00	n/a
s13	3	1	3	2	3	n/a	2	2	2	1	1.00	n/a
s14	153	136	58	18	40	10	7	26	7	22	35.00	43
s16	41	30	5	20	5	23	12	2	12	4	6.00	n/a
s17	213	183	35	21	73	26	12	32	12	45	n/a	610
s18	48	16	14	6	41	3	4	37	4	8	10.00	1288
s19	79	77	63	56	60	2	2	42	2	2	n/a	n/a
s20	60	55	30	7	16	5	2	6	2	4	n/a	n/a
s23	10	3	2	2	4	5	4	10	4	2	1.00	n/a
s24	286	222	141	21	189	80	69	91	69	67	n/a	1194
s25	52	41	5	3	20	n/a	2	10	2	2	41.00	n/a
s27	153	121	88	20	64	28	10	79	10	36	n/a	n/a
s29	125	48	32	2	61	6	3	88	3	2	n/a	1314
s30	163	163	94	3	123	2	2	67	2	1	139.00	n/a
Avg. Rank		94.667	59.810	20.952	58.952	40.556	10.238	42.952	10.238	13.905	30.538	558.375
Improvement			34.857	73.714	35.714	54.111	84.429	51.714	84.429	80.762	64.128	-463.708
Improvement (%)			0.368	0.779	0.377	0.572	0.892	0.546	0.892	0.853	0.677	-4.898

Table G.1: POEM System Comparison

Bibliography

Aronson, A. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proc AMIA 2001*, pages 17–21.

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

Bernstam, E., Herskovic, J., Aphinyanaphongs, Y., Aliferis, C., Sriram, M., and Hersh, W. (2006). Using Citation Data to Improve Retrieval from MEDLINE. *Journal of the American Medical Informatics Association*, 13:96–105.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*.

Bryant, L. S. and Ringrose, T. (2005). Clinical question answering services: What users want and what providers provide. Poster.

Buckley, C., Singhal, A., and M., M. (1995). New retrieval approaches using SMART: TREC 4. In Harman, D., editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, NIST Special Publication, pages 25–48.

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

Chambliss, M. L. and Conley, J. (1996). Answering clinical questions. *J Fam Pract*, 43(2):140–144.

Collaboration, T. C. (2011). About the cochrane library.

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.

Davies, K. (2007). The information-seeking behaviour of doctors: a review of the evidence. *Health Information and Libraries Journal*, 24(2):78–94.

Davies, K. (2009). Quantifying the information needs of doctors in the UK using clinical librarians. *Health Information and Libraries Journal*, 26(4):289–297.

Davies, K. (2011). Physicians and their use of information: a survey comparison between the United States, Canada, and the United Kingdom. *Journal of the Medical Library Association : JMLA*, 99(1).

Demner-Fushman, D. (2006). *Complex question answering based on a semantic domain model of clinical medicine.* PhD thesis, University of Maryland at College Park, USA.

Demner-Fushman, D. and Lin, J. (2005). Knowledge extraction for clinical question answering: Preliminary results. In *Proc. AAAI-05 Workshop on Question Answering in Restricted Domains*, pages 1–10.

Demner-Fushman, D. and Lin, J. (2007). Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics*, 33(1):63–103.

Demner-Fushman, D. D. and Lin, J. (2006). Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 841–848, Morristown, NJ, USA. Association for Computational Linguistics.

Ebell, M. H. (2009). How to find answers to clinical questions. *American Family Physician*, 79(4):293–296.

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. *Journal of the American Board of Family Practice*, 17(1):59–67.

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. *J Am Med Inform Assoc*, 12(2):217–224.

Ely, J. W., Osheroff, J. A., Ebell, M. H., Bergus, G. R., Levy, B. T., Chambliss, L. M., and Evans, E. R. (1999). Analysis of questions asked by family doctors regarding patient care. *BMJ*, 319(7206):358–361.

Ely, J. W., Osheroff, J. A., Ebell, M. H., Chambliss, M. L., Vinson, D. C., Stevermer, J. J., and Pifer, E. A. (2002). Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *BMJ*, 324(7339).

Fiszman, M., Demner-Fushman, D., Lang, F. M., Goetz, P., and Rindflesch, T. C. (2007). Interpreting comparative constructions in biomedical text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, pages 137–144, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5):378–382.

Fontelo, P., Liu, F., and Ackerman, M. (2005). askMEDLINE: a free-text, natural language query tool for MEDLINE/PubMed. *BMC Med Inform Decis Mak*, 5(1).

Garfield, E. (1990). How ISI selects journals for coverage: Quantitative and qualitative considerations. *Current Contents*, 22:5–13.

Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J. (1998). Indexing with WordNet synsets can improve text retrieval. *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*.

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.

Grandage, K., Slawson, D., and Shaughnessy, A. (2002). When less is more: a practical approach to searching for evidence-based answers. *Journal of the Medical Library Association*, 90(3):298–304.

Harzing, A.-W. K. and van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, 8:61–73.

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 : JAMIA*, 1(6):447–458.

Huddleston, R. and Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge (UK) [etc.].

Jasco, P. (2006). Deflated, inflated and phantom citation counts. *Online Information Review*, 30(3):297–309.

Jasco, P. (2008). Google Scholar revisited. *Online Information Review*, 32(1):102–114.

Jindal, N. and Liu, B. (2006a). Identifying comparative sentences in text documents. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 244–251, New York, NY, USA. ACM Press.

Jindal, N. and Liu, B. (2006b). Mining comparative sentences and relations. In *Proceedings of the 21st national conference on Artificial intelligence - Volume 2*, pages 1331–1336. AAAI Press.

Kwok, K. L., L., G., Dinstl, N., and Chan, M. (2000). TREC-9 cross language, web and question-answering track experiments using PIRCS. In *Proceedings of the Text Retrieval Conference (TREC-9*, pages 26–35.

Lu, Z., Kim, W., and Wilbur, W. (2009). Evaluation of query expansion using mesh in pubmed. *Information Retrieval*, 12:69–80. 10.1007/s10791-008-9074-8.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2).

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, 1st edition.

Monz, C. (2004). Minimal Span Weighting Retrieval for Question Answering. In *Proceedings of Information Retireval for Question Answering Workshop (SIGIR 04)*, pages 23–30.

Morrison, A., Ross, G., and Chalmers, M. (2003). Fast multidimensional scaling through sampling, springs and interpolation. *Information Visualization*, 2(1):68–77.

Richardson, W. S. and Wilson, M. C. (1997). On questions, background and foreground. *Evidence-Based Healthcare Newsletter*, November 6.

Richardson, W. S., Wilson, M. C., Nishikawa, J., and Hayward, R. S. (1995). The well-built clinical question: a key to evidence-based decisions. *ACP J Club*, 123(3).

Rindflesch, T., Fiszman, M., and Libbus, B. (2005). Semantic Interpretation for the Biomedical Research Literature. In *Medical Informatics*, pages 399–422.

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.

Roediger, H. I. (2006). The h index in science: a new measure of scholarly contribution. *The Academic Observer*, 19(4).

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72.

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 (Book with CD-ROM)*. Churchill Livingstone, 2nd edition.

Sarawagi, S. (2004). Crf project page.

Scheible, S. (2008). Annotating superlatives. In Calzolari, N., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 28–30. European Language Resources Association (ELRA).

Segurabedmar, I., Martinez, P., and Segurabedmar, M. (2008). Drug name recognition and classification in biomedical textsA case study outlining approaches underpinning automated systems. *Drug Discovery Today*, 13(17-18):816–823.

Slawson, D. C., Shaughnessy, A. F., and Bennett, J. H. (1994). Becoming a medical information master: feeling good about not knowing everything. *J Fam Pract*, 38(5):505–513.

Smith, R. (1996). What clinical information do doctors need? *BMJ*, 313(7064):1062–1068.

Smith, R. (2002). A POEM a Week for the BMJ. *BMJ*, 325(7371):983.

Verhoeven, A. A., Boerma, E. J., and Meyboom-deJong, B. (1995). Use of information sources by family physicians: A literature survey. *Bulletin of the Medical Library Association*, 83(1):85–90.

Vincent, S. (2006). National service standards for clinical question answering services (CQAS). Presentation. Presentation.

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 61–69, New York, NY, USA. Springer-Verlag New York, Inc.

Ward, L. (2005). Survey of UK clinical librarians june 2005. final report and contribution to the audit of rapid response clinical question answering services in England and Wales 2005.

Zhao, Y. and Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 515–524, New York, NY, USA. ACM.