View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by RMIT Research Repository

IMPROVING PATIENT RECORDS SEARCH

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Iman Amini
Ms.CS RMIT University,
School of Science,
College of Science, Engineering, and Health,
RMIT University,
Melbourne, Victoria, Australia.

Supervisors:
Mark Sanderson, David Martinez, Xiaodong Li

June 29, 2017

RMIT

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; and, any editorial work, paid or unpaid, carried out by a third party is acknowledged. I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Iman Amini
School of Science
RMIT University
June 29, 2017

Acknowledgments

This journey would have been impossible without the consistent support of my respected supervisors Prof. Mark Sanderson and Dr. David Martinez. I am very grateful to have the experience of working under their supervision that not only taught me to do research but also to be an outstanding teacher and supervisor in the years to come, for which I can not thank them enough. More importantly, Mark and David have been my true inspiration to aim high, work hard, and pay attention to important details for that shall result in quality research to impact the community and beyond. Despite my lack of knowledge in the field I was never made to feel that limitation in their presence which gave me a sense of teamwork at times, and allowed me to share, openly argue and discuss my ideas.

I am also thankful to the valuable help and support of the following people. Xiaodong Li, who later joined the panel of supervision, my lovely parents Farhad and Leila, and my dear brother Amin. I would like to thank my adorable relatives Feridoon and Nina, and my close friend Marge Fernando whose kind and consistent help from the start of my PhD through to the end has been praiseworthy. I would like to thank my landlord and good friend Steven Crowl whose beautiful house I had the pleasure to live in since the start of my PhD.

Credits

Portions of the material in this thesis have previously appeared in our following publications:

- I. Amini, M. Sanderson, D. Martinez, and X. Li. Search for Clinical Records: RMIT at TREC 2011 Medical Track. In *Proceedings of Text Retrieval Conference*, 2011
- I. Amini, M. Sanderson, D. Martinez, and X. Li. Using Meta-data to search for Clinical Records: RMIT at TREC 2012 Medical Track. In *Proceedings of Text Retrieval Conference*, 2012b
- I. Amini, D. Martinez, and D. Molla. Overview of the ALTA 2012 Shared Task. In *Proceedings of ALTA 2012*, volume 7, pages 7–9, 2012a
- D. Mollá, D. Martinez, and I. Amini. Towards information retrieval evaluation with reduced and only positive judgements. In *Proceedings of the 18th Australasian Document Computing Symposium*, pages 109–112. ACM, 2013
- D. Mollá, I. Amini, and D. Martinez. Document Distance for the Automated Expansion of Relevance Judgements for Information Retrieval Evaluation. In *ACM SIGIR Workshop on Gathering Efficient Assessments of Relevance (GEAR)*, 2014
- I. Amini, D Martinez, X Li, and M Sanderson. Improving patient record search: A Meta-data Based Approach. Journal of *Information Processing & Management (2015)*

The thesis was written in the Eclipse editor on Mac OS X EL Capitan, and typeset using the LaTeX 2_ε document preparation system.
All trademarks are the property of their respective owners.

Contents

List of Figures

List of Tables

Abstract

Improving health search is a wide context which concerns the effectiveness of Information Retrieval (IR) systems (also called search engines) while providing grounds for the creation of reliable test collections. In this research we analyse IR and Text Processing methods to improve health search mainly that of Electronic Patient Records (EPR). We also propose a novel approach to evaluate IR systems, that unlike traditional IR evaluation does not rely on human relevance judgement. We find that our meta-data based method is more effective than query expansion using external knowledge sources, and that our simulated relevance judgments have a positive correlation with man-made relevance judgements.

CHAPTER 1

Introduction

The science of finding digital information, conventionally known as Information Retrieval (IR), and the emergence of search engines have brought a new dimension beyond the previously limited database search and structured query language. Search has become an inevitable part of a large number of software applications, where a pattern or a query given by the user is used to initiate a search.

Search over various kinds of information, including image or video, has been made possible over the recent years, but the most common form of retrieval is textual search. Here the task involved is a kind of string matching between a given pattern, commonly known as a query, and a set of documents.

Text on its own can be categorised into various types that one may perform search on. Each might have certain properties that are important to consider from the search engine's perspective such as the structure, language, etc. While there are certain techniques that can generally be applied to all kinds of texts, there are some characteristics of certain types of text (ex. health records) that set them apart.

Health search, in particular, has been of growing interest in recent years. According to a study by the Pew Research Center, 72 percent of U.S Internet users, in the past year, have used public search engines to carry out health related search [Pew 2012].

Although the scattered health data in the world wide web are mixed and indexed together with other types of data, it has been shown that health data embody certain characteristics implying the need for a different way of handling it. This can be observed from the recent health-related shared tasks [Voorhees and Tong 2011, Voorhees and Hersh 2012, Goeuriot et al. 2013; 2014b]. For instance, the involvement of Natural Language Processing (NLP) in the retrieval of health text is considered important so as to identify negative statements in text (More on handling negation in Chapter 4).

Within the health domain, we particularly focus on two types of textual health data, which have been largely used by the IR and NLP community in the past. The two types vary in terms of the structure and the style of the language they are written in. The first type is clinical records, created by physicians, which are more inconsistent in language and grammatical correctness. The second type is academic papers that are written by health researchers with more attention on consistency and grammatical correctness.

Clinical records in IR research can be used to build a system where academic researchers can look for potential case studies to form a patient cohort. More importantly, these records can be exploited for other applications such as fact extraction, clinical summarization and question answering. In 2011 the well-known venue for testing and evaluating IR systems, TREC, organised two shared tasks (tracks) in this domain. We discuss the details in the subsequent chapters.

Academic papers have been one of the main sources for finding answers to clinical queries where medical practitioners are advised to base their decisions on

the latest medical literature. Using the literature for the purpose of medical decision making, and critically assessing and applying it to patient-care are part of Evidence Based Medicine (EBM) [Sackett et al. 1996].

Whether we are concerned with the retrieval of clinical records or any types of data, our ultimate goal is to satisfy end users. In other words, we want to retrieve as many relevant documents as possible, and retrieve them early on in a ranked list of results, before the retrieval of non-relevant documents.

Plenty of IR metrics have been widely used, amongst which precision and recall. They correspond to the level of accuracy (Precision), which is about retrieving relevant documents prior to the non-relevant documents, and comprehensiveness (Recall), which pertains to the retrieval of all the relevant documents that exist in the collection.

However, given the fact that the most popular IR metrics are based on a judgement of a subset of relevant documents, ensuring reliable IR evaluation is itself a demanding research area.

Moreover gathering a set of ground truth to determine the state of a retrieved document as relevant or non-relevant requires much manual work. This is an area in IR which, to this day purely relies on human intervention, and few attempts have been made to propose an automated algorithm to replace it [Soboroff et al. 2001, Büttcher et al. 2007, Sakai and Lin 2010, Mollá et al. 2014].

We focus on two important aspects of the health search in this thesis. Improving IR effectiveness (see Chapter 4 and 5) to ensure the retrieval of relevant documents, and to provide grounds for the creation of more query relevance judgements (qrel) by suggesting automatic (see Chapter 6) and semi-automatic approaches (see Chapter B and appndx3).

We envisage a situation where scientists reduce their search frequency in finding relevant documents, both saving time and allowing the exploitation of the bulk of

textual data available in digital format.

This thesis is an endeavour to address the following research questions.

1. Patient Cohort Search: How to effectively find patient cohorts for research studies. (Effective Retrieval)

2. What are the possible ways to alleviate manual query relevance judgement for health data. (Effective Evaluation)

To tackle the first research question, we compare a classical method of expanding queries using external knowledge sources with a novel method which exploits a form of a meta-data in the underlying collection. We discuss the meta-data and explain the details of our novel query expansion methodology.

Our second research question is an attempt to provide grounds for the creation of more clinical test collections by building semi-automatic relevance judgement.

1.1 Summary

Chapter 2 has an overview of the related past research, with references to the state of art. Moreover, it provides details of the datasets that we used to perform our experiments.

The clinical dataset that was used to carry out the experiments in Chapters 4, 5 and 6 was the only publicly available clinical IR test collection at the time, and due to the raising confidentiality concerns, the owners of this dataset decided to withdraw all the granted permissions upon the completion of the shared task, preventing further experimentation on this IR test collection. The sensitive nature of this kind of data has unfortunately led to a high degree of data scarcity in this field of IR. This dataset is explained in Chapter 3.

In Chapter 4 we focus on properties and distinct characteristics of clinical records to build an effective setting for the retrieval of clinical records. A query expansion approach is introduced here that makes use of some publicly available knowledge sources to further improve the effectiveness of the underlying IR system.

Chapter 5 is a highlight of a type of meta-data that are present in the collection. This meta-data is used in a novel way to enhance the effectiveness of IR search engines. This work is a more promising alternative to the external query expansion we introduced in the previous chapter. Here we show that these meta-data have a unique characteristic that can be exploited for significantly improving IR performances.

In Chapter 6, after the success in the integration of meta-data to improve the performance of IR systems, we explored a different application with the same meta-data. We perform a set of experiments to find out whether they can be used to judge the relevancy of the medical records for the given set of queries. Here we map the queries to their equivalent set of meta-data and, depending on the overlap with the meta-data in the collection and queries, judge the relevancy of documents.

Appendix A describes a shared task that we organised in 2012 on sentence classification for the problem of Evidence Based Medicine [Amini et al. 2012a].

This research has been published but it represent a work in progress and the following is our planned future work.

We built on the previous work, in the context of sentence classification [Boudin et al. 2010, Chung 2009, Kim et al. 2011, Demner-Fushman and Lin 2007] by using a new framework called PIBOO which stands for: Population/Problem, Intervention, Background, Outcome, Other. Every label represents one of the five conventional categories that a sentence from an abstract may belong to. Our intention of implementing this framework is to explore one way to find different aspects of a medical abstract and be able to generate automatic queries which will, in turn, be run against our collection of clinical records. The goal is to find ways to effectively establish

connections between two different health related corpora: academic papers and electronic clinical records, and to identify particular information that doctors are unable to find from publicly available medical collections.

Appendix B, similar to Chapter 6 concerns the automatic creation of relevance judgement. However, here instead of producing judgements from scratch we extend a limited manual judgement using a document distance-based approach, where the dataset is a collection of medical abstracts [Hersh et al. 1994]. In this work, we demonstrate the value of evaluation based on expanded relevance judgements.

Appendix C is an extension to Appendix B where we show that this approach improves the quality of evaluation *both* for medical and news reports, and we, therefore, add further evidence of the plausibility of this method.

Note that all the retrieval experiments rely on the Terrier open source search engine [Macdonald et al. 2012].

The entire work presented in this thesis, are based on our following publications with respect to their relevant chapters:

1. Chapter 4: Search for Clinical Records : RMIT at Medical TREC 2011 [Amini et al. 2011]

2. Chapter 5: Using Meta-data to search for Clinical Records : RMIT at Medical TREC 2012 [Amini et al. 2012b]

3. Appendix A: Overview of the ALTA 2012 Shared Task [Amini et al. 2012a]

4. Appendix B: Towards Information Retrieval Evaluation with Reduced and Only Positive Judgements: ALTA [Mollá et al. 2013]

5. Appendix C: Document Distance for the Automated Expansion of Relevance Judgements: SIGIR (GEAR) workshop [Mollá et al. 2014]

6. Chapter 5 and 6: Improving Patient Record Search : Journal of IP&M [Amini et al. 2011]

CHAPTER 2

Background

Regardless of the context of search, in IR, searching requires a pattern and an underlying space of objects where the actual search takes place. While the idea remains the same for all types of search, a more fine-grained search system may be more effective based on the type or the context of search. Therefore most of the health search engines are customised to the characteristics of health-related data.

Health search is generally a broad concept which concerns various types of data, such as clinical notes and academic papers. Focusing on textual data, health search in this thesis is mainly concerned with the textual contents taken from medical articles (academic journals) and doctors' notes (clinical records).

Medical articles have been the main source for finding medical answers where doctors and medical practitioners are encouraged to base their decisions on the latest research [Haynes et al. 2002]. There has been a considerable amount of research on ways to facilitate EBM through text processing and information retrieval. An example of such research is the classification of sentences (journal abstracts) [Kim et al.

2011, Huang et al. 2011] to the predefined Patient, Intervention, Comparison, Outcome (PICO) categories, in order to improve the retrieval of medical papers [Boudin et al. 2010].

Other outstanding work on the collection of academic papers include identifying relationships between concepts and terminologies, such as genes, proteins or drugs in the biomedical literature i.e. scientific papers [Cohen et al. 2005, Yu et al. 2002, Yu and Agichtein 2003]. Similar to this work was called the "Hypothesis Generation" proposed by Swanson [Swanson 1988] which infers medical hypothesis and uncovers unknown relationships by looking at large collections of biomedical literature. Swanson performed this task manually for the first time, however, several researchers have tried to automate it [Gordon and Lindsay 1996, Weeber et al. 2000].

While much research have been carried out on a collection of academic papers, the ability to conduct research on retrieval of clinical records has been limited in previous years due to the lack of a publicly available dataset of appropriate size. The bulk of the work in this area has been focused on Natural Language Processing challenges, such as extracting specific information from a small number of clinical records [Özlem Uzuner 2012], or data management and mining radiology reports [Apostolova et al. 2009].

The IR research on publicly available medical articles has a long history. Here, researchers have focused on improving search on medical literature [Bernstam et al. 2006, Hersh et al. 1994, Abdou and Savoy 2008]. Nonetheless, the integration of more recent IR techniques such as page rank, and using citation to rank search results is relatively recent and has made some improvements to the performance of health search systems in finding relevant and high-quality documents [Bernstam et al. 2006].

In 2011, a set of clinical records was used in an IR shared task which generated interest in search over Electronic Health Records [Voorhees and Tong 2011,

Voorhees and Hersh 2012]. This relatively large test collection promoted research on Patient Cohort Search which has been the main focus of this research.

This chapter gives an overview of the related past work and provides details of the datasets, evaluation and resources that we used to perform our experiments.

2.1 Biomedical Processing Resources and Semantic Search

In order to analyse, simplify and cluster health-related text, several resources have been developed and used extensively in the retrieval and text processing of health records in the past. Throughout the course of this research, the following resources will be used and referred to frequently: *MEDLINE, UMLS, PubMed, MetaMap, MeSH, SNOMED-CT, ICD codes*.

MEDLINE (Medical Literature Analysis and Retrieval System Online) is a bibliographic database containing academic journals from a range of fields in medicine and health care. The inclusion of journal articles is a supervised task based on a recommendation from a technical review committee. The content of MEDLINE is publicly searchable through a free search engine known as the *PubMed*.

To facilitate understanding and processing of the abundance of medical and health literature, the Unified Medical Language System (UMLS) [Bethesda 2009] was designed and maintained by the NLM (National Library of Medicine) in 1986. The UMLS was developed as a large collection of biomedical vocabularies, extensively used by medical informatics, providing a mapping structure between all the present vocabularies. In order to exploit the UMLS and automatically map medical text to the UMLS concepts the MetaMap program was developed by Aronson [Aronson 2001] in 2001. MetaMap is extensively used in the biomedical text processing and IR communities to map medical terms into concepts held in the (UMLS) Metathesaurus. This knowledge base integrates different controlled vocabularies,

such as ICD codes, Medical Subject Headings (MeSH), and Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT);[1] which we describe below.

From the collections of biomedical vocabularies present in the UMLS, some are of particular interest for the retrieval of health records. Medical Subject Headings (MeSH), for instance, is a controlled vocabulary from the National Library of Medicine[2]. MeSH was developed to aid the retrieval of journal abstracts in PubMed and is used in the indexing process. Every article is assigned 10 to 15 subject headings, and the PubMed interface automatically assigns subject headings to user queries for a more comprehensive and accurate retrieval.

Furthermore, ICD and SNOMED-CT are designed for different purposes. SNOMED-CT is a comprehensive clinical health terminology comprising a collection of medical terms covering diseases, findings, procedures, micro-organisms, substances, etc. It is used by health care providers to encode the meaning of health information. ICD, on the other hand, is a disease classification system, maintained by the World Health Organization (WHO) and used by coding professionals, mainly for morbidity and mortality statistics. ICD is periodically revised and is currently at the tenth version. In this thesis, we use the 9th version (ICD-9 codes) , as this was the available version at the time of running the experiments. ICD, MeSH and SNOMED-CT are the incorporated vocabularies in the UMLS database.

The utility of applying the aforementioned resources, although not specifically for patient records, have been extensively studied. Despite the clear advantage of exploiting semantic knowledge and conceptual search in clinical IR and biomedical retrieval [Büttcher et al. 2004], in general IR, the use of lexical-semantic relations has been shown to provide less benefit. Voorhees [Voorhees 1994] demonstrates that the effectiveness of semantic query expansion depends on the length of queries, such

[1] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
[2] http://www.nlm.nih.gov/mesh

that, longer and more detailed queries are not improved.

Koopman et al. [Koopman et al. 2011b] highlight the benefits of semantic search over keyword-based approach in the clinical domain. This work was done prior to the medical TREC but used the same records provided by the BLULAB NLP repository, albeit without the TREC queries and relevance judgements. Therefore, empirical evaluation was their first challenge. Out of 3249 queries, they selected 54 queries which had a significant number of relevance judgement. They included both general and specific queries.

SNOMED-CT was chosen as the only domain specific knowledge for its wide coverage and non-applicational focus. SNOMED-CT was proven successful for biomedical applications [Zhou et al. 2007, Liu and Chu 2007] before. Their term conversion into SNOMED-CT concepts was done in two steps: first by mapping terms to UMLS concepts using the Metamap and then querying their SNOMED-CT equivalents using the UMLS. They show that the existence of inference types (i.e. associational, deductive and abductive) between queries and relevant records, imply the need for a conceptual approach. In such cases, two terms can relate to each other without being synonymous.

Their concept-based system significantly outperformed a keyword-based baseline by 25% in map score. More importantly, their system obtained a small gain across a large number of queries which indicated its suitability for practical scenarios. As an extra experiment, by further dividing the queries into 2 sets of hard vs. easy groups, based on the baseline results, they found that the concept-based system performed much better (104% improvement) for the hard queries, compared to the baseline system.

One shortcoming of this work, however, is that some UMLS concepts did not map to any SNOMED-CT concept. If concept recognition tools such as the Metamap provide a direct approach for term conversion to SNOMED-CT concepts, however,

this issue is less likely to occur. However, the fact that a subset of queries performs better using the keyword-based approach indicates that ways to decide when to use conceptual versus keyword-based approach are needed to improve this work.

Ravindran and Gauch [Ravindran and Gauch 2004] go one step beyond, by combining keyword and conceptual similarity for ranking documents, using a dataset from the TREC web track. They utilise the hierarchical relationships between queries and documents using the *Open Directory Project* (ODP)[3]. They demonstrated improved results by filtering out retrieved documents that did not have the same hierarchical relationship with the queries.

In a more recent work, Koopman et al. [Koopman et al. 2016] proposed a novel semantic approach called the Graph Inference Model (GIN), where information units, be they, terms or concepts are represented as nodes. Information units from the documents and queries are connected through edges which are created based on existing relationships in the SNOMED-CT, or the co-occurrence statistics from the underlying corpus. Nodes in their graph have a many-to-many relationship.

In this approach, retrieval is an inference process where each node represents a document. The evidence for query-document similarity is taken from query nodes which are connected to document nodes in the graph.

Surprisingly, however, the initial results of the GIN do not show any significant improvement over the keyword-based baseline. Albeit, the analysis of failure, uncovers an important fact. They found that the GIN was fetching unassessed documents (Medical TREC reports) that have not been retrieved by TREC systems. By recruiting 4 medical assessors, a total of 1030 documents which were retrieved as top-20 documents by the GIN were assessed for relevancy. Indeed, 29% of these documents that were not initially assessed, were judged to be relevant. The extended version of the qrel which included the newly found relevant documents demonstrated that the

[3] http://www.dmoz.org

effectiveness of the GIN was significantly underestimated with the old qrel.

The weakness of the GIN is, however, that, although it tends to improve the effectiveness of a set of hard queries (i.e. queries with the lowest median scores in TREC), it degrades the performance of easy queries, by introducing noise. The development of an adaptive system to use inference per-query basis, is, however, left to future work.

An example of the use of UMLS was presented by Jain et al. [Jain et al. 2010], who proposed a framework for symptom-based retrieval of medical records. They used several sources of knowledge, including semantic relationships from the UMLS, and terms suggested by medical experts. Five hypothetical patient records were randomly generated to create a dataset. Each record contains 200 nursing assessment notes. Also, ten queries which related to various symptoms were selected by a medical expert. The results showed that the terms suggested by medical experts were the most effective source for query expansion, followed by clinically associated terms from the UMLS. However, the combination of all the expanded terms from different sources yielded the highest score.

There is a large body of research on applications of IR techniques on a subset of MEDLINE. Abdou and Savoy [Abdou and Savoy 2008] evaluated 10 different IR models on a large subset of MEDLINE, showing the effect of indexing MeSH on different ranking models [Amati and Van Rijsbergen 2002]. This work mainly focused on finding the best ranking algorithm and did not experiment automatic expansion of queries with MeSH terms. However, the effect of excluding and including MeSH terms to the index was tested for all ranking models.

They demonstrated an average increase of almost 8.5% when indexing MeSH terms, and found that some ranking models benefited more. Okapi, for instance, showed the greatest improvement of 11.1% in MAP scores.

Srinivasan [Srinivasan 1996] evaluated 3 query expansion methodologies on a

MEDLINE test collection, promoting a retrieval feedback technique to add MeSH terms to user queries. Hersh et al. [Hersh et al. 2000] found that only a subset of queries was improved by using relationships from a thesaurus-based approach for a subset of MEDLINE. They pointed out that future work must focus on finding instances of query expansion (based on word synonyms, hierarchies or related terms) which can improve retrieval.

Lu et al. [Lu et al. 2009] automatically mapped query words to MeSH terms to extend queries. This work was conducted on collections from two TREC genomics tracks (2006,2007) [Hersh et al. 2006; 2007]. They found that their replication of the PubMed's [4] Automatic Term Mapping (ATM) to MeSH terms was effective in finding more relevant documents, while, it did not improve the precision in top retrieved documents. Jalali and Borujerdi [Jalali and Borujerdi 2008] used synonyms of the identified MeSH terms or direct descendants of them as expansion terms. They reported improvement over various retrieval systems such as those which use general-purpose ontologies.

2.2 Test Collections

Since the early 1950s, the IR community has been producing test collections. Cleverdon and Salton [Cleverdon 1991, Lesk and Salton 1968] are amongst the first researchers to introduce test collections, giving rise to empirical evaluation of IR systems.

An IR test collection is generally composed of 3 main components: a set of queries (also called topics) where each query is given a unique identifier, a collection of documents of different content, and a set of gold standard (AKA qrel) determining which documents are relevant/non-relevant to the given queries.

[4] http://www.ncbi.nlm.nih.gov/pubmed

2.2.1 Health Related Datasets

To the best of our knowledge the first medical IR test collection was called the MED collection, consisting of 1033 MEDLINE abstracts and in 1994 as part of a shared task a larger collection was created by Hersh et al [Hersh et al. 1994]. It consisted of 348,566 MEDLINE references including of a set relevance judgements (qrel) and a set of topics or queries. The collection was created according to the conventional Cranfield settings [Cleverdon 1991] and it was called the OHSUMED [Hersh et al. 1994] after the Oregon Health Sciences University.

The OHSUMED queries were generated to address actual information needs for clinicians and the assessed documents were retrieved in two iterations, by relying on the MEDLINE search interface and the SMART retrieval system respectively [Hersh et al. 1994]. Retrieved documents were judged by a group of domain experts to the group performing the search.

Starting in 2003, Genomics track involved multiple tasks, namely: ad hoc retrieval, text categorization, summarization and passage retrieval. The aim of the track was to address the growth of biological data and to exploit real information need of biomedical research scientists in the genomic domain [Hersh and Bhupatiraju 2003, Hersh 2005, Hersh et al. 2006; 2007]. This track became one of the largest TREC tracks in terms of the number of participants which ran until 2007.

More recently, however, with the growth of publicly available Electronic Health Records (EHR), one potential new application is to use a collection of such records as a source for finding patients for a medical trial. For this task it is necessary to search over large numbers of EHRs to find patients matching certain criteria, such as suffering a given disease or belonging to a demographic group. However, because of the unique structure and vocabulary of the records, search over such content presents new research challenges.

In order to start exploring this problem, TREC[5] organised medical IR tracks in 2011 and 2012, where the goal was to identify patient records that fulfil the characteristics of given queries (e.g. "Patients with hearing loss"). The queries were built by targeting a list of research areas that the U.S. Institute of Medicine has considered priorities for comparative effectiveness research. Participation in these tracks was strong with 54 research groups submitting runs over the two years of the conference. This generated much interest in search over EHRs [Voorhees and Tong 2011, Voorhees and Hersh 2012].

Apart from the TREC challenge, more recently Conference and Labs of the Evaluation Forum (CLEF) organised a shared task, with the purpose of fostering ways to access health data by lay people, in order to understand their health problem [Goeuriot et al. 2013]. In the 2013 shared task, discharge summaries were used by the organisers to generate queries and the collection was built by a crawl of certified health web pages. Nine groups participated in this task and overall 48 runs were submitted. Although no run significantly improved over a PRF baseline, there were 4 runs (from the same team) which outperformed the baseline system in terms of precision at 10. The following year, however, the CLEF 2014 retrieval task [Goeuriot et al. 2014b] which provided a cleaner data collection than 2013 proved more successful, showing both improved baselines and results. The top three systems took advantage of a language modelling approach and query expansion. The best performance was obtained using the UMLS Metathesarus for concept-based retrieval, and mutual information to identify related terms for query expansion [Shen et al. 2014].

In 2014, TREC introduced another medical track, called the Clinical Decision Support (CDS) which attempts to fill the gap between the genomic literature and medical records, promoting research to facilitate access to biomedical literature for clinical experts, supporting the practice of evidence based medicine [Roberts et al.

[5] http://trec.nist.gov

2015a]. The queries represent medical case reports, and document collection is a subset of PubMed containing almost 700,000 articles. 26 different groups participated in this track, receiving over 100 submissions.

2.2.2 Evaluation

Basic Evaluation

In order to compare any two or more techniques we need to have a set of gold standard/ground truth, that we conventionally refer to as query relevance judgments (qrel) in IR. It is a process of deciding how relevant are the ranked results or how similar they are compared to the given query. The IR community has mostly relied on human intelligence to find relevant documents to a particular set of queries. A binary judgment is one which specifies whether a document is relevant or not, alternatively sometimes decimal numbers are used to specify the degree of relevance as follow:

- 3: Highly relevant: document thoroughly covers the topic
- 2: Relevant: document contains some information about the topic
- 1: Marginally relevant: document contains no information beyond the given topic
- : 0: Non-relevant

Having collected the manual assessment, there are a few popular measures to evaluate the effectiveness of each technique. Comparison between IR systems has been historically based on two different aspects of search results, precision and recall. Precision measures what fraction of the retrieved answers are correct. it reveals the preciseness of a technique. Recall or sensitivity on the other hand measures what

fraction of the correct answers is retrieved. Recall is the metric of comprehensiveness, measuring how many relevant documents have been returned out of all relevant documents. Mean Average Precision (MAP) which is the most popular evaluation metric in IR, is based on precision and recall, where the average of precision and recall are calculated at every position in the ranked retrieval and then divided by the number of queries.

However, one limitation of MAP is the underlying assumption that the relevance judgment is complete and that the unassessed/unjudged documents are non-relevant, which is not true for many IR test collections. In 2004 an alternative metric known as Binary Preference (Bpref) [Buckley and Voorhees 2004] was proposed whereby the differences between systems are measured based on the number of judged non-relevant documents prior to the relevant documents. Bpref was shown to be more reliable for test collections with a limited number of assessed documents. For this reason, Bpref was chosen to be the main evaluation metric for the first medical shared task in 2011 [Voorhees and Tong 2011].

The formula for Bpref is given in Equation 2.1, where R is the number of relevant documents for a query, r is a relevant document and n is a member of the first R judged non-relevant documents retrieved by an IR system.

$$Bpref = \frac{1}{R} \sum_{r} \frac{|n \; ranked \; higher \; than \; r|}{R} \tag{2.1}$$

There are other limitations to IR's evaluation, such as the degree of disagreements between the assessors. However, in 1998, in a series of experiments, Zobel [Zobel 1998] demonstrated that despite the ever existing disagreements between the assessors and the shortage of human assessments, the measured relative performance of systems are reliable [Zobel 1998]. For a detailed review of metrics and test collection based evaluation refer to Sanderson's survey [Sanderson 2010].

Automatic and semi-automatic Evaluation

Despite the sheer dependence on human assessment, the attempts to automatically build a complete set of qrel(e.g. [Soboroff et al. 2001]) and to extend a limited set of qrels(e.g. [Büttcher et al. 2007]) is not new.

To the best of our knowledge, Soboroff et al. [Soboroff et al. 2001] were the first to make an attempt at automatically creating relevance judgments by using patterns of occurrence of documents retrieved by multiple IR systems. However, this work suggested that some level of manual intervention is required when forming relevance judgments.

More recently Sakai [Sakai and Lin 2010] showed that by relying on documents retrieved frequently by a diverse set of IR systems, it is possible to build qrels automatically which produce high correlated ranking compared to manually judged data. However, this approach relies on a set of runs from different research groups, which is not always available.

While the previous work relied on ranking of many IR systems, others used category structures to substitute for relevance judgements(e.g. [Harmandas et al. 1997]). Koopman et al. [Koopman et al. 2011a] in the clinical domain exploited the descriptions of ICD codes to build queries. Finally, to generate the qrel, documents containing those codes were marked as relevant.

Not very different to creating a whole new set of qrel, others tried to expand a small set of real qrels by different means. Büttcher et al. [Büttcher et al. 2007] showed that by automatically expanding an initial set of document assessments, a more accurate evaluation of IR systems is possible. They used Machine Learning methods trained over a subset of relevance judgements to expand a set of relevance judgements. It was shown that evaluation results improved when a limited set of qrel was expanded. A number of IR systems were ranked by the expanded qrel and compared against the system ordering produced by the original qrel. In the clinical

domain, Martinez et al. [Martinez et al. 2008] also explored the use of re-ranking methods based on reduced judgements and found that the use of automatic classifiers considerably reduced the time required for clinicians to identify a large portion of relevant documents. Both works reported limitations of classifiers in cases where the initial number of documents is small.

2.3 The Current State of Patient Cohort Search (PCS)

In order to review the state of PCS in IR, the TREC Medical Records Track of 2011 and 2012 give the best picture of the state of the art at this point since several research groups participated in a shared task over the same patient repository. In this section, we present a summary of work on the only two Medical TREC shared tasks.

The best performing runs in 2011 [King et al. 2011] and 2012 [Zhu and Carterette 2012] focused on different aspects of search. King et al. in 2011 relied heavily on text processing and information extraction. Due to the shortage of relevance judgement at the start of the shared task, they tuned their system using their own manually created relevance judgment of approximately 190 reports per query. In 2012, Zhu and Carterette relied on evidence aggregation, external query expansion and Markov Random Fields. They employed 3 levels for merging the results of IR systems by evaluating visits, based on the best evidence from the reports, aggregation of reports to a visit, and finally the combination of both approaches. In 2012 the system that achieved the best results benefited from the availability of training data from 2011, and they performed optimisation of parameters over the early query set. Both groups gained improvement by using external knowledge sources for query expansion, however, many other configurations contributed to the performance of their final systems.

The retrieval tasks in both the 2011 and 2012 Medical TREC tracks highlighted that vocabulary mismatch is one of the key problems in the domain. A common way

to alleviate the problem is to use some external resources (e.g. SNOMED-CT) such as a biomedical knowledge base or a catalogue of terminologies (e.g. ICD codes).

Koopman et al.[Koopman et al. 2011b] converted all the terms in queries and documents of the medical TREC to SNOMED-CT concepts automatically, such that, a single SNOMED-CT concept could capture all the terms it associated to. This was an attempt to eliminate the need to introduce new terms to match more semantically related terms. Their results showed significant improvements using their concept-based method in comparison to a keyword baseline. Martinez et el. [Martinez et al. 2014] also used UMLS as a graph to find concepts for expansion, with promising results over the same TREC collections.

Goodwin et al. [Goodwin et al. 2011] focused on query analysis and reformulation to extract age and gender-related terms. To bridge the vocabulary gap between queries and documents they selected keywords from the given topics using Wikipedia. The selected terms were then expanded with Pubmed, UMLS and SNOMED-CT. While most groups including RMIT [Amini et al. 2011] performed sanity checks at the pre-retrieval stage, they take an extra precautionary step of filtering out retrieved documents that contradict the inclusion criteria, using age, gender and negation detection algorithms.

Schuemie et al. [Schuemie et al. 2011], after their analysis of the collection, found that discharge summary sections are not always reliable and instead they found the *postoperative diagnosis* section to be more informative. Their best run was based on balancing weights between the terms expanded from Wikipedia, using their Match Score Maximization algorithm [Schuemie et al. 2011] which was designed to make certain that top results contain most aspects of the queries. This was inspired by the conclusion made by the Reliable Information Access (RIA) Workshop [Harman and Buckley 2004] where a task of failure analysis per topic basis was explored. After exhaustively analysing 45 topics, 10 categories to describe the limi-

tations of the IR systems were drawn and Buckley [Buckley 2004] concluded that all categories share the same reason. The reason being the failure to retrieve documents that contain all aspects of queries.

The University of Glasgow [Limsopatham et al. 2011a] experimented with a novel voting model approach, and proposed a simple way to implement negation handling by adding a prefix to negated concepts in the reports, similar to two other teams from RMIT [Amini et al. 2011] and NICTA [Karimi et al. 2011]. In the same work [Limsopatham et al. 2011a], for query expansion, the weight of the expanded terms from Wikipedia and MeSH were calculated using EMIM (Expected Mutual Information Measure) [Rijsbergen 1979]. Their best run maps the ICD codes in the visits and expands the available concepts with Wikipedia. This was a unique way to expand documents with a knowledge source rather than queries.

Fushman et al. [Demner-Fushman et al. 2011] focused on identifying implicit fields in the reports, to automatically weight terms based on the section they appear in. They manually looked at a random sample of the reports and found that different types of records such as Radiology or Discharge Summary have some unique fields. However, finding an ideal set of fields, and tuning the weighting parameters to correctly capture the relative importance of the fields is yet an unsolved research problem in the medical IR.

We also participated in both editions of the challenge. In TREC-M1 we mainly focused on external knowledge sources, such as the UMLS, and DBpedia for query expansion [Amini et al. 2011]. In 2012 we took a different approach by locally expanding the queries with the collection (using pseudo-relevance feedback based on ICD codes), and by detecting and modifying the negated text in the reports [Amini et al. 2012b]. Our 2012 submission is the highlight of this thesis where we extend our systems by exploring the use of ICD for query expansion and automatic/pseudo relevance judgement. We also present diverse ways of mapping queries into ICD codes

and evaluate its performance systematically over the 2011 and 2012 medical TREC collections.

In addition to the above TREC submissions, we identified other distinct work that has been submitted to TREC but requires further investigation to consolidate reported findings.

Diaz et al. [Diaz et al. 2012] focused on the effect of negation handling on retrieval of clinical records. They defined a comprehensive syntactic information which was different to regular expression based algorithms such as NegEx. However, when compared to NegEx, no significant difference was reported.

Their finding which is in line with ours [Amini et al. 2011], states that negation handling does not significantly improve the performance, because only a limited number of reports are affected by negation and the negated terms do not occur in the queries.

Pastor et al. [Pastor et al. 2012] implemented a bag of concepts approach in which all terms are converted to the UMLS concepts. Their goal was to reduce the semantic gap between queries and reports, as the UMLS concepts encapsulate most of the term synonyms in a single concept. For retrieval they exploited language modelling with a two-stage smoothing method. First, Dirichlet prior is used to smooth the document language model and then the language model is smoothed by query background model at the second stage.

They implemented a unique strategy to handle demographic related information in which the UMLS was used to map age-related terms into relevant concepts. They created two distinct fields: age and gender in the documents and gave them higher weights in Indri's query language. For most of the queries, they outperformed the TREC median scores, reporting at least 10% improvement on every measure.

Dinh and Tamine [Dinh and Tamine 2011] experimented with different ranking models. They found that manually removing redundant terms from queries, and

query expansion using the DFR term weighting model [Amati and Van Rijsbergen 2002], produce their best results.

Qi and Laquerre [Qi and Laquerre 2013] produced a competitive run which was a combination of medical concept detection using Metamap, vector space retrieval and query expansion with PRF. Their preprocessing step included eliminating boilerplate text and removing punctuations caused by the de-identification process. Their best run achieved the highest infAp score for 5 topics and ranked the second best of all the 82 automatic runs submitted to TREC.

Daoud et al. [Daoud et al. 2011] relied on a medical annotation tool called the BioLabler[6] , and a medical synonym dictionary, known as the Polysearch [Cheng et al. 2008], for semantic matching. However, they reported poor results with BioLabler due to concept extraction accuracy errors, and the limitation of keyword matches between query concepts and the documents. For future work, they intend to improve on word sense disambiguation to improve their medical concept indexing.

Cogley et al. [Cogley et al. 2011] highlighted the effect of re-ranking and post-processing of documents. This task involved finding the number of rare concepts of an expanded query in a document and boosting the rank of the documents which contained those unique terms. However, query expansion was done with manual intervention through submitting queries to the Pubmed interface and performing a manual semantic lookup for synonyms, which were then added to the queries. They chose Indri search engine as it allowed structured query language which permits term weighting and phrasal searching. Their concept re-ranking run performed slightly above the median score of the TREC.

[6]http://www.biolabeler.com/bioLabeler

2.4 Summary

The related past work demonstrates the progress of IR research on health search but the ongoing work on patient records is recent and is of particular interest to this research.

The overview of recent work in this area illustrate that significant improvement in the effectiveness of search systems are dependent on smart query expansions that make appropriate use of both internal and external medical resources. However, despite the commonality of resources used by researchers, the outcomes were indeed different. This is partly due to the various engineering effort used (e.g. parameter optimisation) in conjunction with the usage of different ranking algorithms.

As a successful alternative to query expansion, the bag of concepts approach was also experimented [Koopman et al. 2011b, Pastor et al. 2012]. The consistent improvement reported by the use of query expansion or the bag of concepts approach, clearly demonstrates the importance of bridging the semantic gap in clinical IR.

Query expansion without selectively expanding specific terms, however, can lead to introducing noise to queries, deteriorating search outcomes. The Metamap was extensively used for this purpose to aid in the identification of non-informative terms [Karimi et al. 2011, Amini et al. 2011].

Despite the wide usage of external sources and biomedical tools, it is evident, however, that previous work neglected the potential value of the internal source of knowledge: ICD codes. Introducing external terms through query expansion while advantageous runs the risk of gathering information that does not exist in the collection, regardless of its relevance.

The other important but scarce area of research is the development of pseudo relevance judgement. Initial effort on this line of research in the clinical domain [Koopman et al. 2011a] does not harness real world queries formulated by medical professionals, and has not been evaluated against a manual relevance judgement.

We learn from past work and experiment with different expansion methods in Chapter 4. The importance of a medical coding system is highlighted in Chapter 5 and Chapter 6. We discuss our findings in the subsequent chapters and finally present the conclusion in Chapter 7.

CHAPTER 3

Datasets

3.1 Medical TREC Test Collections

In this short chapter, we present the two medical test collections that we used to perform our experiments. The collections were originally used as part of two medical IR shared tasks [Voorhees and Tong 2011, Voorhees and Hersh 2012].

The two collections shared the same set of documents (we refer to as reports) and only the queries were different.The reports are medical records collected in 2007 during the course of one month, from multiple hospitals in the U.S. Throughout the chapters in this thesis, we refer to the 2011 and 2012 Medical TREC test collections as *TREC-M1* and *TREC-M2* respectively.

The dataset consists of 93,552 clinical *reports* of patients visiting departments within hospitals. A patient could *visit* multiple departments during his/her time at the hospital. The TREC organisers provided a one-to-many mapping table from multiple reports to single visits. There are 17,265 visits in the collection. Nearly one-fifth of

the visits consist of a single report; the rest have multiple reports ranging from two to one hundred.

The unit of retrieval was chosen to be *visit* by the TREC organisers. However, a single visit may consist of reports that do not necessarily relate to the same medical condition. The participants had the choice to either index the collection based on visits by aggregating all the reports belonging a to a single visit, or indexing at the report level. In the case of the latter, they had to map each report id to its relevant visit id for retrieval. The indexing for our experiments was done at the report level.

Each report contains four informative XML tags: Two tags are reserved for the assignment of the ICD codes, namely, *Admit Diagnosis* and *Discharge Diagnosis*; the third tag is a short text (truncated to 40 characters) naming the chief complaint, and the main body of the text is given in a separate tag. Although the main text of the reports is not systematically structured, it has headings, which represent the start of a new section. We refer to these sections as fields and the following are some instances: *Family History*, *Present Illness*, *Allergies*. Note that the ICD codes in the TREC reports which are used throughout the thesis refer to the ICD-9-CM which is a US adaptation of ICD.

We analysed the collection to gain familiarity with the structure of the documents. Using simple pattern matching we extracted section headings and identified segments pertaining to different population and age groups. We found that 12,006 reports had one visit associated while 2,387 of the reports had more than or equal to 10 visits.

In the first year of the medical track, there were 34 queries , and 47 in the second year. The numbers were originally 35, and 50, however, four queries were removed due to the insufficient number of relevant visits (i.e. documents) in the collection for those queries.

The queries were built by targeting a list of research areas that the U.S. Institute

of Medicine has considered priorities for comparative effectiveness research[1]. The relevance judgment was done by groups of clinicians after pooling documents for each query. The queries included different pathologies and treatments, as well as demographic constraints.

As mentioned in Chapter 2 Bpref was chosen as the main evaluation metric for the TREC-M1 due to the limited number of assessed documents and therefore we report the results of all our retrieval experiments in terms of Bpref.

[1] http://www.iom.edu/Reports/2009/ComparativeEffectivenessResearchPriorities.aspx

101 Patients with hearing loss
102 Patients with complicated GERD who receive endoscopy
103 Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis
104 Patients diagnosed with localized prostate
cancer and treated with robotic surgery 105 Patients with dementia
106 Patients who had PET,MRI, or computed tomography (CT) for staging or monitoring of cancer
107 Patients with ductal carcinoma in situ (DCIS)
108 Patients treated for vascular claudication surgically
109 Women with osteopenia
110 Patients being discharged from the hospital on hemodialysis
111 Patients with chronic back pain who receive an intraspinal pain-medicine pump
112 Female patients with breast cancer with mastectomies during admission
113 Adult patients who received colonoscopies during admission which revealed adenocarcinoma
114 Adult patients discharged home with palliative care / home hospice
115 Adult patients who are admitted with an asthma exacerbation
116 Patients who received methotrexate for cancer treatment while in the hospital
117 Patients with Post-traumatic Stress Disorder
118 Adults who received a coronary stent during an admission
119 Adult patients who presented to the emergency room with with anion gap acidosis secondary to insulin dependent diabetes
120 Patients admitted for treatment of CHF exacerbation
121 Patients with CAD who presented to the Emergency Department with Acute Coronary Syndrome and were given Plavix
122 Patients who received total parenteral nutrition while in the hospital
123 Diabetic patients who received diabetic education in the hospital
124 Patients who present to the hospital with episodes of acute loss of vision secondary to glaucoma
125 Patients co-infected with Hepatitis C and HIV
126 Patients admitted with a diagnosis of multiple sclerosis
127 Patients admitted with morbid obesity and secondary diseases of diabetes and or hypertension
128 Patients admitted for hip or knee surgery who were treated with anti-coagulant medications post-op
129 Patients admitted with chest pain and assessed with CT angiography
130 Children admitted with cerebral palsy who received physical therapy
131 Patients who underwent minimally invasive abdominal surgery
132 Patients admitted for surgery of the cervical spine for fusion or discectomy
133 Patients admitted for care who take herbal products for osteoarthritis
134 Patients admitted with chronic seizure disorder to control seizure activity
135 Cancer patients with liver metastasis treated in the hospital who underwent a procedure

Table : List of all the TREC-M1 queries. query 130 has been later excluded by the TREC organisers for the lack of relevant documents

136 Children with dental caries
137 Patients with inflammatory disorders receiving TNF-inhibitor treatments
138 Patients with acute tubular necrosis due to aminoglycoside antibiotics
139 Patients who presented to the emergency room with an actual or suspected miscarriage
140 Patients who developed disseminated intravascular coagulation in the hospital
141 Adult inpatients with Alzheimer's disease admitted from nursing homes with pressure ulcers
142 Patients admitted with Hepatitis C and IV drug use
143 Patients who have had a carotid endarterectomy
144 Patients with diabetes mellitus who also have thrombocytosis
145 Patients with lupus nephritis and thrombotic thrombocytopenic purpura
146 Patients treated for post-partum problems including depression, hypercoagulability or cardiomyopathy
147 Patients with left lower quadrant abdominal pain
148 Patients acutely treated for migraine in the emergency department
149 Patients with delirium, hypertension, and tachycardia
150 Patients who have cerebral palsy and depression
151 Patients with liver disease taking SSRI antidepressants
152 Patients with Diabetes exhibiting good Hemoglobin A1c Control 8.0
153 Patients admitted to the hospital with end-stage chronic disease who are offered hospice care
154 Patients with Primary Open Angle Glaucoma (POAG)
155 Heart Failure (HF): Beta-Blocker Therapy for Left Ventricular Systolic Dysfunction (LVSD)
156 Patients with depression on anti-depressant medication
157 Patients admitted to hospital with symptomatic cervical spine lesions
158 Patients with esophageal cancer who develop pericardial effusion
159 Patients with cerebral edema secondary to infection
160 Patients with Low Back Pain who had Imaging Studies
161 Patients with adult respiratory distress syndrome
162 Patients with hypertension on anti-hypertensive medication
163 Patients treated for lower extremity chronic wound
164 Adults under age 60 undergoing alcohol withdrawal
165 Patients who have gluten intolerance or celiac disease
166 Patients who have hypoaldosteronism and hypokalemia
167 Patients with AIDS who develop pancytopenia
168 Patients with Coronary Artery Disease with Prior Myocardial Infarction on Beta-Blocker Therapy
169 Elderly patients with subdural hematoma
170 Adult patients who presented to the emergency room with suicide attempts by drug overdose
171 Patients with thyrotoxicosis treated with beta-blockers
172 Patients with peripheral neuropathy and edema
173 Patients over 65 who had Pneumonia Vaccination Status presently or previously
174 Elderly patients with ventilator-associated pneumonia
175 Elderly patients with endocarditis
176 Patients with Heart Failure on ACE Inhibitor or ARB Therapy for Left Ventricular for LVSD
177 Patients treated for depression after myocardial infarction
178 Patients with metastatic breast cancer
179 Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression
180 Patients with cancer who developed hypercalcemia
181 Patients being evaluated for secondary hypertension
182 Patients with Ischemic Vascular Disease
183 Patients presenting to the emergency room with acute vision loss
184 Patients with Colon Cancer who had Chemotherapy
185 Patients who develop thrombocytopenia in pregnancy

Table : List of all the TREC-M2 queries, queries 138, 159, and 166 have been later excluded by the TREC organisers, and query 176 was slightly truncated to fit in the table

CHAPTER 4

Building a Search System for Clinical Records

Exploiting the characteristics of clinical records in order to improve search result, is the basis for customising an effective search system to handle clinical queries. In this chapter, we build and combine a number of techniques that have proved essential for the retrieval of clinical records in past research. These methods make a default setting and form a strong baseline system that we use in the next chapter to compare our novel methodology with.

Clinical records are loosely structured but they contain certain sections that are identifiable by their heading. They may relate to patient's condition, background and his/her current and past treatments, such as, *patient history*, *family history*,*discharge diagnosis*, etc. From the retrieval point of view, these sections may vary in terms of the degree of relevancy to queries. For instance, if the term *migraine* occurs under the *family history*, it may not be entirely relevant to a query that targets patients who

currently suffer from *migraine*. Therefore, learning the differences between these sections may hold the key to a more effective retrieval.

The style of writing is also a unique characteristic of clinical records that we take into account. Practitioners often report the absence of a disease or symptoms in medical records to reject patient's initial complaint. This implies that the condition is not relevant to the record. While this is clear to a human reader, it can potentially mislead any ranking algorithm. Therefore, identifying the polarity of a term is crucial in retrieving a medical record.

Similar to the problem of negation, and often at the initial stages of a patient's visit, doctors may also wish to describe uncertainty, to cast doubt on the existence of a particular condition/disease. Unless the query is targeting patients imminent to the given condition/disease, the uncertain condition does not imply relevance.

Another important factor that contributes in shaping the unique characteristic of clinical records is the wide usage of synonyms. There are many ways to describe a clinical concept, and the word selection by one practitioner may vary from others. Also, the existence of abbreviations additionally pose a new challenge on the accurate retrieval of clinical records. Not knowing these variations can result in missing relevant records.

Understanding the structure and the language of clinical records are indeed vital, however, retrieval is also about the users and how they pose queries to interact with retrieval systems. Given the potential gap between the language of queries and clinical records, the traditional keyword matching or the bag of words approach although often effective, may not be sufficient to overcome this problem.

If terms have the same stem or root, the trivial approaches such as stemming or lemmatization that transform terms into a universal stem or root may suffice. Stemming is mostly a default setting of all search engines. However, in the case of complex synonyms, a simple word reconstruction of queries is not enough. In particular,

when queries in clinical IR are generated by laypeople, given their lack of expertise, the need to transform queries is more pronounced.

A classical approach to bridge the gap between user queries and documents is known as query expansion. Query expansion potentially increases the chance of retrieving more relevant documents by adding extra terms to the query. It assumes that the initial query does not completely express user's need, and that the collection has relevant documents that may not necessarily have a keyword overlap with the query.

Broadly speaking, the existing query expansion methods can be categorised as either local or global. They differ in the sources where extra information are selected from. Local expansion methods exploit terms only from the underlying collection, while global methods use information from external knowledge sources.

Based on their popularity and past usage we choose three external knowledge sources for our global query expansion. We compare Wikipedia, Dbpedia and hyponyms from the UMLS. We also measure the degree of overlap between the expanded and original query, and the relevant documents, for every query.

We focus on techniques that have a potential impact on the sound retrieval of clinical records. Query expansion, to overcome the abundance of terminologies and variations. Negation handling, to avoid misleading search systems while preserving important information. We also learn about the structure of clinical records by text processing and extracting section headings from the collection.

4.1 Background

Although Natural Language Processing (NLP) proved largely unsuccessful in information retrieval [Voorhees 1999], medical IR has been shown to benefit from implementing NLP approaches such as negation detection [King et al. 2011, Koopman et al.

2010], to better distinguish relevant from non-relevant documents.

If a query targets a population without a particular disease D or treatment T for instance, search engines have no way of penalising documents for containing the word D or T by default. For example, for the following query: *Patients without dementia*, search engines will look for any occurrence of dementia in the documents by default. This may eventually lead to the unwanted retrieval of patient records *with dementia*.

Identifying negated terms from clinical text is commonly done by NegEx [Chapman et al. 2001a] which is a simple regular expression algorithm. NegEx is reported to be reliable, obtaining high scores on sensitivity and precision when finding negative phrases in clinical text [Chapman et al. 2001b].

Participants of the Medical TREC conference performed different experiments with Negation. University of Glasgow [Limsopatham et al. 2011b] introduced a new tokeniser (called NegExTokeniser) and Cengage Learning [King et al. 2011] used their own algorithm to find uncertain conditions from the text. Both groups take advantage of negation handling and report marginal improvements.

NLP can also aid in identifying and exploiting the existence of explicit and implicit structure in clinical records to distinguish different sections in clinical records. Such a structure can be used in a fielded search, by ranking terms with regards to their respected sections. However, this can only be useful if there exist informative sections/fields whose importance is different in comparison to other sections/fields.

Fushman et al. [Demner-Fushman et al. 2011] focused on identifying implicit fields in the reports to automatically weight terms based on the context they appear in. They manually looked at a random sample of reports and found that different types of records such as Radiology and Discharge Summary have some unique fields. However, a conclusion that could be drawn from this work is that finding an ideal set of fields, and tuning the weighting parameters remains an unsolved research problem

in the clinical IR.

In addition to the reports, queries posed by users need further processing that may involve incorporating external knowledge sources to bridge the gap between queries and reports. Choosing the right knowledge source for expansion is indeed a critical task for global query expansion. If the source for query expansion is external and the language between queries and the knowledge source do not match sufficiently, expanding queries is most likely to deteriorate the search outcome than to improve it.

A resource widely used on its own is SNOMED-CT, which is a subset of UMLS, comprising a collection of medical terms covering diseases, findings, procedures, micro-organisms, substances, etc. Koopman et al.[Koopman et al. 2011b] converted all the terms in queries and documents to SNOMED-CT concepts automatically, such that, a single SNOMED-CT concept could capture all the terms it associated to. This was an attempt to eliminate the need to introduce new terms or to perform any kind of relevance feedback to match more semantically related terms. Their results showed significant improvements using their concept-based method, in comparison to a keyword-based baseline.

However, mapping every term to an SNOMED-CT concept may incur the cost of losing specific information that does not have a mapping to any SNOMED-CT concepts. Such information can be related to the age of patients. For instance, in the query: *"Children with dental caries"*, an IR system must have a way of filtering results to include only those related to children. However, practitioners tend to use different language to refer to the population in the reports and they may choose to write the age in number form e.g., *(18 years old)*, or written as in: *Adult,Young*. Handling population criteria, however, requires document pre-processing and many of the teams in the first TREC shared task [Voorhees and Tong 2011] ignored this setting. Nonetheless, two methods that have been tested before, include reformulating

query terms before the retrieval stage [Karimi et al. 2011], and filtering documents at the post retrieval stage [Goodwin et al. 2011].

Apart from the difficulty in finding a reliable source for expansion, is the problem of selecting the most informative terms from the sources. While pseudo relevance feedback, as a local expansion method, exploits the assigned weights to the terms in the document and effectively select only the top n terms, words extracted externally are not weighted, which requires additional processing to make an elite selection.

Researchers have employed different ways to incorporate terms from external knowledge sources. Goodwin et al. [Goodwin et al. 2011] manually assigned weights to terms from different Knowledge Sources (KSs) for expansion, whereas Zhu and Carterette [Zhu and Carterette 2011] used an algorithm known as CORI [Callan 2002] to automatically assign weights to external KSs by means of assigning similarity weights based on term overlaps. In their work three sources that were not used by other teams, to the best of our knowledge, were tested. According to their results, the following sources: TREC Genomics track and one day PubMed query log were most and least effective respectively, while imageCLEF had a moderate improvement over their baseline system. Note that for image CLEF, they used the captions and crawled the text related to all images from their corresponding URL.

4.2 Experimental Settings

In this section first, we present the methodologies for our analysis on structural fields in medical reports and handling negation. We lastly describe the steps to perform query expansion using external knowledge sources. We use the TREC-M1 test collection explained in Chapter 3 for all the experiments.

4.2.1 Structural Fields in Medical Reports

Query terms appearing in non-relevant documents can potentially fool ranking algorithms into retrieving false positives, imposing a negative effect on the overall score of search engines. Therefore we intend to address the question, whether contextual or implicit fields in the clinical records can be used to discard non-relevant documents that embody one or some aspects of queries.

We identified 2 explicit fields to be common across the collection: *family history* and *past history*, which can be searched by regular expressions as they are typed in capital letters followed by a carriage return. However, we also hypothesise that other fields exist in the collection that are not explicitly identifiable. In order to verify this and examine if query terms are indeed classifiable based on the context they appear in, we manually looked at a sample of relevant and non-relevant reports.

We observed that some documents have phrases that are misplaced and do not occur in their expected fields. For instance, a phrase referring to the past history of a patient with a particular disease is expected to occur under the *past history*, however, there are many cases where these instances occur elsewhere, and this is often observable by reading the complete sentence. This suggests that in order to get the most accurate count of query terms with regards to their context, we must perform this task manually. However, in section 4.4 we propose a simplified way for counting the query terms which allows us to automate this task, and discover useful implications.

Based on a manual look-up of query terms with respect to their contexts, we defined 4 fields in addition to the previously identified explicit fields. We refer to them as implicit fields: *Speculation, Secondary Problem, Etc, Text Body*.

Speculation is where the term is used with uncertainty, *Secondary Problem* means that the clinician has mentioned a particular disease or symptom to be secondary to the patient's primary problem, *Etc* implies that this word was just men-

tioned in the report with no regards to patient's own or family history and this was a very rare case, *Text body* is the main part of a report that does not belong to any other sections.

The number of relevant reports for each topic is not substantial and it was possible to rely on a limited human resource, however, the number of non-relevant reports are much too high for our limited human resources to manually count the occurrence of all the query terms. Therefore we randomly sampled 20% of the relevant and non-relevant reports in order to manually count the number of terms with regards to their context.

It was essential to convert the queries into keywords that are informative and distinguishable. We used Metamap to process all the 35 queries retaining all the terms that have a medical mapping to a concept while discarding the rest. However, we noticed that 3 words remained that fell under a generic category and were present in most of the reports, which we stopped. Those terms were: "Admission" , "Received" and "Treated". Finally, we only retain the terms that have at least one occurrence in the non-relevant documents, as otherwise, a large portion of non-relevant documents would have zero counts for all the fields.

The counting was performed on the collection level. We created two large files, one with all the relevant reports, and the other included all the non-relevant reports. The assessor could keep on pressing 'next' on a Unix-based terminal once for every query term. Counting at the document level, however, would have required us to stop and write the counts for every single document. This would have taken a considerable amount of time with our limited resources and the given deadline to the TREC participants to discard the collection for previously discussed confidentiality problems.

4.2.2 Negation

A simple approach for handling negation will discard the identified negated terms from queries and a more advanced method will alter the negated terms for instance by adding a prefix in the queries as well as documents prior to indexing.

We use NegEx from the MetaMap program to identify negated terms from the records and we settle for using the pre-indexing approach and use a rule-based algorithm to substitute the negated terms with a new prefix. Our implementation of Negation has an assumption that if a term is negated once in the documents, it will be negated for the scope of the entire document. Such that, all occurrences of a negated terms will be converted to the negated form. For example, if NegEx detects that the term *dementia*, has been negated once, and once it has not in the same document, we replace both terms into a negated form, assuming that the second mention of *dementia* was also a negation but failed to be detected by NegEx. We refer to this technique as Neg-Aggressive.

An example of a query with a negated term is as follows:
Query: patients with a BMI $>$ 40 without hypertension. Our approach prefixes all occurrences of hypertension with a prefix "no". So the term *without hypertension* would be replaced by: "nohypertension".

4.2.3 Query Expansion

The first step for the external query expansion was choosing a reliable knowledge source. We explored two publicly accessible sources: Wikipedia and Dbpedia. Note that Dbpedia allows users to access the content of the Wikipedia in a form of structured information, however, as we experimented with both separately, we refer to them as two sources. Moreover, we explored expansion using hierarchical relations from the UMLS Metathesaurus by selecting all terms in the hyponym concepts. A

hyponym of a word has a more specific meaning, and we found that some medical terms in the queries can have up to 4000 hyponyms.

To enhance the quality of expansion, later we used *MetaMap-2010* to identify phrases linked to terms in the UMLS Metathesaurus (version 2010AA). We only kept terms that had a mapping to a medical concept. For example, for the query below we kept the following phrases: "liver metastasis" "treated hospital" "cancer patient" "procedure".

Cancer patients with liver metastasis treated in the hospital who underwent a procedure

To this end, we have a number of phrases and terms ready to be expanded. We create three different sets of expansion for each of the knowledge sources by using a perl script. The algorithm is straightforward: first, we look up for a matching web page for the online sources: Wikipedia and DBpedia. Next, we crawl all the text, filtering out the unwanted terms.

For the UMLS expansion, however, we directly queried the relevant UMLS table to fetch all the hyponyms related to each term. In the case of DBpedia[1] we extracted all terms listed under the "category" section. After examination, we found that strings with the following terms indicated noise and we, therefore, removed them: "code", "history", "mechanism", "poisoning", "toxicity", and "withdrawal".

Exploiting Semantic Types from the MetaMap

In order to further refine our expansion, we perform a second experiment with STs from the MetaMap. For every ST, we keep the terms associated to it to form a query. Those queries are then run against the collection. Based on their scores, we chose the

[1] http://wiki.dbpedia.org/OnlineAccess

top 4 STs out of all the STs in the queries by comparing the individual performances of each ST. A list of the 21 semantic types are presented in table 4.1 with their scores. In fact, these STs were chosen as they obtained the highest 4 Mean Average Precision (MAP) on the TREC relevance judgement. The scores ranged from (0.29 to 0.34), and due to the absence of any statistical difference, we only chose the highest scoring STs, as using all the STs would obviously not solve the over-expansion problem. This experiment was done in preparation for the first TREC in 2011, where we had TREC's initial relevance judgement before the official metric was chosen to be Bpref.

STs are assigned to medical concepts in the UMLS relational tables and they are extractable from the output generated by the Metamap. Figure 4.1 is an output generated by Metamap and the terms appearing in brackets refer to semantic types associated to each concept. Identified concepts will be expanded by external knowledge sources to form a verbose query.

Phrase: "with hearing loss"
Meta Candidates (6):
1000 C1384666:Hearing Loss (hearing impairment) [Finding]
1000 C2029884:hearing loss (hearing loss by exam) [Finding]
861 C0018767:Hearing [Physiologic Function]
861 C1455844:Hearing (Hearing examination finding) [Finding]
861 C1517945:Loss [Quantitative Concept]
861 C2015933:hearing (outcomes otolaryngology hearing) [Finding]
Meta Mapping (1000):
1000 C1384666:Hearing Loss (hearing impairment) [Finding]
Meta Mapping (1000):
1000 C2029884:hearing loss (hearing loss by exam) [Finding]

Figure 4.1: Output generated by metamap

Nicta [Karimi et al. 2011] was one of the TREC participants who took an in-

Number	Semantic Type	Score
1	Body Substance	0.340
2	Organic Chemical	0.328
3	Health Care Related	0.329
4	Educational Activity	0.340
5	Professional or Occupational	0.340
6	Age Group	0.309
7	Functional Concept	0.341
8	Family Group	0.340
9	Pharmacologic Substance	0.340
10	Clinical Attribute	0.340
11	Neoplastic Process	0.339
12	Health Care Activity	0.324
13	Manufactured Object	0.305
14	Finding	0.342
15	Pathologic Function	0.319
16	Spatial Concept	0.339
17	Patient or Disabled Group	0.323
18	Medical Device	0.343
19	Amino Acid, Peptide	0.340
20	Disease or Syndrome	0.347
21	Mental Dysfunction	0.295

Table 4.1: 21 semantic types from the output of Metamap and their MAP scores

tuitive approach by choosing two semantic types and discarding the remaining expanded terms with other types. However, we filtered out terms before expansion, meaning that we only expanded terms that had a particular semantic type, on the contrary, Nicta chose acceptable terms from the new terms, after expansion.

To this end, we tested all the knowledge sources separately, however, our last experiment involved combining all the knowledge sources which then became our

final run for the TREC-M1 competition. The pipeline for combining the expansion types are shown in Figure 4.2. It demonstrates this experiment in 7 steps, from filtering unwanted terms by through the Metamap to combining all the expanded terms and finally the retrieval of the relevant documents.

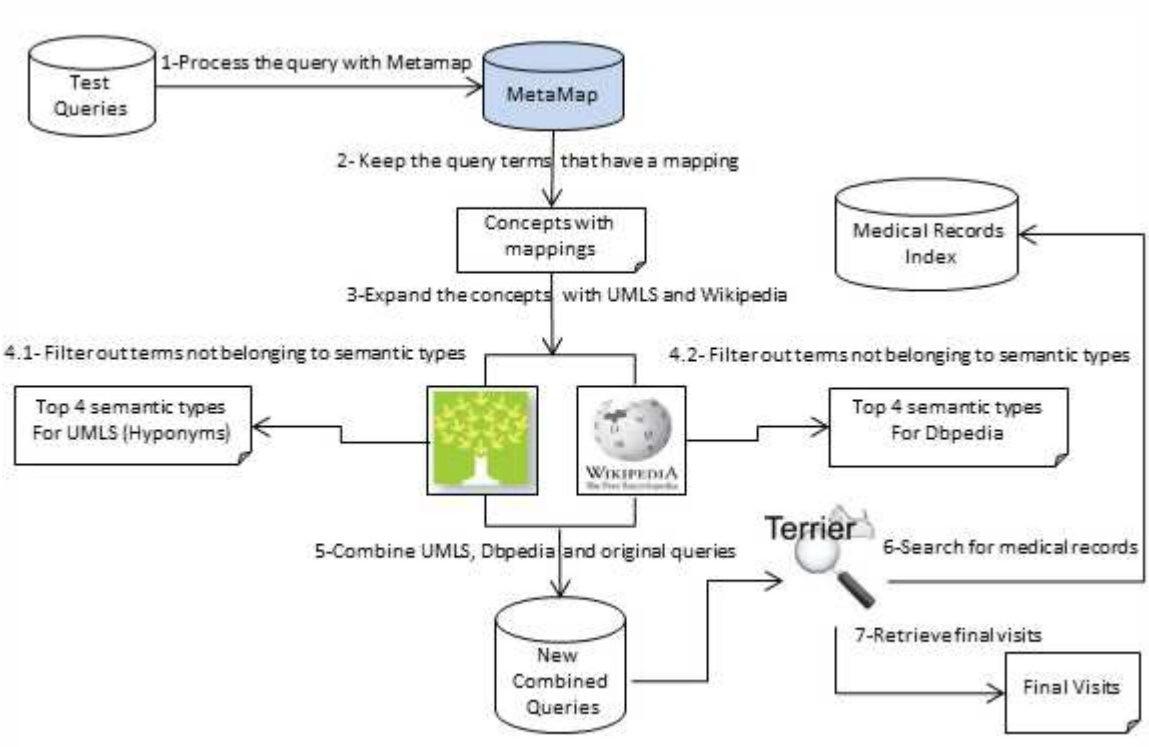


Figure 4.2: Combining knowledge sources

4.3 Results

Here we present the results of experiments on query expansion. Note that for all the following experiments we handle negation by default. We used the Terrier [Macdonald et al. 2012] search engine using the PL2 ranking algorithm keeping the default values for all the parameters [Amati and Van Rijsbergen 2002].

We rely on the Porter stemmer to strip all suffixes of the English words in queries and documents to allow for matching of the terms with the same root [Porter 1980a].

Our first attempt shows that using any knowledge source deteriorates the original query. Looking at the performances given in the last two columns, we can see that all the knowledge sources, when used on their own right or combined with the original query terms, exacerbate the results. This implies the presence of noise introduced from all the sources which lower the scores.

Knowledge Source (KS)	Without original query	With original query
Original Query Terms	N/A	0.3143
Dbpeida	0.2831	0.3048
Wikipedia	0.0903	0.2866
UMLS (Hyponyms)	0.2044	0.2874

Table 4.2: Mean Bprefs using different knowledge sources

As discussed before, in order to reduce the amount of noise from the expanded queries, we test retrieval with all the STs. That is, for every ST, we only keep the terms it is associated to.

Table 4.3 shows the top 4 semantic types for UMLS and Dbpedia that provide the best results after being added to the plain queries. All the scores are given in Bpref and the highest results are differentiated in bold. For instance, *Disease or*

Semantic Types (STs)	Dbpedia	UMLS(hyponym)
Disease or Syndrome	0.3371	**0.3145**
Finding	0.304	0.3111
Functional Concept	0.3061	0.3077
Medical Device	**0.3374**	-
Baseline(No Expansion)	0.3143	0.3143

Table 4.3: Bpref scores using different semantic selection to expand queries, - indicates that the Knowledge Sources did not have any term related to the unexpanded term associated to the given ST

Syndrome ST is the best for UMLS, while it is the second best for terms extracted from the Dbpedia. We only show the top four scoring STs in this table.

The low performances obtained by some of the STs, those for UMLS in particular, were tracked down to the large size of the expansion for a fraction of the queries which negatively affected the outcome. This implies that a more precise approach is needed for the refinement of such cases.

As shown in figure 4.2, our final run is a combination of all refined knowledge sources, which was submitted to TREC-M1. This run obtained 0.3457 Bpref score, and was the only run to statistically outperform other runs.

4.4 Discussion

The experiments that we presented in this chapter were performed as part of the TREC challenge, using the test collection that had to be used for a limited period of time for the duration of the task. They, however, give us a picture of the important features of this kind of data.

We discuss three aspects of the TREC-M1 test collection. First, we look at the percentage of distribution of query terms in the relevant documents. Next, we look

at the distribution of negated terms across the relevant and non-relevant documents. Finally, we focus on the 6 explicit and implicit fields across the collection and draw a conclusion on the distribution of the query terms with respect to the fields they appear in.

4.4.1 Query Expansion

Observing the effects of the knowledge sources in query expansion, it is clear that the cost of introducing noise outweighs the benefits of expanding the original queries with domain specific knowledge. The sheer diversity of lexical variants lead to the problem of over-expansion. We found that some medical terms in the queries can have up to 4000 hyponyms.

The retrieval performance suffers more in the case of Wikipedia. DBpedia, on the other hand, offers a more structured way to access the content of Wikipedia, which allowed us to better discard the non-informative terms.

In order to shed light on the reason for the limitation of our knowledge sources, we measure the degree of overlap between the language of the query, before and after expansion, and the collection. Our aim was to determine firstly, which of the several knowledge sources have the highest degree of overlap with the documents and secondly, how much does the overlap reflect the performance of retrieval systems?

The original terms used for this experiment are a list of keywords that we gathered from Metamap which is made available as an appendix to this chapter.

We show in Table 4.4 the percentage of the coverage of query terms across the relevant documents.

The numbers indicate that the vocabulary that has generated the queries is more likely to be the same as the documents compared to other knowledge sources. This is followed by the hyponyms extracted from the UMLS.

In the case of the original queries, the percentage of the coverage reflect the performance. We can see that the best performance is obtained by just using the plain queries. Nonetheless, DBpedia and UMLS hyponyms despite the similar coverage have different performances. DBpedia leads to a better performance, while Wikipedia has the worst impact on the original queries. This indicates that the Wikipedia and the UMLS are more likely to introduce noise to the plain queries.

Knowledge Source (KS)	Count(no-stem)	Count(stemmed)	Bpref
Original Query Terms (without KS)	56.83%	59.52%	0.3143
Dbpeida	42.61%	55.95%	0.2831
Wikipedia	23.11%	23.48%	0.0903
UMLS (Hyponyms)	55.51%	56.56%	0.2044

Table 4.4: Percentage and mean Bpref of exact matches using query terms and extended terms from different knowledge sources

4.4.2 Negated Terms

In the previous section, we presented the percentage of the query terms which have been used in a positive form across the relevant reports. The presence of almost 60% of the positive terms is exploited by the search engines to identify relevant reports. However, there are many factors that lead to the retrieval of false positive reports, as well. One of which is the query terms that are mentioned in negative forms. Looking at the distribution of the negated terms will give us a different information as to whether the handling of negated terms is necessary for the retrieval of clinical records.

Here we show the statistics collected by counting the number of negated occurrences of the query terms in the relevant and non-relevant sets.

To count the matches between the queries and the reports we used the same terms that were used for the analysis of query expansion 4.4.1. Therefore, the count for each query is the summation of the counts of all the query terms for that particular query.

The numbers in the table 4.5 show that the query terms have not been used in a negative sense across the relevant documents.

This implies that handling negation can be mainly effective for filtering out non-relevant documents. It also indicates that it is safe to discard a document that contains a negative form of a query term, as relevant documents tend to be devoid of negated query terms.

The 35 queries are shown with their corresponding mean. Relevant documents are those found from the relevance judgement and all other documents in the collection are considered to be non-relevant.

Query Number	Relevant	Non-relevant
101	0	26
102	0	18
103	0	141
104	0	9
105	0	61
106	0	251
107	0	2
108	0	573
109	0	21
110	0	0
111	0	4
112	0	15
113	0	7
114	0	0
115	0	102
116	0	0
117	0	1
118	0	0
119	0	1
120	0	105
121	0	495
122	0	0
123	0	25
124	0	30
125	0	27
126	0	169
127	0	728
128	0	0
129	0	3720
130	0	1
131	0	0
132	0	0
133	0	15
134	0	299
135	0	7
Mean	0	195.8

Table 4.5: Number of negative medical terms found in each relevant and non-relevant set for 35 topics

4.4.3 Fields

Here we present the analysis of performing a manual and automatic search for query terms with respect to their corresponding fields. Table 4.6 and 4.7 show the count and percentages of 6 distinguishing sections: explicit – sometimes with headings, and implicit – inferring from the context, showing the total count of all the fields.

Most of the terms in the queries occur under the *Text Body* followed by the *Past History* and the rest of the sections take a fraction of the whole. Therefore we break down the fields into *Past History*, *Text Body* and a combination where terms occur in both fields for a given report. With this simplification we were able to automate the counting process and calculate a statistical test to draw a meaningful conclusion out of the figures.

Fields	**Counts**	**Overall Percentage**
Total	1350	100.0%
Text Body	1091	80.80%
Speculation	17	1.26%
Family History	14	1.03%
Secondary Problem	5	0.37%
Past History	216	16.03%
Etc	7	0.51%

Table 4.6: Percentage and count of query terms across the relevant documents with regards to their context

Fields	Counts	Overall Percentage
Total	6479	100.0%
Text Body	5476	84.75%
Speculation	41	0.62%
Family History	120	1.82%
Secondary Problem	90	1.37%
Past History	741	11.28%
Etc	11	0.16%

Table 4.7: percentage and manual count of query terms across the non-relevant documents with regards to their context

Table 4.8 shows the result of our automatic counts of the query terms. Terms occurring only under the *Past History* and *Text Body* individually are displayed first followed by the terms occurring under both *Past History* and *Text Body*. The result of chi-squared test also confirmed a significant difference for the two cases.

Indeed, query terms related to the *Past History* of a patient occur mostly in non-relevant documents. On the other hand, query terms that occur in both *Past History* and *Text Body* of a document tend to belong to the relevant documents. From this, we can conclude that if a query term occurs only under the *Past History*, the underlying document is most likely to be non-relevant. In other words, terms occurring under *Past History* should be penalised, unless they occur under both fields in the same document.

Fields	Relevant	Non-relevant
Past History	3.98%	21.85%
Past History + Text Body	2.74%	3.20%

Table 4.8: percentage of query terms occurring only under past history and under past history and text body across the relevant and non-relevant documents on document level

4.4.4 Conclusion

The work presented in this chapter was mostly the development of our runs submitted to the TREC-M1 competition. Our runs performed above the median of the 47 judged and 80 unjudged submissions. Combination of knowledge sources, integration of Metamap, and handling negation seem to yield the best outcome.

We learned about the challenges of external query expansion. Extreme size of expansion and the degree of noise introduced for some particular queries posed a challenge that we overcame by the usage of Semantic Types. In the next chapter we will focus on an alternative to external query expansion: Pseudo Relevance Feedback using ICD codes.

Appendix

”hearing loss” patients
complicated receive gerd endoscopy patients
treated endocarditis ”methicillin-resistant staphylococcus aureus” ”hospitalized patients” mrsa
”localized cancer” robots diagnosed prostate patients ”prostate cancer” treated surgery localised
dementia patients
staging ct monitoring pet patients cancer ”positron-emission tomography” ”computed tomography” ”magnetic resonance imaging”
”ductal carcinoma” dcis patients
treated vascular surgically claudication patients
women osteopaenia
discharge hospital haemodialysis patients
”chronic back pain” patients intraspinal pump receive ”pain medicine”
female mastectomies admission ”breast cancer” patients
patients received adenocarcinoma admission revealed adult colonoscopies
patients hospice discharge home adult ”palliative care” admit adult asthma patients
hospital methotrexate received ”cancer treatment” patients
”post-traumatic stress disorder” patients
admission received ”stent, coronary” adults
”anion gap acidosis” patients ”emergency room” ”insulin dependent diabetes” secondary presented adult
chf treatment admit patients
patients give ”emergency department” ”acute coronary syndrome” presented cad plavix
”parenteral nutrition, total” hospital received patients
received hospital diabetic ”diabetic education” patients
episodes hospital patients present acute glaucoma secondary vision loss
infected hiv ”hepatitis c” patients
diagnosis admit ”multiple sclerosis” patients
”obesity, morbid” diabetes ”secondary disease” hypertension patients admit
patients medications ”knee surgery nos” post ”coagulant, nos” hip admit treated
”ct angiography” admit ”chest pain” assessed patients
”physical therapy” children received ”cerebral palsy” admit
”abdominal surgery” invasive patients
fusion patients discectomy admit surgery ”cervical spine”
take patients herbals osteoarthritis products admit care
patients ”chronic disorder” ”seizure disorder” seizure ”seizure activity” chronic admit control
”liver metastasis” treated hospital ”cancer patient” procedure

Table : List of all the words and phrases identified by Metamap and used for all the statistical analysis

CHAPTER 5

Query Expansion using ICD codes

Pseudo Relevance Feedback (PRF) is a type of query expansion that identifies salient terms from the local collection and adds them to the initial query based on the the assumption that top n retrieved documents are relevant. PRF is, therefore, two-fold, running the initial query and then using the retrieval results to expand and re-run the query.

Exploiting the presence of ICD codes (Chapter 2) in clinical records has not been extensively explored for IR. However, since the workers who assign the codes to patient records are required to follow strict guidelines, the use of ICD codes could help to alleviate some of the imprecisions present in a bag-of-words representation of such records. This is particularly important in patient records, as often the free text part of the record will contain speculation, negations (e.g. "the patient does not have X"), references to past conditions, family history of the patient, etc. ICD codes, on the other hand, refer to the current conditions of the patient.

Our focus here is to enhance ranking methods for clinical IR by relying on ICD

V58.66	Long-term (current) use of aspirin
596.7	Hemorrhage into bladder wall
585.6	End stage renal disease
786.8	Hiccough
941.13	Erythema due to burn (first degree) of lip(s)
783.40	Unspecified lack of normal physiological development
V44.4	Status of other artificial opening of gastrointestinal tract
952.08	C5-c7 level with central cord syndrome

Table 5.1: A sample of ICD codes with descriptions

codes in a novel approach through PRF. In this chapter, we explain the ICD codes in more detail and examine a variation of PRF based on mapping of query terms into ICD codes.

5.1 Background

ICD is a disease classification system for health care, providing a system of diagnostic codes with a modest diversity of symptoms, signs and medical findings. ICD is mainly used operationally to assign diagnosis codes for insurance claims in most countries [Puckett 2011].

Other usages of ICD codes are to help with statistics related to the general health of a country, monitor the prevalence of diseases. It is also used for the compilation of national mortality and morbidity statistics. ICD codes have been shown to have problems of completeness and bias [Roque et al. 2011], and this could harm IR effectiveness. The codes are also challenging to work with, as they have a hierarchical structure with different levels of specificity. For instance, *hearing loss* can be linked to many ICD codes, including but not limited to 389.03 (middle ear), 389.0 (conductive hearing loss), and 380.01 (external hearing loss).

The utility of ICD codes is illustrated through the use of this source of information by most participants in the TREC tracks. The most common approach to their exploitation was to expand the text in the medical records, in an attempt to increase the word overlap with the queries (which have no assigned codes). Each ICD code has a short text description and a simple approach used by a number of the TREC participants was to replace each code with its written description.

A different approach was implemented by Limsopatham et al. [Limsopatham et al. 2011a] who expanded the text in medical records with ICD descriptions and words taken from Wikipedia pages related to the ICD codes. Their system performed well (in the TREC-M1 run of the track) in terms of Bpref, where a marginal improvement was gained over their baseline. However for other two measures (R-prec and P@10), it was outperformed by the baseline. In Bedrick et al.'s [Bedrick et al. 2012] work, ICD codes were assigned to queries using an automated method based on a parser, and no clear improvement was reported in this case.

In order to exploit ICD codes for IR, an important step is the automatic mapping of queries into ICD codes. The previous work on assigning these codes to text fragments has focused on document-level evaluation (e.g. patient records); there is no evaluation at the query level that we are aware of.

For patient records, in 2007 Pestian et al. [Pestian et al. 2007] curated a shared task with the goal of fostering research on automatic assignment of ICD codes at the document level. Here, the problem of automatically assigning ICD codes to medical records has been tackled as a classification problem, where a number of training instances were used in a shared task to develop machine learning classifiers to predict the ICD codes in the test data. They provided training data of approximately 1,000 records with 45 ICD codes which made 94 distinct combinations. Various approaches including negation, machine learning and symbolic processing were used by the top participants. However, we do not have training data for classifying medical queries

into ICD codes. We treat this task as an unsupervised problem for which we develop three automatic approaches, that we call ICD coders (see Section 5.2.3.

Later Aronson et al. [Aronson et al. 2007] found that combining the evidence from multiple classifiers and a pattern matching algorithm in a stacking setting, is indeed more efficient than any individual classifier. They also observed a consistent improvement by using negation to discard the negated text from the records.

Prior to the classification of queries to ICD codes, however, finding candidate phrases that are convertible to a matching ICD code requires NLP. Indeed automatic concept recognition is a crucial process and there have been an extensive effort in mapping natural language to medical concepts, mostly those in the UMLS. Programs such as the Metamap, MicroMesh [Lowe and Barnett 1987], KnowledgeMap [Denny et al. 2003], ProMiner [Hanisch et al. 2005] have been developed that facilitate the usage of the UMLS to expand the semantic context of medical text, be it query or document.

Concept recognition programs have been compared in terms of the speed of processing, and their performance in terms of recall and precision. For instance, IndexFinder [Zou et al. 2003] which allows users to specify semantic and syntactic filters, bypasses the expensive NLP processing of the Metamap, and is therefore, faster.

Denny et al. [Denny et al. 2003] compared the performance of the Metamap against their concept mapping program: KnowledgeMap (KM). This was the first comparative analysis of medical curricular documents which were taken from a university lecture notes, handouts and presentations. In this evaluation, KM outperformed the Metamap, both in terms of recall and precision. KM had more success in a number of processes, such as handling of acronyms, abbreviations, and the use of heuristic disambiguation to predict the correct sense of medical terms or acronyms. A large percentage of KM's failure, however, was attributed to terms not being present in the UMLS file: MRCON.

An extensive use of concept identification tool was demonstrated by Gurulingappa et al. [Gurulingappa et al. 2011]. For their medical TREC participation, they utilised a number of tools and dictionaries to identify concepts and the relationships between them. Automatically found concepts and relations from the medical queries were searched against the collection where concepts and relations were indexed separately.

Additionally, they trained a CRF concept identifier over a training data of almost 800 health records. After analysis of results, they found that their system using the CRF identified concepts considerably outperformed other runs. On the other hand using Metamap identified concepts and SemRep relationships in a separate run obtained poor results. This was tracked down to SemRep or Metamap identifying false positive concepts for some cases. However, using an acronym disambiguation strategy by ProMiner helped to alleviate this problem in a different run.

Qi and Laquerre [Qi and Laquerre 2013] experienced another issue with Metamap for certain queries. The following example, taken from their paper, illustrates this difficulty in mapping natural language text to medical concepts. The phrase: "TNF-inhibitor treatments" in a query was converted to 2 concepts: "inhibitor" and "treatments", by Metamap. Missing an important medical concept: *Tumor Necrosis Factor* which was used as an acronym: "TNF" in the query.

Mapping natural language to structured queries is a challenging process. Choosing the correct sense of the term for a given context is not trivial and can lead to erroneous term mapping. The ubiquitous use of synonyms, abbreviations and acronyms in the clinical domain, only compounds this challenge. Moreover, mapping clinical queries to structured concepts such as SNOMED-CT or ICD is usually not a direct process. Indeed, biomedical term mapping tools such as the Metamap do not directly map query terms to structured concepts such as SNOMED-CT or ICD codes. Terms must be converted to UMLS concepts first before being mapped to SNOMED-CT

or ICD codes. Consequently, as some UMLS concepts do not have an equivalent code, some terms can not be mapped at all [Koopman et al. 2011b], generating false negatives.

Amongst the aforementioned programs, however, MetaMap has been consistently improving and is most widely used in IR and text processing of clinical records. In the absence of a realistic scale evaluation of the Metamap [Aronson and Lang 2010], the evidence for Metamaps's usefulness indeed outweighs its shortcomings [Karimi et al. 2011, Koopman et al. 2011c, Amini et al. 2011, Koopman et al. 2016] and makes it a popular choice for medical IR researchers.

5.2 Experimental Setting

In this section first, we describe the process of manual and automatic mapping of queries into ICD codes. Next, we present our IR baseline systems followed by the methodology for ICD based PRF. We used the data from TREC-M1 and TREC-M2 (see Chapter 3) test collections to run the experiments.

5.2.1 Manual Coding of queries to ICD

First, we manually map TREC queries into sets of ICD codes for evaluation of our automatic approaches 5.2.3. Every code has a description which we extracted from the ICD9Data web site[1]; see Table 5.1 for examples of the codes. Note that, the examples are only of type "disease", as the presence of other kinds of ICD codes such as "procedure" was indeed limited in the TREC collections.

The manual look up for ICD codes involved two annotators querying all the disease names found in the TREC queries in order to locate the best ICD matches. Each annotator performed this task separately, and the gold standard was built after

[1] http://www.icd9data.com/

discussion of each disagreement. Some of the queries contained different boolean operators linking the diseases (e.g. "patients with AIDS who develop pancytopenia"), and the way to represent such cases was the main source of disagreement. We decided to represent the queries with ICD codes linked via boolean operators.

In order to measure the level of agreement between the two ICD coders we calculate the overall percentage of the overlap between the ICD assignments per query. The agreement considers whether the annotators assign exactly the same codes, and we report the percentage of overlap between the ICD assignments of the two annotators.

The percentages of agreement for TREC-M1 and TREC-M2 were 55 and 44 respectively. In order to be unanimous in our ICD assignments, we decided not to assign any codes for cases where the matching is an approximation. In other words, we only assigned ICD codes for diseases that were explicitly mentioned and not implied by the intervention. For instance, prior to the given guideline, for the query "Patients who underwent minimally invasive abdominal surgery", one annotator did not assign any ICD code and the other annotator assigned 789.0, 550 and 553 assuming that if the patient had surgery, she must have had abdominal pain. However, based on the aforementioned guideline, the above query did not receive any ICD code.

After the first annotation phase, the annotators discussed all the disagreements and reached joint decisions for each case. We used this final set as the gold standard.

5.2.2 Evaluating the Quality of the ICD Gold Standard

To ensure the quality of our gold standard presented in Section 5.2.1 we gathered all the ICD codes from the relevant documents per query basis. We refer to it as the ground truth and compare it against our gold standard.

The intersection of the ICD codes in the relevant documents turned out to be null for most of the queries. Hence the ground truth for each query is the union of

Query-Number	Query
113	Adult patients who received colonoscopies during admission which revealed adenocarcinoma
122	Patients who received total parenteral nutrition while in the hospital
137	Patients with inflammatory disorders receiving TNF-inhibitor treatments
146	Patients treated for post-partum problems including depression, hypercoagulability or cardiomyopathy
174	Elderly patients with ventilator-associated pneumonia
183	Patients presenting to the emergency room with acute vision loss

Table 5.2: Six queries that do not have matching ICD codes in the collection

the ICD codes found in the documents that are judged relevant for a particular query.

With the exception of six queries given in Table 5.2, all other queries aligned with our gold standard. That is, the entirety or a subset of the gold standard were found in the ground truth.

On aggregate across the whole set of queries there has been 75.46% alignment between the ground truth, and the gold standard. This breaks down to 3 sets of queries, some with no alignment (6 queries), and the other two sets with either 100% (50 queries) or less than 100% alignment (25 queries).

We verified the accuracy of our ICD code assignment, however, we learned that the TREC collection lacks a number of procedural and disease related ICD codes, which is the case for the aforementioned six queries. Furthermore, in alignment with our finding, Bedrick et al. [Bedrick et al. 2012] mention that, as an artefact of the TREC data export process, the number of ICD codes per visits may have been truncated to a certain number, which indicates a possible loss of important ICD codes in some records.

5.2.3 Automatic Coding of Queries to ICD

To build an automatic ICD coder for our PRF, we explored automatic mapping of queries into ICD codes. We developed three ICD coders for the task, each using different resources and means of matching query and the text in target codes. The

first coder relied on word overlap between the ICD description text and query terms. An IR system was used to find the ICD description that best matched the query. The system was configured to use a PL2 [Amati 2003] weighting model. The terms in the queries and ICD descriptions were stemmed using the Porter stemmer [Porter 1980b], and a stop word list from Goodwin et al. [Goodwin et al. 2011] was used. This method assigns a single ICD code to each query.

Our second approach used the information boxes in Wikipedia to obtain the ICD codes of the concepts in the query. The process includes the following steps: (i) apply MetaMap to identify a set of medical concepts in the queries, (ii) automatically retrieve the Wikipedia page for each concept, and (iii) extract the ICD codes found in the information box for each of the retrieved pages.

The editors of Wikipedia often include redirects to a medical term from synonym terms. This means that by searching for any of the variant forms of a given disease, Wikipedia will return the main page describing the concept. For example, Wikipedia does not have a page match for the term *hearing loss*, however an attempt to look up such a page automatically redirects to the page about *deafness*[2], which provides appropriate ICD codes. In this case, each of the phrases identified in the query can provide ICD codes, and we assign all of them to the query. The maximum number of ICD codes assigned to a query by this coder in TREC-M1 and TREC-M2 were 3 and 6, and the averages were 1.45 and 1.95, respectively.

While the second approach uses the MetaMap indirectly to look up for ICD codes, in the third approach, MetaMap is used directly to query UMLS concepts from a UMLS table called *MRCONSO*. The queries are first mapped into UMLS concepts, and the concepts are then used for querying ICD codes.

We illustrate the process with the following query: “patients with AIDS who develop pancytopenia”. In this case, MetaMap recognizes the terms *Patient*, *Devel-*

[2]http://en.wikipedia.org/wiki/Hearing_Loss http://en.wikipedia.org/wiki/Deafness

Method	Average	Min	Max
Manual	1.44	0	6
MetaMap	0.77	0	4
Wikipedia	1.08	0	6
Ranked	1.00	1	1

Table 5.3: Average, minimum and maximum number of ICD assigned to each query using different methods, for TREC M1 and M2 combined

oping, *AIDS* and *pancytopenia*, that is, four UMLS mappings are assigned to the query. The Wikipedia coder uses each of these four terms separately and extracts the ICD codes if they are present in the wiki page. In the case of the MetaMap coder, four SQL queries are submitted to the *MRCONSO* table to find the relevant ICD codes. The mapping for the Ranked coder is straightforward, and the entire query is used to search over the dataset containing all the ICD descriptions.

In this case, we assigned 042, 284.1 manually in the gold standard and the Wikipedia coder assigned 284.1, 042, 044 Ranked method: 248.1 and the direct approach using MetaMap did not find a match.

Table 5.3 provides further statistics on the number of ICD codes assigned to each query. Note that the Ranked-based coder always assigns one ICD code per query.

5.2.4 Baseline Systems

The ranking algorithm for all the baseline systems as well as the ICD based is fixed to be the Inverse Expected Document Frequency model with Bernoulli after-effect and the normalisation 2 from the DFR framework which is available in the Terrier [Macdonald et al. 2012] open source search engine. The DFR ranking models performed the best for the TREC-M1 and TREC-M2 datasets [Amini et al. 2012b].

The DFR models are instantiated by three components of the framework: selecting a basic randomness model i.e., Inverse Expected Document Frequency, in this case, applying the first normalisation, and normalising the term frequencies, for which we use the second normalisation form given in the Equation 5.1.

$$tfn = tf \cdot log(1 + c \cdot \frac{sl}{dl}) \tag{5.1}$$

Where tfn is the normalised term frequency, tf is the term frequency of t in the document, sl is the standard length and dl is the document length and c is the hyper parameter whose value was fixed to 1, which is the default setting.

Our baseline systems incorporated negation detection using the NegEx algorithm with the rule-based Neg-Aggressive setting [Amini et al. 2012b] whereby a concept and its further occurrences are prefixed with the string "no", if the concept is found to be negated at least once within the same report. Altering the word prevents it from being matched when retrieving a positive query term, but it allows to find it in cases where the query is negated (e.g. "Patients taking atypical antipsychotics without a diagnosis of schizophrenia").

Documents were stemmed using the Porter stemmer, and both queries and documents were filtered using the stop-word list recommended by King et al. [King et al. 2011]. All the ICD codes were expanded in the documents replacing the code by the textual descriptions. This was the most common approach to exploit the ICD codes over the last two medical TREC competitions.

We implemented two baseline systems. We refer to our first baseline as ICD-naïve, because the only usage of the ICD codes is via mapping of the numeric representation into the text it refers to. In addition to ICD-naïve, for the second baseline we implement the pseudo relevance feedback system from the Terrier package, to see how our modified PRF performs in comparison. We refer to this system as Traditional-PRF. This system has the same setting as the ICD-based PRF. The key

difference, however, is that the Traditional-PRF uses the text representation of the queries to select the top documents for expansion, rather than the ICD codes.

5.2.5 ICD-based PRF

We illustrate our query expansion IR approach in Figure 5.1. First, the best automatic ICD coder is used to map a query into one or more ICD code(s). These codes are passed to a 'Document Selector' that gathers relevant reports containing at least one of the codes assigned to the query. We then pick the best report per visit by ranking each report separately against the queries. The explanation for which is, adding terms from all the reports can potentially introduce noise as the reports belonging to a visit have the same ICD codes, but do not necessarily have the same context.

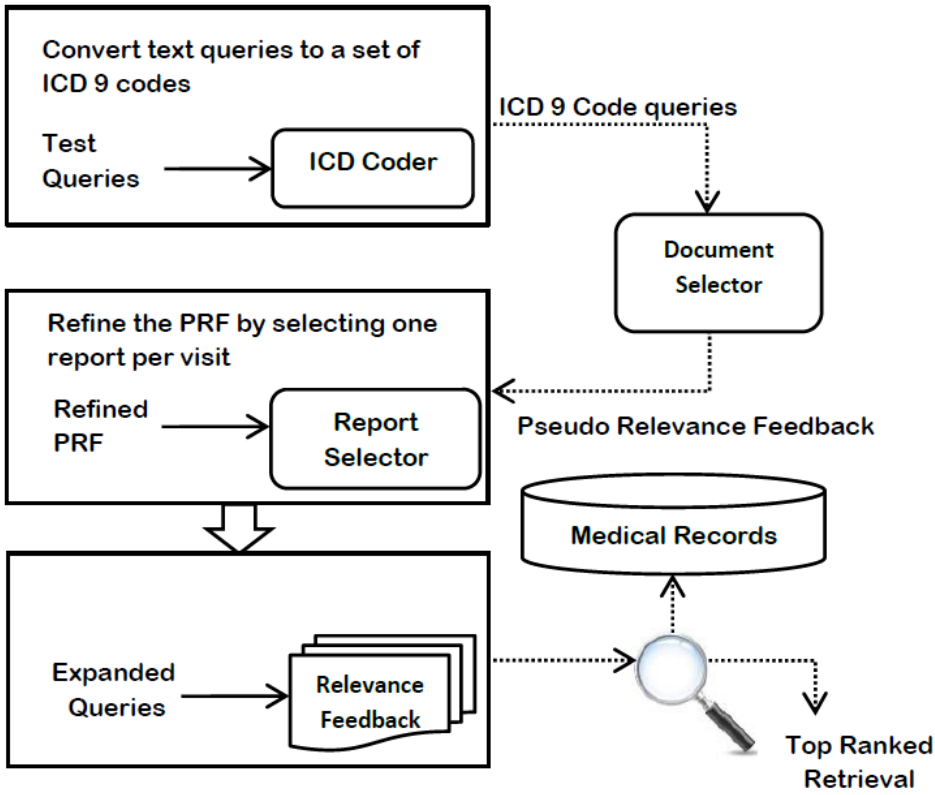


Figure 5.1: The scheme used for expanding the original queries based on the ICD Pseudo Relevance Feedback

We want to select the most informative words to expand a query for which we need to find the most relevant report (best report) within a given visit.

Reports were ranked using the Inexp_B2 ranking function in Terrier against the original query, and the scores were used to determine the best report per visit. We assumed that if a report did not appear in the ranking list, it was not relevant and

hence removed from the PRF. However, there is no limit to the number of visits used to expand each query. However, only the top ranked n terms will be selected to expand each query, such that we avoid the over-expansion problem. The minimum, maximum and mean of visits used for query expansion across two TREC collections are, 0, 91.02 and 611 respectively. In terms of efficiency when the number of records is much higher we may have to limit our search to the top n documents.

Note that each visit in TREC collections maps into multiple reports and not all the reports within a visit relate to one query necessarily, although they all have the same ICD codes.

All the terms in the reports chosen from PRF were weighted using their normalised term frequency by BoseEinstein-1 [Amati 2003] the DFR model for expansion. Finally, we selected the default top forty terms for expansion. We refer to this method as *ICD-based PRF*. We chose to select the default setting of our search engine to add the top forty terms and did not exhaustively tune all the ranking parameters, since the parameter space could be extremely large, and we did not intend to solve an optimisation problem, but rather trying to see the effect of ICD codes in certain conditions.

5.3 Results

Results of the automatic ICD coders, baseline systems and ICD-based PRF are presented here. The evaluation for all the retrieval systems is based on the qrel provided by the TREC-M1 and TREC-M2 test collections.

5.3.1 ICD Coders

We compared the performance of our three ICD coders based on precision and recall using the manually assigned ICD codes as a gold standard. Since the coders do

	Precision	Recall	F1-measure
Ranked	0.39	0.32	0.35
Wikipedia	0.66	0.50	0.57
MetaMap	0.44	0.33	0.37

Table 5.4: Evaluation scores for the three automatic ICD coders on the combined set of TREC M1 and M2

not parse the query for boolean operators and simply return one or more codes, we evaluated them as a multiclass classification problem. However, if the ICD code for a term is within a range, such as the code *140-239* for *Cancer*, we count them as one code for the sake of evaluation.

Table 5.4 shows the performance for each of the techniques. The scores are low, but they suggest that the Wikipedia coder is more reliable for our next experiment. Later we will explore whether the low query-mapping performance can still lead to improved IR results.

5.3.2 Baseline

For better comparison, we measured the effectiveness of the systems against the mean Bpref scores of all submitted runs to TREC. Table 5.5 shows that our baselines are already above the TREC mean for both years of the TREC track. The difference is statistically significant for TREC-M1, but not TREC-M2, suggesting that more sophisticated systems competed in the second edition. For TREC-M1 our baseline is close to the winning participant. Note that we employ a 2 tailed student t-test for all the statistical differences reported in this paper.

System	TREC-M1	TREC-M2
TREC-Mean	0.404	0.305
TREC-Median	0.427	0.328
TREC-Best	0.552†	0.451†
ICD-naïve	0.509†	0.330
Traditional-PRF	0.530†	0.343

Table 5.5: Evaluation of baseline and TREC Best Automatic run and TREC Mean: † indicates significance $p < 0.01$ compared to the TREC-Mean

	ICD-naïve	Traditional-PRF	ICD-PRF
TREC-M1	0.509	0.530	0.547
TREC-M2	0.330	0.343	0.374††
TREC-Combined	0.406	0.422	0.446†

Table 5.6: first two columns present the Bpref scores of our baselines, and are followed by the performances of our ICD-PRF method. The scores are followed by up to two signs. † implies significance ($p \leq 0.01$) difference over the ICD-naïve and †† means significance difference ($p \leq 0.01$) over both ICD-naïve and the Traditional PRF.

5.3.3 ICD-based PRF Evaluation

We tested the systems on three different sets of queries: the 34 queries of TREC-M1, the 41 of TREC-M2 and a combination of both. Table 5.6 shows the comparison between the two PRF methods and the baseline (ICD-naïve). The improvement of the ICD-PRF method over the ICD-naïve is consistent across the different query sets. However, the level of significance is higher for the larger set of queries when the test queries are combined. The ICD based PRF system yields higher scores than the traditional PRF, and the difference is statistically significant for TREC-M2.

While our ICD expansion method does not outperform the best run in TREC-M2 and it performs equally to the best run in TREC-M1, we demonstrated that a different

	ICD-naïve	Traditional-PRF	ICD-PRF
TREC-Combined	0.393	0.386	0.456††

Table 5.7: Performance of systems for a set of perfectly aligned ICD codes double † implies significance difference ($p \leq 0.01$) over both ICD-naïve and the Traditional PRF.

way of incorporating the ICD code is indeed superior to the current approach. Due to the complexity and the engineering efforts required to replicate the best runs in TREC and the fact that TREC 2011 best run was tuned with the author's manually created training data, we were unable to reproduce these systems. However, we implemented two strong baselines, both incorporating the ICD implementation of the TREC best runs. We showed that performing PRF based on the ICD codes is more effective than the conventional method (Traditional PRF) that relies on the text representation of the queries.

To discover the strength of ICD expansion, we hypothesised that this method performs favourably for those queries which have a strict alignment with the codes where all the medically related terms have at least one ICD code.

For instance query 123 and 133 in Table 6.2 are instances of relaxed alignment with ICD codes. They contain terms such as *diabetic education and herbal products* which do not have any related ICD codes. On the other hand the query *Patients with hearing loss* is one with strict ICD alignment.

Overall we found 36 queries with strict alignment across two TREC collections. Table 5.7 shows the result of this experiment. The score shows that the ICD-PRF is significantly superior to other two baselines for these queries.

As queries for finding clinical trials may not always map to a corresponding ICD code, an ideal IR system would take advantage of both the traditional and ICD based PRF, where in case of no ICD mapping the former would be utilised.

5.4 Conclusion

We found that ICD-based PRF outperforms a strong IR baseline, in particular, for queries with perfect ICD alignment. One of the reasons for the good performance of our approach may be that some relevant reports do not have term overlap with the queries. Tinsley et al. [Tinsley et al. 2012] found that for one of the TREC queries, only one visit out of all the relevant visits contained a string from the query, while all the relevant visits contained the corresponding ICD codes. We suspect that there are more of such cases across the collections leading to improvement in effectiveness over the Traditional PRF.

We also showed that the significant improvements over the ICD-naïve proves that there is something to gain by going beyond the simple mapping of ICD codes into their text representation.

Another challenge presented by the queries were *AND* and *OR* conditions found in the formulation of the query. By default, our automatic systems used a logical OR operation to gather all the documents with at least one mention of the ICD codes from the queries. We believe that simple rule-based regular expressions to identify and apply such conditions can yield further improvements.

CHAPTER 6

Pseudo Relevance Judgement (PRJ)

At the heart of every IR test collection is its query relevance judgement (qrel); the only component that relies entirely on the judgement of human assessors. Obtaining human relevance judgement nonetheless is a challenging and expensive process.

It has been said that a reliable IR evaluation often requires a large amount of queries [Urbano et al. 2013]. Obtaining more queries is a simple task but the extra cost of human assessment is often unbearable. Moreover, relevance assessment in the clinical domain, unlike most IR collections, requires domain expertise.

A recent study on the reliability of IR evaluation showed that some of the previously built test collections had low reliability rates based on a series of statistical tests [Urbano et al. 2013]. Furthermore, some important IR evaluation metrics could not be used in the first year of Medical TREC for the lack of relevance assessment [Voorhees and Tong 2011].

A low reliability rate for a test collection is indeed undesirable as it suggests that conclusions drawn on the performance of IR systems are not reliable, and that

can prevent an additive improvement of search systems in the field of IR.

The insufficient depth of assessed documents due to the lack of human resources in the first Medical TREC (TREC-M1) challenge has been the main motivation for us to test the feasibility of modelling a PRJ for a clinical test collection. A system that does not depend entirely on human assessment. We intend to provide grounds for a more cost-effective and comprehensive evaluation framework.

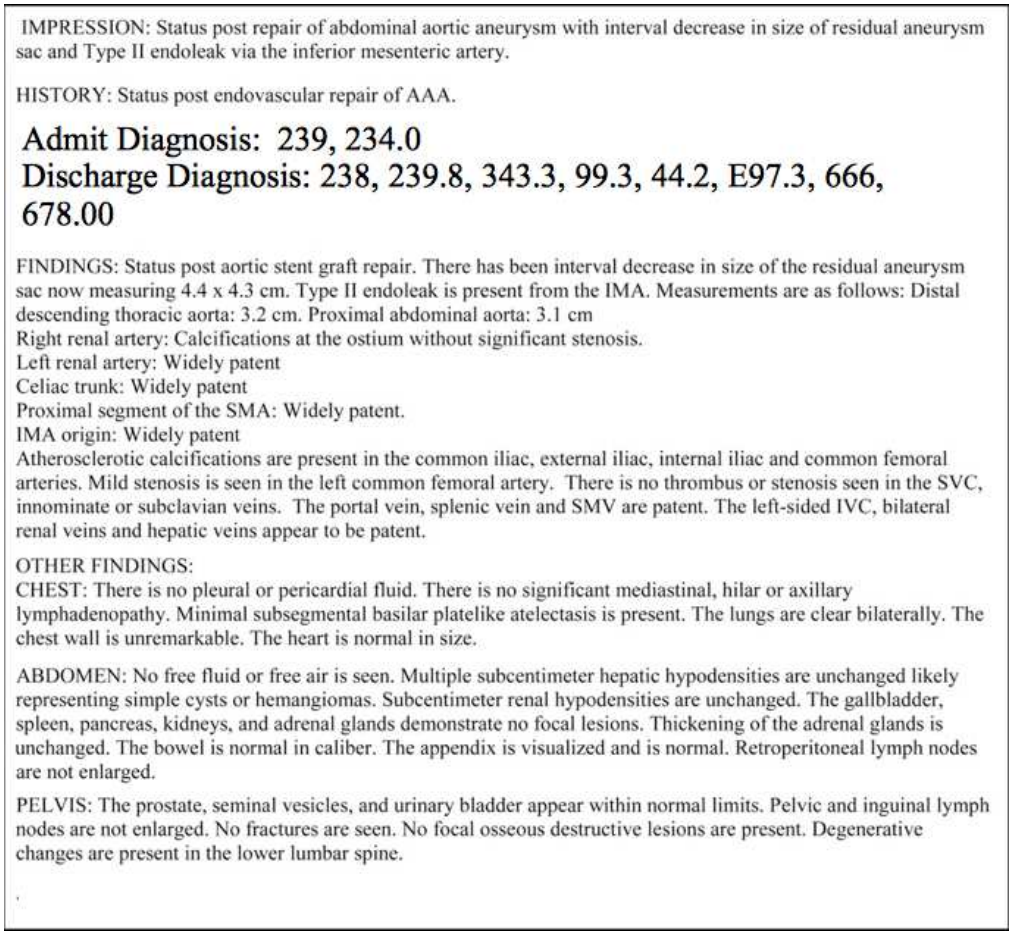
In the previous chapter, we explored an application of ICD codes with regards to the retrieval of EHR. However, due to their unique characteristics, we believe that ICD codes can be a suitable replacement for the qrel, hence the name, pseudo relevance judgement. ICD codes are assigned to all the reports in the collection. They can potentially be used to describe the main theme of the reports, and unlike the text in the records, they are concise, specific and contain no misleading embellishment such as negation or speculation.

Figure 6.1 is an example of a made up electronic health record which shows a typical format of such records with ICD codes occurring in two distinct sections: *Admit Diagnosis* and *Discharge Diagnosis*.

We hypothesised that, given the characteristics of ICD codes, if correct ICD codes were assigned to the queries, matches between the ICDs in the queries and the reports can be a positive signal for relevance.

ICD codes are first assigned to the queries and the records with the matching ICDs are identified to build a pseudo qrel. The final stage is to rank all the TREC runs [1] based on the PRJ and the real qrel and compare the two. The question was, whether the evaluations of the runs by the PRJ correlate sufficiently with the official qrel. For our experimentation, we collected a total of 127 runs from 29 groups in TREC-M1 and a total of 88 runs from 24 participating groups in TREC-M2.

[1] A TREC run is a ranking of reports produced by a participating system of TREC, in other words a run is a submission and each group could submit up to 4 runs.



IMPRESSION: Status post repair of abdominal aortic aneurysm with interval decrease in size of residual aneurysm sac and Type II endoleak via the inferior mesenteric artery.

HISTORY: Status post endovascular repair of AAA.

Admit Diagnosis: 239, 234.0
Discharge Diagnosis: 238, 239.8, 343.3, 99.3, 44.2, E97.3, 666, 678.00

FINDINGS: Status post aortic stent graft repair. There has been interval decrease in size of the residual aneurysm sac now measuring 4.4 x 4.3 cm. Type II endoleak is present from the IMA. Measurements are as follows: Distal descending thoracic aorta: 3.2 cm. Proximal abdominal aorta: 3.1 cm
Right renal artery: Calcifications at the ostium without significant stenosis.
Left renal artery: Widely patent
Celiac trunk: Widely patent
Proximal segment of the SMA: Widely patent.
IMA origin: Widely patent
Atherosclerotic calcifications are present in the common iliac, external iliac, internal iliac and common femoral arteries. Mild stenosis is seen in the left common femoral artery. There is no thrombus or stenosis seen in the SVC, innominate or subclavian veins. The portal vein, splenic vein and SMV are patent. The left-sided IVC, bilateral renal veins and hepatic veins appear to be patent.

OTHER FINDINGS:
CHEST: There is no pleural or pericardial fluid. There is no significant mediastinal, hilar or axillary lymphadenopathy. Minimal subsegmental basilar platelike atelectasis is present. The lungs are clear bilaterally. The chest wall is unremarkable. The heart is normal in size.

ABDOMEN: No free fluid or free air is seen. Multiple subcentimeter hepatic hypodensities are unchanged likely representing simple cysts or hemangiomas. Subcentimeter renal hypodensities are unchanged. The gallbladder, spleen, pancreas, kidneys, and adrenal glands demonstrate no focal lesions. Thickening of the adrenal glands is unchanged. The bowel is normal in caliber. The appendix is visualized and is normal. Retroperitoneal lymph nodes are not enlarged.

PELVIS: The prostate, seminal vesicles, and urinary bladder appear within normal limits. Pelvic and inguinal lymph nodes are not enlarged. No fractures are seen. No focal osseous destructive lesions are present. Degenerative changes are present in the lower lumbar spine.

Figure 6.1: A sample of an anonymized health record with a set of ICD code

Throughout this chapter, we refer to two versions of relevance judgements, namely pseudo and real qrels. Real qrels are the relevance judgements provided by the TREC challenge organisers, and they are built by manually assessing documents. Pseudo qrels are built automatically for each query, by associating ICD codes to the

query, and then considering all the documents in the collection which have at least one of those codes as relevant for the query.

6.1 Background

Generating relevance judgments is always time-consuming and due to limitations on human resources, the TREC community conventionally assesses only a portion of the collection; usually by incorporating the top k documents from all the runs submitted by the participants. This is known as *depth-k pooling*. Shallow pooling techniques are those where the average number of judged documents per query is low. In the TREC 2011 medical track, shallow pooling was employed, and the cost of obtaining manual judgments was highlighted by the track organisers.

Since shallow pooling makes evaluation less reliable, means of alleviating the problem of obtaining judgments are always being sought.

There is very limited research on evaluating IR systems without manual relevance judgment to the best of our knowledge and we understand that IR has not yet arrived at a reliable approach of automating it. However, we seek to promote further research on this topic for clinical IR.

Soboroff et al. [Soboroff et al. 2001] produced a widely cited paper, which described an attempt to automatically build relevance judgments by using patterns of occurrence of documents retrieved by multiple IR systems. The results appeared promising although there were limits to the quality of the judgments, particularly for measuring highly effective retrieval systems. A conclusion that can be drawn from Soboroff et al.'s work is that some level of manual intervention is required when forming relevance judgments.

More recent work on pseudo-relevance judgment [Sakai and Lin 2010] has shown that by relying on documents retrieved frequently by a diverse set of systems it is pos-

sible to build relevance judgments automatically, and achieve high correlation with manually judged data. They concluded that a simple method based on the ordering of the documents in the pool by the number of runs that returned each document at or above rank 30, performed as well as any other existing system. Their method yielded higher correlation on the NTCIR collection than TREC. Nonetheless, all the past work relied on the pool of runs gathered from the participants in the shared tasks, which is not a realistic option at the time of building a test collection.

Building on ideas tried before, where category structure was used as a substitute for relevance judgments (e.g. [Harmandas et al. 1997]) Koopman et al. [Koopman et al. 2011a] treated the written descriptions of ICD codes as queries and the documents containing those codes as the relevant set for that query. However, the queries were artificially created, and it is not clear whether they would be representative of real world questions formulated by medical professionals. In addition, the quality of these wholly simulated judgments could not be compared to the manual judgments produced for TREC, as they used different queries.

While other attempts revolved around simulating the entire qrel, Mollá et al. [Mollá et al. 2014], assuming relevant documents bear some degree of resemblance, automatically completed partial and limited qrels, by gathering unjudged documents that are similar to the judged relevant documents. They reported positive correlation when the number of available qrel is very limited.

6.2 Experimental Setting

In this section first, we describe the evaluation process for the PRJ against the official qrel, and then present the development of the PRJ approach. We used the data from TREC-M1 and TREC-M2 (see Chapter 3) test collections to run the experiments.

6.2.1 Evaluation: Kendall's tau

Evaluating the performance of the PRJ involves testing how similar the pseudo rankings are to the ranking produced by the official relevance judgment. Kendall's tau has been widely used in IR for such purposes. It measures the number of pairwise swaps between the rankings until the two are the same, and is normalised in a way that it produces 1.0 if the ranking are the same and -1 where the rankings correlate reversely. Equation 6.1 is the formula for Kendall's τ that we used, which is the proportion of concordance pairs C versus the proportion of discordance pairs D, where n is the total number of pairs.

$$\tau = \frac{C + D}{\frac{1}{2}n(n-1)} \tag{6.1}$$

6.2.2 Pseudo Relevance Judgment derived from ICD Codes

We tested whether relevance judgments for a medical test collection can be derived from ICD codes. We hypothesised that a report that shares the same ICD code with a query is relevant to that query. While all the records have assigned ICD codes, the TREC queries do not have any ICD codes.

First, we assigned the codes to each of the queries in TREC-M1 and M2 by relying on both manual and automatic ICD coders described in Chapter 5. Documents containing at least one of these codes were assumed to be relevant for the given query (pseudo-relevant from now on). We tested this approach by comparing the way the two forms of qrels (the real ones and pseudo-qrels) ranked retrieval systems against each other; a correlation-based method that is widely applied in the IR literature (e.g. Büttcher et al. [Büttcher et al. 2007]).

For this experiment, we rely on the assignment of ICD codes to queries (manual or automatic) to build our pseudo-relevance judgments. Once we assign the codes to

the query, we consider all documents carrying at least one of the codes as relevant (and the rest as non-relevant).

After building the pseudo relevance judgments in this manner, we collect all the runs submitted to the TREC-M1 and TREC-M2 evaluations (downloadable from the TREC website) and use the same set of 34 and 47 queries to rank all the systems based on the Bpref scores obtained by using the official qrels and the qrels from the PRJs. We use Bpref because it was the metric of choice for TREC-M1. Bpref was selected because of its robustness for incomplete judgments sets, since it is computed on the basis of judged documents only [Buckley and Voorhees 2004]. Bpref is inversely related to the fraction of judged non-relevant documents that are retrieved before judged relevant documents. The inferred metrics chosen for TREC-M2 had stability problems when applied in TREC-M1.

6.3 Results

The evaluation for PRJ is based on the official qrel provided by the TREC-M1 and TREC-M2 test collections.

To calculate the final results we need to measure the Kendall τ correlation between the ranking of runs based on the PRJs (from manual and automatic ICD code assignments), with the ranking of runs based on the original TREC manual judgments. The values of τ are shown in the top two rows of Table 6.1. We find positive correlations in all cases and the results are similar to the previous pseudo-relevance correlation scores [Soboroff et al. 2001], where τ correlation numbers were reported to range from 0.369 to 0.571. The correlation was much higher for TREC-M2 than for TREC-M1, and this suggests that the PRJ would be more appropriate for the queries in TREC-M2 (both with manual and automatic assignment). For TREC-M1 the correlation scores are low, and surprisingly the use of manual ICD codes performs

slightly worse than the automatic system.

To this end, we can see that the rankings obtained by the PRJ have a low correlation with the rankings of the official qrel. However, it is important to note that even human assessors do not reach a perfect correlation of 1. As an additional step, we wanted to see how well the official qrel correlates with itself, and to use this correlation as a reference in our experiment.

To do this, we calculated the correlation of the system rankings, when measured by different splits of the queries using only the official qrels. This is referred to as *Data-based reliability indicator* and has been used to measure test collection reliability in the past [Urbano et al. 2013].

The intuition behind this idea is that the rankings resulting from the different subsets of queries should present a reasonably high correlation, given that the queries and relevance judgments originate from the same source, and that the assessments have been done manually.

Ideally, however, we would have two human assessors, separately judging the relevancy for each set of the query, and find the correlation. A costly process, which does not require the query split and iterations of the *Data-based reliability* test, but demands extensive manual assessment.

For each collection (TREC-M1 and TREC-M2), the queries were randomly divided into two equal subsets. However, since we can only use half of the queries in each collection, the random division of the queries was performed for 1,000 iterations to find the highest score, to make sure that this reference score is not underestimated.

For TREC-M2, to make the number of queries equal in each subset, we eliminated the last query, i.e., query 182. We obtained the Kendall τ for every iteration and selected the highest correlation score for TREC-M1 and TREC-M2. The Kendall τ for this reference approach is given in the bottom row of Table 6.1, named "Max(Official Query Split)". We can see that the τ scores are close to the manual

pseudo-relevance for TREC-M1, and for TREC-M2.

Note that, we take into consideration that in "Max(Official Query Split)" we have to split the queries as we do not have two separate qrels for each set of query. Therefore, we conclude that the use of PRJ performs similarly to the reliance on half of the queries of the collection with real qrels. For a safe evaluation of the results, however, we adhere to the de-facto minimum of the τ of 0.9 established by Voorhees [Voorhees 2000b]. According to which, the PRJ correlations are not ideal for a reliable evaluation.

	TREC-M1	TREC-M2
PRJ-Manual	0.35	0.59
PRJ-Automatic	0.37	0.50
Max(Official Query Split)	0.41	0.53

Table 6.1: τ Correlations between PRJ and official relevance judgement

Table 6.3 presents a subset of queries, including their manually-assigned ICD codes and their corresponding τ correlations between the PRJ and the official relevance judgment. This subset represents the four queries with the highest and the lowest correlations, and the descriptions of the codes are given in Table 6.2. The queries at the top and the bottom do not seem very different, with most of the queries having close matches to ICD codes, and containing restrictions that cannot be directly captured with ICD codes (e.g. "Imaging studies"). However, the differences in performance suggest that some restrictions have a greater effect in the relevance of documents.

Figure 6.2 and 6.3 graph the official TREC Bpref scores on a per query basis, which we compare to the Bpref scores obtained using our manually created qrels for TREC-M1 and TREC-M2 respectively. It can be seen that TREC-M2 is more consistent with the official scores and the top and low systems are mostly similar.

ICD code	Description
296.2	Major depressive disorder single episode
296.3	Major depressive disorder recurrent episode
311	Depressive disorder, not elsewhere classified
724.2	Lumbago
410	Acute myocardial infarction
411	Other acute and subacute forms of ischemic heart disease
412	Old myocardial infarction
413	Angina pectoris
414	Other forms of chronic ischemic heart disease
346	Migraine
250	Diabetes mellitus
715	Osteoarthrosis and allied disorders
155.0	Malignant neoplasm of liver, primary
155.2	Malignant neoplasm of liver, not specified as primary or secondary
584	Acute kidney failure
585	Chronic kidney disease (ckd)

Table 6.2: ICD codes, and their corresponding definitions.

Figure 6.4 and 6.5 show the same correlation using the Wikipedia-based automated ICD coder explained in Chapter 5, where it seems that the signal is weaker in this case.

Query ID	Query	ICD codes	Tau Correlation
156	Patients with depression on anti-depressant medication	(296.2 OR 296.3 OR 311)	0.71
160	Patients with Low Back Pain who had Imaging Studies	724.2	0.685
182	Patients with Ischemic Vascular Disease	410-414	0.681
148	Patients acutely treated for migraine in the emergency department	346	0.669
123	Diabetic patients who received diabetic education in the hospital	250	0.052
133	Patients admitted for care who take herbal products for osteoarthritis	715	0.059
135	Cancer patients with liver metastasis treated in the hospital who underwent a procedure	(155.0 OR 155.2)	0.068
110	Patients being discharged from the hospital on hemodialysis	(584 OR 585)	0.076

Table 6.3: Four highest and lowest τ correlations between the official and the PRJ using the manually assigned ICD codes

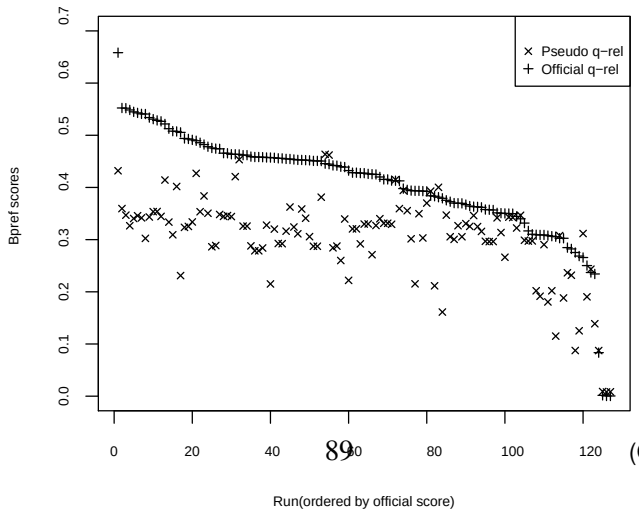


Figure 6.2: Official Bpref scores vs. ICD based Pseudo Relevance judgment for TREC-M1 with manual ICD assignment

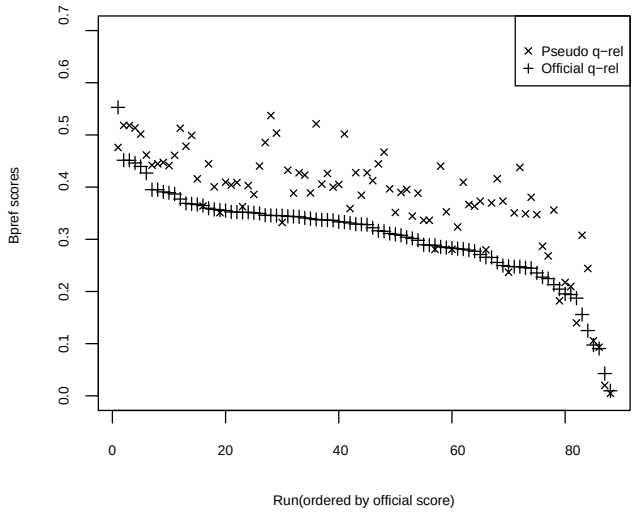


Figure 6.3: Official Bpref scores vs. ICD based Pseudo Relevance judgment for TREC-M2 with manual ICD assignment

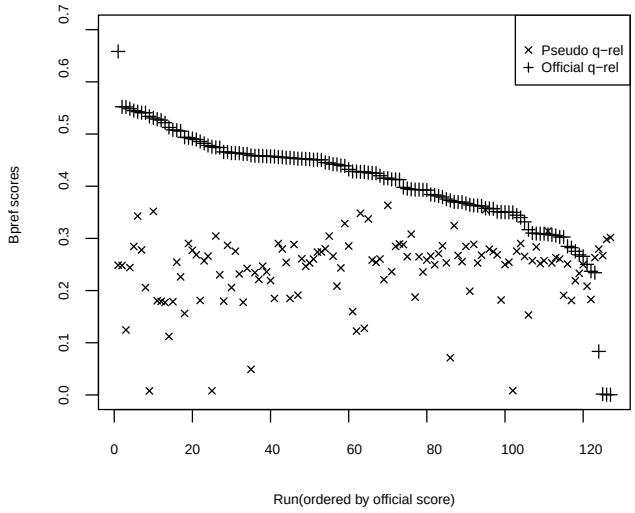


Figure 6.4: Official Bpref scores vs. ICD based Pseudo Relevance judgment for TREC-M1 with automatic ICD assignment

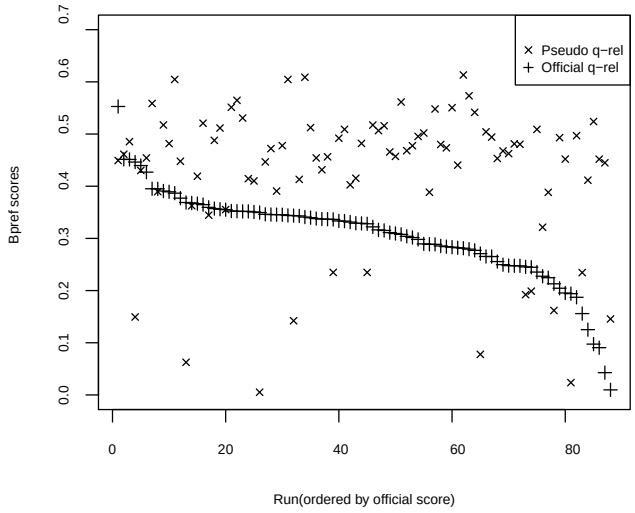


Figure 6.5: Official Bpref scores vs. ICD based Pseudo Relevance judgment for TREC-M2 with automatic ICD assignment

6.4 Discussion

Our intention for performing 1,000 iterations of splitting the queries for official relevance judgement was to see how well the official qrel correlates with itself. To evaluate the PRJ, however, we directly calculate the correlations with the official qrel without splitting of the queries. Despite the mediocre correlations of 0.35 and 0.59 for TREC-M1 and TREC-M2, the lower number of the former i.e. TREC-M1 was questionable.

There is room for investigation of the causes behind the correlation differences.

Qualitative analysis between the human assessments across the collections is difficult to measure. Nonetheless, there are important variables that we can analyse.

For instance, the variability across the TREC runs, can be a clue to discover the reason behind this difference. The mean for the TREC-M1 for 1,000 iterations was 0.33 with the standard deviation of 0.0364, the same values for TREC-M2 were 0.41 and 0.0392 respectively. However, the low variability rules out system variation as a reason to affect the correlation scores.

We, therefore, look at other variables, namely: size of the queries, qrel, and both together. The average size of the judgment set for TREC-M2 (512) was almost twice the size of TREC-M1 (260.7) and TREC-M2 contained 15 more queries. Therefore we look at the effect of the query and the qrel size by truncating them in TREC-M2.

Table 6.4 shows the tau correlations for TREC-M2 when queries and qrel are increasingly reduced. We randomly discard a percentage of queries and qrels, per query basis, starting from 20%, cutting down to almost half, equalising it to the size of relevance judgements in TREC-M1.

percentage	query	qrel	qrel-and-query
0	0.59	0.59	0.59
20	0.48	0.51	0.49
40	0.51	0.52	0.51
60	0.49	0.52	0.50

Table 6.4: τ Correlations with query and qrel reduction in TREC-M2

There is an 11% drop when queries are reduced by 20%, and the correlations for qrel and query-qrel reduction, range from 0.49 to 0.51. However, the correlation scores do not consistently drop by increasing the percentage of discarded qrels or queries. The results show that smaller size of queries and qrels can affect the correlation with the PRJ. However, this effect doesn't seem to cause a drastic change in

the correlations.

Therefore, it is evident that other factors affect the lower correlations in TREC-M1. In fact, the analysis of variance and studies on inter-assessor disagreements, have been done extensively in general IR tracks [Voorhees 2000a, Webber et al. 2012], including the TREC medical track [Urbano et al. 2013] that was used for this experiment. In this study [Urbano et al. 2013], based on a series of statistical tests it was shown that the TREC-M1 had one of the lowest reliability measures, which reassured our assumption about the quality of this test collection.

Nonetheless, the correlations reported in our research were both relatively low, and that is what we were concerned with. This experiment needs to be done on other medical collections, ideally those with higher reliability factors and more official qrel when a collection of appropriate size become available.

6.5 Conclusion

We analysed the feasibility of a potential alternative to replace the time-consuming process of human relevance judgment. ICD codes can designate the primary theme of the medical reports and therefore they were found to be suitable for simulating traditional manual relevance judgments.

The research by Koopman et al. [Koopman et al. 2011a] was a short research paper to facilitate the creation of a clinical evaluation platform using ICD codes to designate relevant documents. This work was important as it introduced the first empirical evaluation framework for medical IR in the context of clinical records. The collection that was later used as the dataset for the medical TREC.

Since this study was conducted prior to the TREC medical track, it did not include any queries or relevance judgement. In fact, the automatic process to build the qrel and queries was the contribution of this work. ICD codes were extracted

from the collection and converted to their textual descriptions as defined in the ICD taxonomy. The textual descriptions were used as medical queries and the record ids that contained the ICD codes were added to a file to create the qrel. The average query length was reported to be 18 words with each query fetching an average of 231 relevance judgements i.e. relevant report ids.

The quality of their qrel was negatively affected by the way of ICD assignments across the collection. For instance, they found that for some diseases, a choice of a generic ICD code was preferred to a more specific and relevant code, or conversely some generic concepts such as "Kidney" were hardly used across the collection.

In a small experiment, they used the Indri search engine to run the queries, with a number of different indexes which varied in terms of, what parts of documents were excluded for each. They found that excluding administrative ICD codes from the reports produced the highest MAP score.

However, measuring the reliability of the test collection and the feasibility of PRJ was beyond the scope of this paper, and therefore the reported scores may not represent the actual results. Our experiments shed light on these aspects of using ICD codes for an IR evaluation framework in a clinical context.

Despite the positive correlations, we can see that relying solely on ICD codes does not provide a reliable evaluation framework and this can be taken into account by efforts such as Koopman et al. [Koopman et al. 2011a].

We observed that ICD codes do not necessarily capture some aspects of medical queries such as population or medical devices, and that limitation can negatively affect their usage for the PRJ.

In order to incorporate the information about the population or medical devices, if present in the queries, we need to rely on other sources such as SNOMED-CT concepts. However, no publicly available medical records, to the best of our knowledge, contains such meta-data or the latest version of the ICD codes (ICD 10). We recog-

nise this limitation and await the future shared tasks to provide the latest electronic health records to the community.

CHAPTER 7

Conclusion

Improving patients record search involved researching two general aspects of the field that we recognised vital. In a nutshell, the first half of this thesis analyses two different ways of improving the effectiveness of clinical IR, and the second half investigates ways of automating the manual work of human assessment.

The usage of ICD codes was the highlight of this thesis which has opened up possibilities to further research on their usabilities, both by improving the current state and also finding new applications. In particular the automation of relevance assessment is of interest and can benefit the IR community through allowing the creation of more test collections in the clinical domain.

ICD codes were particularly suitable for this research as opposed to other medically available codes such as SNOMED CT concepts [Spackman and Campbell 1998] or Concept Unique Identifiers [Aronson 2001] (CUIs) because they are assigned to clinical reports. They can be used to describe the main theme of the reports and therefore be harnessed to judge the relevancy of them as well.

For the first part of this thesis, we focused on improving IR effectiveness for the retrieval of patient cohorts by two different means.

We tried to address the following research question: *How to effectively find patient cohorts for research studies? (Effective Retrieval)*

The use of external knowledge sources (Chapter 4) to tackle this question, was more challenging and less effective than relying on the local/internal meta-data, the ICD codes(Chapter 5). Using external and internal sources was meant to compensate for the vocabulary mismatch between the queries and the documents.

Our novel method which exploited the ICD codes using PRF was our main contribution towards the enhancement of effective retrieval of patient cohorts. Here the results were indeed promising, and we measured significant improvements over the traditional PRF approach. Another conclusion that we drew from this work was that our approach to using the ICD codes is indeed superior to the simple mapping of the codes that was implemented by most TREC participants.

For the second part of this thesis we asked the question: *What are the possible ways to alleviate automatic query relevance judgement for clinical records? (Effective Evaluation)*

Here we dealt with the possibility of creating more clinical test collection by building automatic relevance judgement. We found that our approach to evaluate the TREC challenge runs using simulated relevance judgments had a positive correlation with the TREC official results.

We believe that this was the first time that the ICD codes were utilised for PRJ over real queries. Although we observed positive correlations for TREC runs, we also noticed the difficulty of relying on this sole source of evidence for a sound evaluation.

There is room for the improvement of PRJ. For instance, one step towards the refinement of our pseudo judgment and pseudo relevance systems is to automatically designate how relevant a report is for a given query. Human assessors can determine

the level of relevance ranging from 0 non-relevant to 2 highly relevant, whereas, our approach indiscriminately assigned 1 to all the relevant visits. In the PRF model, this can potentially help in the reduction of the less informative reports. This could also increase the correlation with the official relevance judgment for the PRJ model. While there are ways to compensate for this we left the automatic grading of the relevant documents for future work.

Further to the main contributions of the thesis, we included three appendices. Appendix A represent our work in progress to establish connections between two different corpora in health: clinical records and medical papers. We aim to build a framework to provide easy access to both medical articles and clinical records with an ability to locate relationships between them. To the best of our knowledge, the task of extracting and linking evidence from clinical records to scientific research has not been tackled before and our goal to initiate it can create a new field of research in the field of bioinformatics and health search.

We organised a shared task to build automatic sentence classifiers to map the content of biomedical abstracts into a set of pre-defined categories. The contribution here besides creating a shareable dataset was finding a system that improved the state of art in sentence classification of the medical abstracts. As stated before, this was a beginning of a broader work, that will use the predictions of the sentence classifier to connect certain segments of the medical abstracts to clinical records.

Appendix B and C demonstrated an alternative work on PRJ. This work differed from the PRJ in Chapter 6 in two ways. The dataset used here was a collection of medical abstracts, and the work carried out was rather extending a limited set of qrels as opposed to creating a new set.

The best results were achieved when the number of available relevance judgements were more limited. While our technique does not fully correlate to the ranking of the real qrel, we suggest that it can be used during the development stage of infor-

mation retrieval systems for the sake of training.

We tested our techniques on a specific test collection in the medical domain, and a potential future work will determine how well these findings carry to other domains.

7.1 Summary

Past work reflects certain challenges in the field of medical IR. We present, in a short summary, what we perceive as important challenges and areas that require further attention. We also present our initial attempt to operationalise the ICD approaches in this thesis for a hospital, showing the possibility of extending this work onto real world datasets.

7.1.1 Challenges and Gaps

Data scarcity is indeed a major obstacle for advancements in the field of clinical IR as there is a need for further research to establish standardised baselines or evaluation metrics [Goeuriot et al. 2016; 2014a]. Nonetheless, there has been a continuous effort, mainly by the TREC community, to alleviate this problem by providing other medical test collections for IR. A shift from patient cohort search [Voorhees and Tong 2011, Voorhees and Hersh 2012] to Clinical Decision Support (CDS) in 2014 and 2015 in TREC shared tasks [Roberts et al. 2015a;b] prompted research on medical IR again. However, in these tracks, only short clinical records were used as queries, and documents were medical articles, taken from a subset of the Pubmed. These tracks assist in advancements of techniques to support evidence-based medicine and do not provide access to actual clinical records.

Clinical records are a complex type of medical data, made up of clinical notes, pathology data, patient history, etc. Exploiting the unique characteristics of such

records that we explained in Chapter 4, demands more concrete work. For instance, taking advantage of the structure and the existing fields in clinical records have not been explored to sufficient lengths [Demner-Fushman et al. 2011] and therefore remains an unsolved problem.

Furthermore, the growing volume of medical terminologies and the variations that it may cause in posing medical queries, not only challenges IR systems, but also pose cognitive work-load to human assessors to judge the relevancy of clinical records [Koopman and Zuccon 2014]. Therefore, to ensure the quality of relevance judgements, cognitive load for human assessors may need to be taken into account [Koopman and Zuccon 2014] in the design of IR mechanisms.

Varying information need in medical IR is yet another challenge that demands extra processing of queries and docuements [Goeuriot et al. 2016]. For instance, a general practitioner might require basic information to advise a patient, but a specialist may need a more in-depth and comprehensive information for deciding on a course of treatment [Goeuriot et al. 2014a]. The recent TREC CDS tracks [Roberts et al. 2015a;b] partly focus on this problem. Koopman et al. [Koopman et al. 2017] use data from the CDS tracks, categorising information need into 3 groups: Treatmant, Diagnosis and Test, and structuring IR systems around them. While reporting gains in retrieval effectiveness and saving work-load for searchers, the obtained results show that more work is needed to further analyse and improve systems for treatment related tasks.

Lastly, to the best of our knowledge, efforts to exploit coding systems to perform pseudo relevance judgement are in their infancy. The early work by Koopman et al. [Koopman et al. 2011a] had some limitations. In this study, ICD codes were used to simulate a test collection out of approximately 81,000 electronic medical records. However, queries were formed artificially from the existing ICD codes in the collection, and it is not clear whether they represent real world queries posed

by medical professionals. In addition to the above shortcoming, the quality of these wholly simulated judgments could not be compared to the manual judgments produced for TREC, as they used a different set of queries.

Although our study on the usefulness of ICD codes for pseudo relevance judgement [Amini et al. 2015] addressed these limitations, it was negatively impacted in situations where query terms did not map to any ICD codes. Future work must entail a trial of approaches to find ways to alleviate this shortcoming.

7.1.2 Real World Applications

Implementing and testing the effectiveness of our developed techniques for real users is of great importance. In order to test the usability of our proposed ICD based approaches, we sought ways to extend this research and endeavours alike on to the real world applications such as hospitals. The history, implications and global usage of ICD codes are thoroughly discussed in the field of medicine [Manchikanti et al. 2011a;b, Beach 2012], and it is known that the ICD that we used for this thesis are very essential to patient records and its current version has been used in about 110 different countries [Manchikanti et al. 2011a].

Our preliminary collaboration with the bioinformatics department of an Australian hospital revealed their mere reliance on relational databases to perform basic searches on their patient records. Their search space is limited to the kind of SQL queries that can be executed on their systems. We are also aware that the dataset they use contains appropriate meta-data such as ICD codes.

Indeed, the reliance to ICD codes in both the PRJ and the PRF approaches gives us a chance to widen and improve their basic search, and to make further progress to enhance research on the PRJ.

Apart from search, however, another important application for this hospital is to replace the expensive process of manual assignment of ICD codes. At this stage, we

can not completely automate this process, however, we plan to produce a list of ICD suggestion for each record from which an ICD coder can select, remove or add other ICD codes.

The applications of ICD codes are generalizable within the field of electronic clinical records. ICD is widely implemented in hospitals of 110 countries, and the contributions of this research is indeed applicable beyond the TREC collections.

CHAPTER A

Appendix 1

A.1 Sentence Classification for Medical Abstracts: ALTA Shared Task

The ALTA shared task ran for the third time in 2012, with the aim of bringing research students together to work on the same task and data set, and compare their methods in a current research problem. The task was based on a recent study to build classifiers for automatically labeling sentences to a pre-defined set of categories, in the domain of Evidence Based Medicine (EBM). The partaking groups demonstrated strong skills this year, outperforming our proposed benchmark systems. In this work we explain the process of building the benchmark classifiers and data set, and present the submitted systems and their performance.

Medical research articles are one of the main sources for finding answers to clinical queries, and medical practitioners are advised to base their decisions on the available medical literature. Using the literature for the purpose of medical decision

making is known as Evidence Based Medicine (EBM).

According to the EBM guidelines, users are suggested to formulate queries which follow structured settings, and one of the most used systems is known as PICO: Population (P) (i.e., participants in a study); Intervention (I); Comparison (C) (if appropriate); and Outcome (O) (of an Intervention). This system allows for a better classification of articles, and improved search. However curating this kind of information manually is unfeasible, due to the large amount of publications being created on daily basis.

The goal of the ALTA 2012 shared task was to build automatic sentence classifiers to map the content of biomedical abstracts into a set of pre-defined categories. The development of this kind of technology would speed up the curation process, and this has been explored in recent work [Chung 2009, Kim et al. 2011]. One of the aims of this task was to determine whether participants could develop systems that can improve over the state of the art.

A.2 Dataset

Different variations and extensions of the PICO classification have been proposed and the schema used for this competition is PIBOSO [Kim et al. 2011], which removes the *Comparison* tag, and adds three new tags: *Background* , *Study Design* and *Other*. Thus, the tag-set is defined as follows:

- *Population*: The group of individual persons, objects, or items comprising the study's sample, or from which the sample was taken for statistical measurement;
- *Intervention*: The act of interfering with a condition to modify it or with a process to change its course (includes prevention);

- *Background*: Material that informs and may place the current study in perspective, e.g. work that preceded the current; information about disease prevalence; etc;
- *Outcome*: The sentence(s) that best summarise(s) the consequences of an intervention;
- *Study Design*: The type of study that is described in the abstract;
- *Other*: Any sentence not falling into one of the other categories and presumed to provide little help with clinical decision making, i.e. non-key or irrelevant sentences.

We rely on the data manually annotated at sentence level by [Kim et al. 2011], which consists of 1,000 abstracts from diverse topics. Topics of the abstracts refer to various queries relating to traumatic brain injury, spinal cord injury, and diagnosis of sleep apnoea. Over three hundred abstracts are originally structured, that is, they contain rhetorical roles or headings such as *Background*, *Method*, etc. For the competition, however, we do not separate abstracts based on their structuring, rather we leave them interspersed in the training and test data. Nonetheless, we provide participants with the headings extracted from the structured abstracts to be used as a set of structural features.

In order to build classifiers, 800 annotated training abstracts were provided, and the goal was to automatically annotate 200 test abstracts with the relevant labels. Table A.1 shows the exact number of sentences and the percentages of the frequency of labels across the data set. We relied on "Kaggle in Class" to manage the submissions and rankings[1], and randomly divided the test data into "public" and "private" evalua-

[1] http://www.kaggle.com/

	All	Struct.	Unstruct.
Total			
- Abstracts	1,000	38.9%	61.1%
- Sentences	11,616	56.2%	43.8%
- Labels	12,211	55.9%	44.1%
% per label			
- Population	7.0%	5.6%	7.9%
- Intervention	5.9%	4.9%	6.6%
- Background	22.0%	10.3%	34.2%
- Outcome	38.9%	34.0%	40.9%
- Study Design	2.0%	2.3%	1.4%
- Other	29.2%	42.9%	9.0%

Table A.1: Statistics of the dataset. "% per label" refers to the percentage of sentences that contain the given label (the sum is higher than 100% because of multilabel sentences).

tion; the former was used to provide preliminary evaluations during the competition, and the latter to define the final classification of systems.

We provided two benchmark systems at the beginning of the competition. The first system is a simple frequency-based approach, and the second system is a variant of the state-of-the-art system presented by [Kim et al. 2011], using a machine learning algorithm for predictions.

A.2.1 Naive Baseline

For the naive baseline we merely rely on the most frequent label occurring in the training data, given the position of a sentence. For instance, for the first four sentences in the abstract the most frequent label is *Background*, for the fifth it is *Other*, etc.

A.2.2 Conditional Random Field (CRF) Benchmark

CRFs [Lafferty et al. 2001] were designed to label sequential data, and we chose this approach because it has shown success in sentence-level classification [Hirohata et al. 2008, Chung 2009, Kim et al. 2011]. Thus we tried to replicate the classifier used by [Kim et al. 2011]. However our systems differ in the selection of features used for training. We use lexical and structural features:

1. **Lexical features:** bag of words and Part Of Speech (POS) tags for the lexical features; and

2. **Structural features:** position of the sentences and the rhetorical headings from the structured abstracts. If a heading *h1* covered three lines in the abstract, all the three lines will be labeled as *h1*.

We used NLTK [Bird et al. 2009] to produce a list of POS tags and for the CRF classifier we utilized the Mallet [McCallum 2002] open source software.

Upon completion of the challenge we learned that our input to the CRF Benchmark did not have a separation between abstracts, causing Mallet to underperform. We rectified the training representation and obtained the accurate score which we refer to as CRF_corrected.

A.3 Evaluation

Previous work has relied on F-score for evaluating this task, but we decided to choose the *receiver operating characteristic* (ROC) curves and corresponding *area under curve* (AUC) value as the main metric. ROC curves plot the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. The AUC

Student Category	Open Category
Marco Lui	Macquarie Test
A_MQ	DPMCNA
System_Ict	Dalibor
	Starling
	Mix

Table A.2: Team names and categories.

score is the area under this plot, and the main benefit of this metric is that it allows us to compare classification outputs that assign probability distributions to labels, instead of a binary decision. We also provide F-scores for a better comparison with the existing literature.

Table A.2 shows the team names and the categories. There were two categories: "student" and "open". Members of the "student" category were exclusively students at any level: undergraduate or postgraduate. None of the members of the "student" category can hold a PhD in a relevant area. Members of the "open" category included those who could not participate in the "student" category. The winner of the student category and winner overall was Marco Lui from NICTA and the University of Melbourne, followed by Team A_MQ (Abeed Sarker) from Macquarie University and Team System_Ict (Spandana Gella and Duong Thanh Long) from the University of Melbourne. The top participants of the open category were Team Macquarie_Test (Diego Mollá, one of the task organisers) from Macquarie University, and Team DPMCNA (Daniel McNamara) from Australia National University and Kaggle. The description of the systems is provided in Section A.4.

Table A.3 shows the final scores obtained by the 8 participants and the baseline systems. The scores for private and public test data are very similar. We can see that the top system improved over our state-of-the-art baseline, and all the top-3 were close to its performance.

We relied on a non-parametric statistical significance test known as random shuffling [Yeh 2000] to better compare the F-scores of the participating systems and benchmarks. We present in Table A.5 the ranking of systems according to their F-scores, and the p-value when comparing each system with the one immediately below it in the table[2]. The p-values illustrate different clusters of performance, and they show that team "Marco Lui" significantly improves the CRF_corrected state-of-the-art benchmark, and that team "A_MQ" and CRF_corrected perform at the same level.

Table A.4 shows the F-scores separately for each class; the best scoring system is superior for most of the 6 classes. We observed that the ranking of the participants as measured by the official AUC score was the same for the top participants, but the ranking at the bottom of the list of participants differed. The *Outcome* and *Intervention* labels have the highest and lowest scores, respectively, which mostly correlates to the amount of available training instances for each.

A.4 Description of Systems

The top participants in the task kindly provided a short description of their architectures, which is given in the Appendix. All these submissions relied on Machine Learning (ML) methods, namely Support Vector Machines (SVM), Stacked Logistic Regression, Maximum Entropy, Random Forests, and CRF. Only one of the top participants relied on sequential classifiers (team "System_Ict" applied CRFs).

Two of the top systems (teams "Marco Lui" and "Macquarie Test") used a two-layered architecture, where features are learned through a first pass (supervised for "Marco Lui", unsupervised for "Macquarie Test"). Team "A_MQ" performed param-

[2]The p-value gives the probability of obtaining such an F-score difference between the compared systems assuming that the null hypothesis (that the systems are not significantly different from each other) holds.

System	Private Test	Public Test	F-score
Marco Lui	**0.96**	**0.97**	**0.82**
A_MQ	0.95	0.96	0.80
Macquarie Test	0.94	0.94	0.78
DPMCNA	0.92	0.93	0.71
System_Ict	0.92	0.93	0.73
Dalibor	0.86	0.92	0.73
Starling	0.86	0.87	0.78
Mix	0.83	0.84	0.74
Benchmarks			
- CRF_corrected	0.86	0.88	0.80
- CRF_official	0.80	0.83	0.70
- Naive	0.70	0.70	0.55

Table A.3: AUC and F-scores for public and private tests. The best results per column are given in bold.

eter optimisation separately for each of the PIBOSO categories, and it was the only team to use Metamap as a source of features. Feature selection was used by teams "Daniel McNamara" and "System_Ict", which also achieved high performances.

A.5 Conclusions

The third shared task aimed at fostering research on classifying medical sentences into the predefined PIBOSO category to aid the practice of EBM. Participants from Australia and world-wide competed on this task and the winning team obtained better results than state of the art where the difference was shown to be statistically significant. The best performing technique was attributed to the usage of the meta-learner feature stacking approach using three different sets of features.

We will endeavor to identify such important research problems and provide a forum for research students to provide their effective solutions in the forthcoming

System	Population	Intervention	Background	Outcome	Study Design	Other
Marco Lui	**0.58**	0.34	**0.80**	**0.89**	0.59	**0.85**
A_MQ	0.51	**0.35**	0.78	0.86	0.58	0.84
Macquarie Test	0.56	0.34	0.75	0.84	0.52	0.80
Starling	0.32	0.20	**0.80**	0.87	0.00	0.82
DPMCNA	0.28	0.12	0.70	0.78	0.48	0.73
Mix	0.45	0.19	0.68	0.82	0.40	0.81
System_Ict	0.30	0.15	0.68	0.84	0.35	0.83
Dalibor	0.30	0.15	0.68	0.84	0.40	0.83
Naive	0.00	0.00	0.59	0.68	0.00	0.15
CRF_official	0.33	0.22	0.55	0.78	0.67	0.81
CRF_corrected	**0.58**	0.18	**0.80**	0.86	**0.68**	0.83
Aggregate	0.38	0.21	0.71	0.83	0.42	0.76

Table A.4: F-scores across each individual label class and the aggregate. The best results per column are given in bold.

shared tasks.

A.6 Description of the top systems

The following text is by the team competitors who kindly agreed to send us their system descriptions.

Team Marco (Marco Lui)

A full description of this system is given in [Lui 2012]. We used a stacked logistic regression classifier with a variety of feature sets to attain the highest result. The stacking was carried out using a 10-fold cross-validation on the training data, generating a pseudo-distribution over class labels for each training instance for each feature set. These distribution vectors were concatenated to generate the full feature vector for each instance, which was used to train another logistic regression classifier. The test data was projected into the stacked vector space by logistic regression

System	F-score	p-value
Marco Lui	0.82	0.0012
CRF_corrected	0.80	0.482
A_MQ	0.80	0.03
Starling	0.78	0.3615
Macquarie Test	0.78	0.0001
Mix	0.74	0.1646
System_Ict	0.73	0.5028
Dalibor	0.73	0.0041
DPMCNA	0.71	0
Naive	0.55	-

Table A.5: Ranking of systems according to F-score, and pairwise statistical significance test between the target row and the one immediately below. The horizontal lines cluster systems according to statistically significant differences.

classifiers trained on each feature set over the entire training collection. No sequential learning algorithms were used; the sequential information is captured entirely in the features. The feature sets we used are an elaboration of the lexical, semantic, structural and sequential features described by Kim et al [Kim et al. 2011]. The key differences are: (1) we used part-of-speech (POS) features differently. Instead of POS-tagging individual terms, we represented a document as a sequence of POS-tags (as opposed to a sequence of words), and generated features based on POS-tag n-grams, (2) we added features to describe sentence length, both in absolute (number of bytes) and relative (bytes in sentence / bytes in abstract) terms, (3) we expanded the range of dependency features to cover bag-of-words (BOW) of not just preceding but also subsequent sentences, (4) we considered the distribution of preceding and subsequent POS-tag n-grams, (5) we considered the distribution of preceding and subsequent headings. We also did not investigate some of the techniques of Kim et al, including: (1) we did not use any external resources (e.g. MetaMap) to introduce additional semantic information, (2) we did not use rhetorical roles of headings for

structural information, (3) we did not use any direct dependency features.

Team A_MQ (Abeed Sarker)

In our approach, we divide the multi-class classification problem to several binary classification problems, and apply SVMs as the machine learning algorithm. Overall, we use six classifiers, one for each of the six PIBOSO categories. Each sentence, therefore, is classified by each of the six classifiers to indicate whether it belongs to a specific category or not. An advantage of using binary classifiers is that we can customise the features to each classification task. This means that if there are features that are particularly useful for identifying a specific class, we can use those features for the classification task involving that class, and leave them out if they are not useful for other classes. We use RBF kernels for each of our SVM classifiers, and optimise the parameters using 10-fold cross validations over the training data for each class. We use the MetaMap tool box to identify medical concepts (CUIs) and semantic types for all the medical terms in each sentence. We use the MedPost/SKR parts of speech tagger to annotate each word, and further pre-process the text by lowercasing, stemming and removing stopwords. For features, we use n-grams, sentence positions (absolute and relative), sentence lengths, section headings (if available), CUIs and semantic types for each medical concept, and previous sentence n-grams. For the outcome classification task, we use a class-specific feature called 'cue-word-count'. We use a set of key-words that have been shown to occur frequently with sentences representing outcomes, and, for each sentence, we use the number of occurrences of those key-words as a feature. Our experiments, on the training data, showed that such a class-specific feature can improve classifier performance for the associated class.

Team Macquarie Test (Diego Molla)

A full description of this system is given in [Molla 2012]. The system is the result of a series of experiments where we tested the impact of using cluster-based

features for the task of sentence classification in medical texts. The rationale is that, presumably, different types of medical texts will have specific types of distributions of sentence types. But since we don't know the document types, we cluster the documents according to their distribution of sentence types and use the resulting clusters as the document types. We first trained a classifier to obtain a first prediction of the sentence types. Then the documents were clustered based on the distribution of sentence types. The resulting cluster information, plus additional features, were used to train the final set of classifiers. Since a sentence may have multiple labels we used binary classifiers, one per sentence type. At the classification stage, the sentences were classified using the first set of classifiers. Then their documents were assigned the closest cluster, and this information was fed to the second set of classifiers. The submission with best results used Maxent classifiers, all classifiers used uni-gram features plus the normalised sentence position, and the second classifiers used, in addition, the cluster information. The number of clusters was 4.

Team DPMCNA (Daniel McNamara)

We got all of the rows in the training set with a 1 in the prediction column and treated each row as series of predictors and a class label corresponding to sentence type ('background', 'population', etc.) We performed pre-processing of the training and test sets using stemming, and removing case, punctuation and extra white space. We then calculated the training set mutual information of each 1-gram with respect to the class labels, recording the top 1000 features. For each sentence, We converted it into a feature vector where the entries were the frequencies of the top features, plus an entry for the sentence number. We then trained a Random Forest (using R's randomForest package with the default settings) using these features and class labels. We used the Random Forest to predict class probabilities for each test response variable. Note that We ignored the multi-label nature of the problem considering most sentences only had a single label.

Team System_Ict (Spandana Gella, Duong Thanh Long)

A full description of this system is given in [Gella and Long 2012]. Our top 5 sentence classifiers use Support Vector Machine (SVM) and Conditional Random Fields (CRFs) for learning algorithm. For SVM we have used libsvm 1 package and for CRF we used CRF++ 2 package. We used 10-fold cross validation to tweak and test the best suitable hyper parameters for our methods. We have observed that our systems performed very well when we do cross validation on train data but suffered over fitting. To avoid this we used train plus labelled test data with one of the best performing systems as our new training data. We observed that this has improved our results by approximately 3%. We trained our classifiers with different set of features which include lexical, structural and sequential features. Lexical features include collocational information, lemmatized bag-of-words features, part-of-speech information (we have used MedPost part-of-speech tagger) and dependency relations. Structural features include position of the sentence in the abstract, normalised sentence position, reverse sentence position, number of content words in the sentence, abstract section headings with and without modification as mentioned in [Kim et al. 2011]. Sequential features were implemented the same way as in [Kim et al. 2011] with the direct and indirect features. After having the pool of features from the above defined features, we perform feature selection to ensure that we always have the most informative features. We used the information gain algorithm from R system3 to do feature selection.

CHAPTER B

Appendix 2

This work was carried out in collaboration with two other authors (Diego Molla and David Martinez) and as one of the authors I have contributed in forming the idea and running the experiments. However, unlike the rest of this thesis I have not been the first author in this work. I would like to acknowledge that the writing of this appendix has been a collaboration between the three authors, mainly Diego Molla who has been the first author of this particular work taken from the following paper: Molla et al. [Mollá et al. 2013].

We propose a document distance-based approach to automatically expand the number of available relevance judgements when those are limited and reduced to only positive judgements. This may happen, for example, when the only available judgements are extracted from a list of references in a published clinical systematic review. We show that evaluations based on these expanded relevance judgements are more reliable than those using only the initially available judgements. We also show the impact of such an evaluation approach as the number of initial judgements

decreases.

B.1 Semi Automatic Relevance Judgement

There are applications that benefit from an information retrieval (IR) stage, but which do not have enough sample documents for a full assessment of the retrieval quality. Furthermore, the few sample documents available only represent positive relevant documents. For example, within the area of Evidence Based Medicine (EBM), clinical systematic reviews provide the medical doctor with clinical evidence together with a list of relevant documents. We envisage the development of tools that will facilitate the production of such systematic reviews. One of the first stages of such an application consists of an IR step that retrieves all key relevant documents. But the references in a systematic review cover only a small sample of all relevant references [Dickersin et al. 1994], and only a fraction of the documents of a systematic review can be retrieved after performing exhaustive searches, mostly due to the fact that there are complex queries and several document repositories [Martinez et al. 2008]. Furthermore, the list of references only indicate relevant documents but there are no lists of non-relevant documents readily available. It is therefore expected that any evaluation metric that is based solely on the references from the systematic review will show unreliable results.

Previous work has shown that by expanding an initial set of document assessments for given queries, one can perform a more accurate automatic evaluation of IR systems. For example, Büttcher et al. [Büttcher et al. 2007] used Machine Learning methods trained over a subset of relevance judgements in order to expand the set of relevance judgements. They showed that evaluation results with the expanded set of relevance judgements had better quality than using the source subset of judgements. Quality of the evaluation was measured by ranking a set of IR systems according to

the new expanded relevance judgements, and comparing it against the system ordering produced by the original set of judgements. In the clinical domain, Martinez et al. [Martinez et al. 2008] explored the use of re-ranking methods based on reduced judgements, and found that the use of automatic classifiers would allow to considerably reduce the time required for clinicians to identify a large portion (95%) of the relevant documents. Both these articles reported limitations of the classifiers when the initial number of documents was small. Furthermore, in the scenario that we contemplate, where we rely on the list of references of a systematic review as the set of relevant documents, we do not have information about negative judgements, and therefore a classifier-based approach to expand the set of relevant documents would have to deal with this issue.

More recent work [Sakai and Lin 2010] has shown that by relying on documents retrieved frequently by a diverse set of systems, it is possible to build relevance assessments automatically, and achieve high correlation with manually judged data. However this approach has been tested by building on a set of competing runs from different research groups, which is not always available; and this method does not benefit from existing qrels.

We propose to automatically expand the set of relevant documents by adding documents that are reasonably close to the original, reduced set. We show the result of several experiments that test the impact of such automatic expansion. For our experiments, we rely on the OHSUMED test collection [Hersh et al. 1994]. This is a corpus containing clinical queries and assessments, and we focus on the set of 63 queries that was used in the TREC-9 Filtering Track. The OHSUMED queries were generated to address actual information needs for clinicians, and the assessed documents were retrieved in two iterations, by relying on the MEDLINE search interface[1] and the SMART retrieval system respectively. The retrieved documents were judged

[1] http://www.ncbi.nlm.nih.gov/pubmed

by a separate group of domain experts to the group performing the search. As document collection we rely on the 1988-91 subset of MEDLINE that was released as test data for the TREC-9 challenge, which contains 293,856 documents. For evaluation we apply a variety of IR systems implemented in the Terrier open source package [Macdonald et al. 2012].

B.2 Distance versus Relevance

The rationale of our work is related to the so-called cluster hypothesis, that is, the assumption that "documents that are in the same cluster behave similarly with respect to relevance to information needs" [Manning 2008]. The cluster hypothesis has been used to improve the results of information retrieval and classification tasks. In contrast, we are not concerned about improving the IR results. Instead, we want to improve the effectiveness of IR evaluation. But this slightly reworded version of the cluster hypothesis may apply: documents *that are similar enough* will behave similarly with respect to relevance to information needs. The question is, how similar must these documents be?

We first examined the impact of similarity between documents with regards to their relevance. For every document associated to any qrel from the OHSUMED test set (3,121 documents), we computed the distance between the document and the closest qrel (other than the document itself) within each query. The resulting (document,question) pairs were sorted by distance and binned into centiles such that the first centile is formed by the top 1% pairs, and so on. Then, within each centile we computed the percentage of relevant documents. Figure C.1 shows the result.

The figure shows a clear relation between distance and relevance. 78% of documents in the first centile are relevant, and the number quickly degrades. The figure has been truncated to the top 10 centiles since virtually none of the documents from

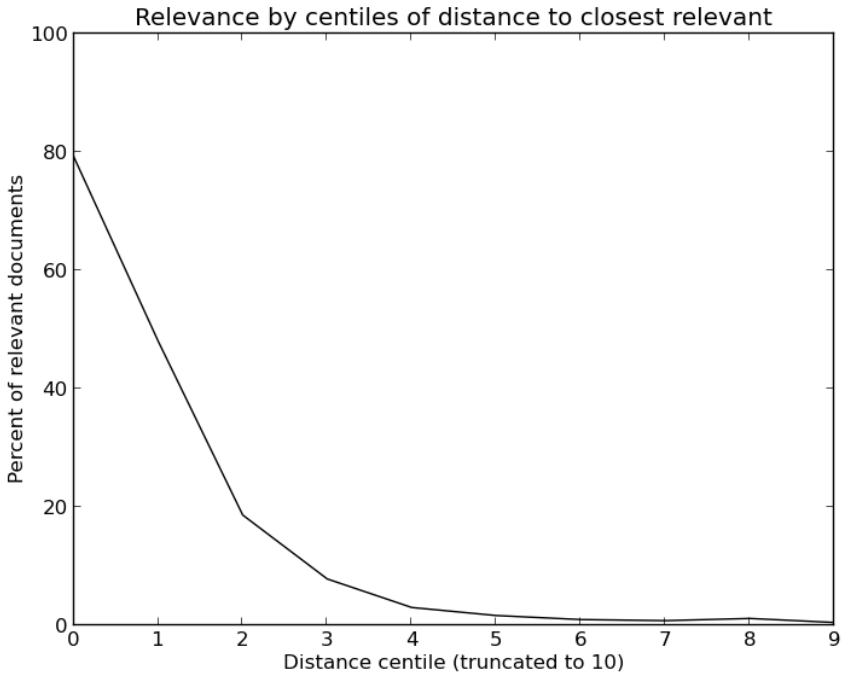


Figure B.1: Distance versus relevance in the OHSUMED test corpus.

the 10th centile onwards are relevant.

For these experiments we used as the distance metric $1 -$ cosine similarity. The vector representations of the documents were formed by obtaining the tfidf values of all words lowercased and with stop words removed, and then taking the top 200 components after performing Principal Component Analysis (PCA).[2] PCA was used as a means to compress the vector space. Using tfidf features without the subsequent PCA stage produced a slightly less marked relation and at the expense of much longer

[2]These experiments were carried out in Python and the scikit-learn library.

processing times.

B.3 Evaluation metrics

The results described in Section C.4 show that distance between a document and a known relevant document may be a good indicator of document relevance. We therefore studied the impact of using document distance as a means to generate new relevance sets which we call pseudo-qrels.

In the following experiments we used several IR systems as described below. We evaluated the performance of each system according to these sets of relevance judgements:

1. Original set of qrels.
2. A subset of qrels. This subset is a baseline that models the situation where the number of qrels is limited.
3. The same subset of qrels, expanded with the pseudo-qrels. These pseudo-qrels are produced based on distance metrics as described below.

B.3.1 Information Retrieval Baselines

In the absence of the official set of runs from the TREC filtering track participants, we resorted to building our own systems using the open source Terrier 3.5 package. We built 16 baselines by choosing 16 different ranking algorithms and used them with their default settings to build the runs.

Terrier offers a range of Divergence from Random (DFR) models which are instantiated by three components of the framework: selecting a basic randomness model, applying the first normalisation, and normalising the term frequencies. We

BB2	BM25	DFR_BM25	DLH
DPH	DFRee	Hiemstra_LM	DLH13
IFB2	In_expB2	In_expC2	InL2
LemurTF_IDF	LGD	PL2	TF_IDF

Table B.1: List of 16 runs from the terrier package

stopped and stemmed all the 63 test queries and the collection, and used the Porter stemmer as the default stemming algorithm. Table B.1 is the list of ranking models corresponding with the baselines used for our experiments.

B.3.2 Pseudo-qrels for Evaluation

The pseudo-qrels of a query are generated by selecting those that are closest to some qrel within the query, using the $1 -$ cosine distance metric described in Section C.4:

1. For every query q:
    a) For every document d in the pool of available documents:
        i. Record the minimum distance between d and the set of qrels within q (except d).
2. Sort the resulting triples (distance,d,q) in ascending order and select the top K.
3. Add the selected documents d to the corresponding q. These are the pseudo-qrels.

For these experiments, the pool of available documents was generated by taking the top N documents retrieved by each query for all of the IR systems that we used. We varied the percentage of available qrels in our experiments, always making sure that each query had at least one qrel.

Note that the above algorithm selects the candidate pseudo-qrels using a threshold that is global to all queries. This means that some queries may receive more pseudo-qrels than others, and in extreme cases only a few queries will receive pseudo-qrels. We thought that this is desirable, since the experiment in Section C.4 shows such a strong impact of document distance in the relevance of the document. If a query only has documents that are relatively far from known qrels, we better not add them as pseudo-qrels.

The approach described above can be seen as a simple one-rule classifier based on distance. We resorted to such a simple classifier instead of a more sophisticated classifier such as the SVM classifier used by Büttcher et al. [Büttcher et al. 2007] because of the scarcity of data and lack of negative judgements in our scenario. If we were to train an SVM classifier we would need to find a means to reduce overfitting.

For a first estimation of the quality of the retrieved qrels, we evaluated our method in the manner of a text classification system, by relying on different splits of the original qrels, and measuring the F-score value for the detection of relevant documents for each query. For this experiment we use different partitions as "training" data (which is used for the documents in the collection to compare against) and "test" data (which is used for evaluation).

In order to retrieve pseudo-qrels, we set $N = 100$ (we retrieved the top-100 documents for each of our IR systems), and $K = 0.2\%$ (we considered as relevant the top 0.2% of the most similar documents). The evaluation of the pseudo-qrels is given in Table B.2, when using up to 50% of the qrels as training data. We can see that overall the performances are low, specially in recall, but Büttcher et al. [Büttcher et al. 2007] found that low F-scores can still lead to large improvements when measuring the correlation between manual and semi-automatic relevance judgements. The results also illustrate that when using only 20% of relevant documents, we achieve the highest F-score, and more than a third of the retrieved documents are relevant.

Train Qrels	Test Qrels	Precision	Recall	F-score
10%	90%	0.360	0.100	0.157
20%	80%	0.345	0.118	0.176
30%	70%	0.290	0.112	0.161
40%	60%	0.282	0.125	0.173
50%	50%	0.244	0.123	0.164

Table B.2: Retrieved pseudo-qrels evaluated against the original relevance set.

B.3.3 Correlation for ranking IR systems

Figure B.2 shows Kendall's tau between the ranking of the IR systems when evaluated using 1) a baseline consisting of original qrels, and 2) varying percentages of qrels extended with the computed pseudo-qrels. The evaluation metric was MAP. The figure presents the results for varying values of N (the number of documents taken from each query in each IR system), and K (the percentage of top documents selected as pseudo-qrels).

The baseline shown in the figure uses the qrels without the pseudo-qrels and it reflects the quality of the evaluation when using the available data. We can observe, as expected, that larger percentages of qrels lead to better correlation figures. The other curves show the evaluation quality when the qrels are expanded with pseudo-qrels.

The figure shows that different choices of values of N and K affect the quality of the evaluation. When we choose relatively few documents (N = 30) to form the pool of available documents, the results do not improve on the baseline. This is presumably due to the lack of enough documents to gather useful statistics. When we choose a larger number of documents (N = 100), then a wise threshold K may lead to improvements. In our experiments, choosing a relatively small percentage of doc-

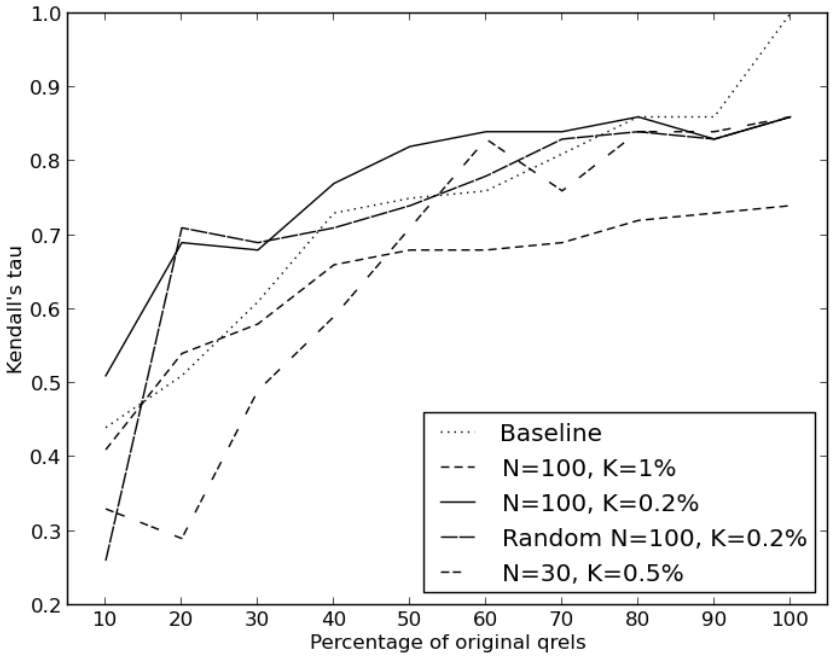


Figure B.2: Kendall's tau of system orderings using MAP. The baseline uses percentages of the original qrels. The other evaluations use percentages of the qrels plus the pseudo-qrels for several choices of N (number of documents chosen per query) and K (percentage of top documents selected as qrels).

uments from the pool (K = 0.2%) leads to results above the baseline, but choosing a larger percentage (K = 1%) leads to a decline of results. This is in line with the analysis shown in Figure C.1, which indicates that the percentage of relevant documents decreases steeply as we increase the distance. Therefore a threshold which is too relaxed may introduce too much noise. With a choice of N=100 and K=0.2%, small percentages of qrels lead to a comparatively greater improvement over the baseline.

These results are very encouraging and support the idea of using distance metrics to compensate for the lack of available relevance judgements and the lack of negative relevance judgements.

When selecting the qrels, all the results described above used the top qrels (those appearing first in the list of qrels). Figure B.2 also includes the results when using a random selection of qrels. It shows wide changes for small percentages of qrels, and it tends to agree with the baseline for larger percentages. This probably means that the choice of qrels really matters, and documents from the top qrels may be quality relevant documents. In future work we will study the impact of the selection of qrels further.

Figure B.3 shows the system map scores using the official qrel combined with the pseudo-qrels for three varying sizes of limited positive qrels. As a reference we also show the curve when using 20% of qrels only. It can be seen that the scores generated by the pseudo-qrels range in the vicinity of the official map scores for qrel = 80%, while the results with the lower percentage of true qrels tend to be underestimated. However, the ordering of the runs which was our ultimate goal, remains stable. When using 20% of qrels, the curves with and without pseudo-qrels look similar, but lead to different rankings, as the Kendall's tau scores in Figure B.2 illustrate.

B.4 Conclusions

We have shown promising results towards the use of a simple distance-based approach to expand a set of relevance judgements. The results are particularly encouraging when the number of available relevance judgements is very limited, and works when there are only positive judgements.

These results suggest the use of distance-metrics extensions of relevance judgements as a quick and cheap evaluation during the development stage of information

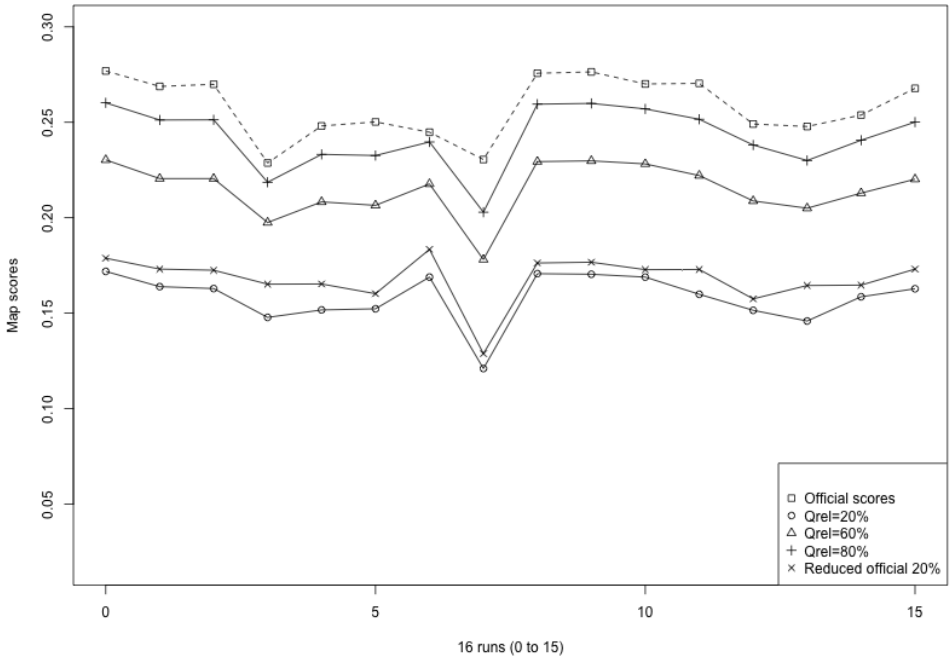


Figure B.3: Official map scores using official qrel versus limited true qrels combined with the pseudo-qrels for N = 100 and K = 0.2

retrieval systems when there are few and only positive relevance judgements. It can therefore be applied for the development of IR systems that search for relevant clinical studies, even when the set of known available relevant documents is just the list of references of a sample clinical systematic review.

Further work includes a more comprehensive study of the thresholds that lead to the best evaluation setting. It is also desirable to determine how well these findings carry to other domains. Also, given that the measure of quality used in this

study is based on the correlation of rankings with an automated evaluation metric, it is desirable to extend this study with real human judgements for a more precise characterisation of the possibilities of this approach.

We have used a very simple distance metric in this study. It will be interesting to explore the impact of additional distance metrics that may use domain knowledge or more sophisticated linguistic information.

CHAPTER C

Appendix 3

This work was carried out in collaboration with two other authors (Diego Molla and David Martinez) and as one of the authors I have contributed in forming the idea and running the experiments. However, unlike the rest of this thesis I have not been the first author in this work. I would like to acknowledge that the writing of this appendix has been a collaboration between the three authors, mainly Diego Molla who has been the first author of this particular work taken from the following paper: Molla et al. [Mollá et al. 2014].

This work reports the use of a document distance-based approach to automatically expand the number of available relevance judgements when these are limited and reduced to only positive judgements. This may happen, for example, when the only available judgements are extracted from a list of references in a published review paper. We compare the results on two document sets: OHSUMED, based on medical research publications, and TREC-8, based on news feeds. We show that evaluations based on these expanded relevance judgements are more reliable than

those using only the initially available judgements, especially when the number of available judgements is very limited.

C.1 Semi Automatic Relevance Judgement: Document Distance

An important bottleneck in the development of information retrieval (IR) systems is their evaluation. Generating human-produced judgements is expensive and time-consuming, and it is not always possible to produce a large set of relevance judgements (qrels henceforth).

We envisage a scenario where the only available qrels are the list of references of a survey paper. For example, within the area of Evidence Based Medicine (EBM), clinical systematic reviews provide the key published evidence that is relevant to a specific clinical query, together with a list of references that backs up the clinical evidence. This list of references, however, covers only a small sample of all relevant references [Dickersin et al. 1994]. Furthermore, only a fraction of the documents of a systematic review can be retrieved after performing exhaustive searches, mostly due to the fact that there are complex queries and several document repositories [Martinez et al. 2008]. Another problem with using the list of references as the only qrels is that negative qrels, that is, judgements about non-relevant documents, are not included. Any attempts to develop IR systems for such a scenario will need to supplement the list of references with something else. In this research we propose to automatically expand the qrels by finding similar documents.

C.2 Related Work

Using document distance as a criterion to expand a list of qrels sounds intuitive. The approach is related to the well-known cluster hypothesis: “closely associated docu-

ments tend to be relevant to the same requests" [Rijsbergen 1979]. This hypothesis has been typically used to improve the quality of the retrieval of documents but there is very limited past work using the cluster hypothesis to improve the quality of the evaluation.

Previous work on the expansion of an initial set of document assessments include the use of Machine Learning. For example, Büttcher et al. [Büttcher et al. 2007] trained over a subset of qrels in order to expand the set of qrels. They showed that evaluation results with the expanded set of qrels had better quality than using the source subset of qrels. Quality of the evaluation was measured by ranking a set of IR systems according to the new expanded qrels, and comparing it against the system ordering produced by the original qrels. In the clinical domain, Martinez et al. [Martinez et al. 2008] explored the use of re-ranking methods based on reduced judgements, and found that the use of automatic classifiers would allow to considerably reduce the time required for clinicians to identify a large portion (95%) of the relevant documents. Both of these articles reported limitations of the classifiers when the initial number of documents was small. Furthermore, in the scenario that we contemplate, where we rely on the list of references of a systematic review as the set of qrels, we do not have information about negative qrels, and therefore a classifier-based approach to expand the set of relevant documents would have to deal with this issue.

More recent work [Sakai and Lin 2010] has shown that by relying on documents retrieved frequently by a diverse set of systems, it is possible to build relevance assessments automatically, and achieve high correlation with manually judged data. However this approach has been tested by building on a set of competing runs from different research groups, which is not always available; and this method does not benefit from existing qrels.

Prior work using document distance criteria for expanding the qrels includes [Mollá et al.

2013], who suggests that this approach may work for a document collection within the medical domain. In this research we show that this approach improves the quality of evaluation *both* for medical and news reports, and we therefore add further evidence of the plausibility of this method.

Our work complements that of related work on the study of the impact of the number of topics and relevance judgements in IR evaluation [Carterette and Smucker 2007].

C.3 Data Sets

We use the OHSUMED collection of medical research publications, and the TREC-8 collection of news feeds.

The OHSUMED collection [Hersh et al. 1994] is a corpus containing clinical queries and assessments. We focus on the set of 63 queries that was used in the TREC-9 Filtering Track. The OHSUMED queries were generated to address actual information needs for clinicians, and the assessed documents were retrieved in two iterations, by relying on the MEDLINE search interface[1] and the SMART retrieval system respectively. The retrieved documents were judged by a separate group of domain experts to the group performing the search. As document collection we rely on the 1988-91 subset of MEDLINE that was released as test data for the TREC-9 challenge, which contains 293,856 documents. The judgement set has an average of 50.87 judgements per query, all of them positive. Since the original runs of the systems participating in the TREC-9 challenge are not available, for evaluation we created 16 IR systems implemented with the Terrier 3.5 open source package [Macdonald et al. 2012]. Table C.1 lists the settings of the Terrier package used for our runs, which are the same settings used by [Mollá et al. 2013].

[1] http://www.ncbi.nlm.nih.gov/pubmed

BB2	BM25	DFR_BM25	DLH
DPH	DFRee	Hiemstra_LM	DLH13
IFB2	In_expB2	In_expC2	InL2
LemurTF_IDF	LGD	PL2	TF_IDF

Table C.1: List of 16 runs from the terrier package

Each document of the OHSUMED collection contains bibliographical data (title, authors, etc) plus the abstract. For the experiments reported in this research we used only the contents of the abstract.

The TREC-8 collection [Voorhees and Harman 2001] comprises disks 4 and 5 of the TREC collection, excluding the *Congressional Record* subcollection. We used the test set, which has 50 queries with an average of 1,736 qrels per query. Of these, since we want to model a scenario where only positive judgements are used, we use only the positive qrels, which average 94.56 positive qrels per query. The qrels were generated using the pooling method, taking the top 100 documents retrieved by the systems participating in the *ad-hoc* task of TREC-8. For evaluation we used the results of the original systems that participated in the *ad-hoc* track of TREC-8.

Each document of the TREC-8 collection contains various XML markups. Given that each of the multiple sources had a different XML tag set, for the experiments reported in this research simply we ignored all lines that had an XML markup. The remaining lines consisted mostly of the main text, but there were still a few lines left that had meta-data.

C.4 Distance versus Relevance

We first examined the relation between similarity between qrel candidates, and their relevance. We obtained the candidates by pooling, as explained below for each dataset. For every query and for every qrel candidate in the query, we computed

the minimum distance between the qrel candidate and a known positive qrel for the query. The resulting (qrel candidate, query) pairs were sorted by distance and binned into deciles such that the first decile is formed by the top 10% pairs, and so on. Then, within each decile we computed the percentage of qrel candidates that were actually positive qrels. Since the OHSUMED data only had positive qrels, for each query we built the list of qrel candidates by pooling the top 100 documents per run. There was an average of 202.80 qrel candidates per query (12,371 qrel candidates in total[2]), and those that were not in the list of known qrels were tagged as negative judgements. For the TREC data, we used the qrels provided by the organisers of TREC. These qrels had been obtained by pooling the top 100 documents per run and contained positive and negative judgements, with an average of 1,736.60 qrels per query (86,830 qrels in total). Due to time and memory constraints we have used the first 100 qrels of each query, giving a total of 5,000 qrel candidates.

Figure C.1 shows the result. The figure shows a clear relation between distance and relevance in both datasets. The relation is not as marked as reported by [Mollá et al. 2013] but, as we will show below, it is sufficient to give an improvement in the evaluation when we expand the original qrels. The reason why the results differ from those of prior work is that the pool of documents in prior work was taken from the global list of known qrels, instead of from the runs of the systems. Our pooling method reflects a more realistic scenario and makes it possible to compare the OHSUMED and the TREC datasets. We observe that, in general, the percentage of relevant candidates drops much quicker in the TREC data than in the OHSUMED data.

For the experiments we used as the distance metric $d(x, y) = 1 - \cos(x, y)$ where $\cos(x, y)$ is the cosine similarity. The vector representations were formed by

[2]Note that the total number of qrels is slightly lower than 63*202.80=12,777 due to the existence of qrels shared among questions.

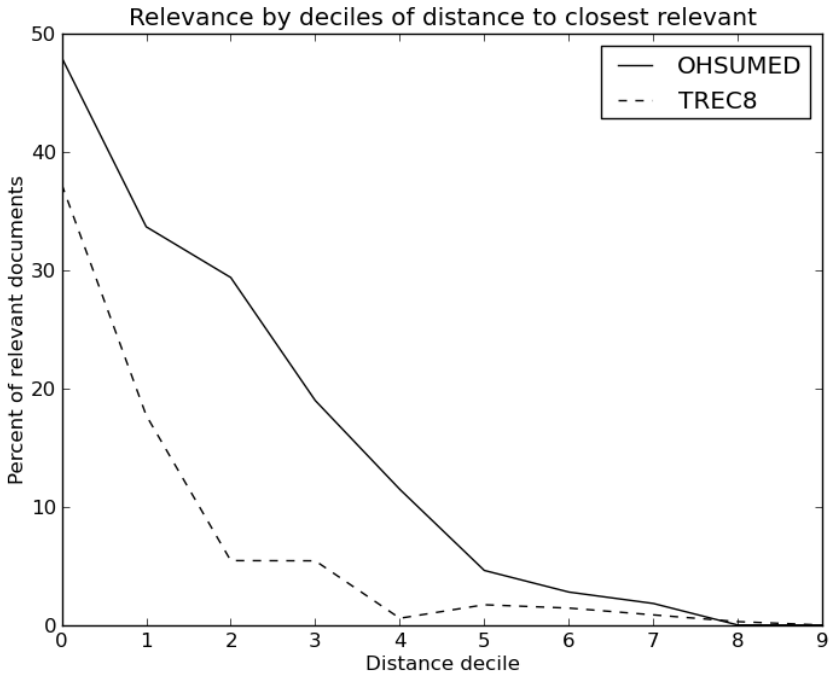


Figure C.1: Distance versus relevance in the OHSUMED and TREC-8 test datasets.

obtaining the *tf.idf* values of all words after lowercasing and removing stop words, and then taking the top 200 components after performing Principal Component Analysis (PCA).[3] These are the same settings as described by [Mollá et al. 2013].

[3]These experiments were carried out in Python and the scikit-learn library.

C.4.1 Pseudo-qrels for Evaluation

We expand the original qrels by introducing qrel candidates that are close enough to a known positive qrel. The specific process to rank the candidates is the same as described in Section C.4. We then apply a percentile threshold to select the pseudo-qrels. In other words, given the list of pairs (qrel candidate, query) sorted by distance to the closest positive qrel of the query, we select the top $K\%$ qrel candidates. We will call these added qrel candidates pseudo-qrels.

The process to find the pseudo-qrels uses a threshold that is global to all queries. This means that some queries may receive more pseudo-qrels than others, and a query may receive no pseudo-qrels. As we reduce the threshold, we will find more cases where a query has no additional pseudo-qrels. We thought that using a global threshold is desirable, since if a query only has documents that are relatively far from known qrels, we better not add them as pseudo-qrels.

To test the impact of the number of available qrels, in our experiments we have varied the number of qrels per query, always making sure that each query had at least one qrel. The selected qrels were drawn randomly from the original set of qrels, using the same random seed in all experiments.

C.4.2 Correlation for ranking IR systems

To determine the quality of the pseudo-qrels, and keeping in mind the scenario envisaged at the introduction, we evaluate and rank the set of runs using the qrels plus pseudo-qrels. The evaluation metric was MAP. We then compare the ranking of systems against another evaluation where we use the complete set of qrels. The system rankings are compared using Kendall's tau.

We conducted several experiments by varying the percentages of qrels extended with the computed pseudo-qrels. We also included a baseline that does not include

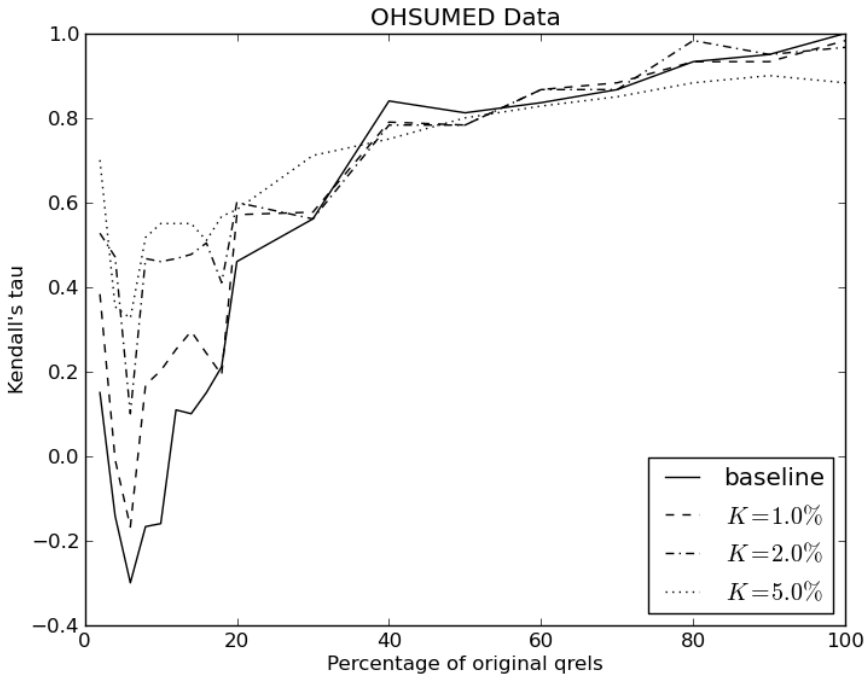


Figure C.2: Kendall's tau of system orderings on the OHSUMED data

the additional pseudo-qrels. The baseline simulates the default case when we only use the available qrels.

Figure C.2 shows the results for the OHSUMED dataset, and Figure C.3 shows the results for the TREC dataset. The figures present the results for varying values of K (the percentage of top documents selected as pseudo-qrels). We can observe, as expected, that larger percentages of qrels lead to higher correlation.

In both cases, we observe a gain of Kendall's tau for small percentages K of the original qrels. The gain is higher in the OHSUMED than the TREC dataset.

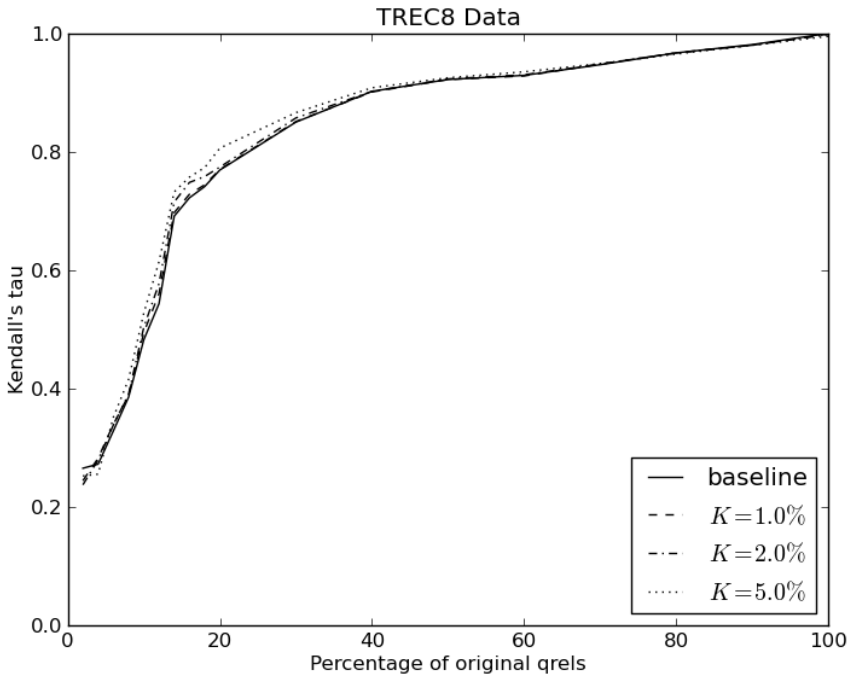


Figure C.3: Kendall's tau of system orderings on the TREC data

Figure C.4 zooms on the lower values of K for the TREC data. We appreciate a greater gain in some of the smaller values of K. Critically, these values represent an original number of qrels that is similar to those encountered in our envisaged scenario.

We observed that selecting a different subset of qrels influences the resulting tau, especially for the smaller percentages of qrels. We tried with several baselines by using different random seeds to select the qrels, and compared them with the expanded versions with the pseudo-qrels. The gain of adding pseudo-qrels varied

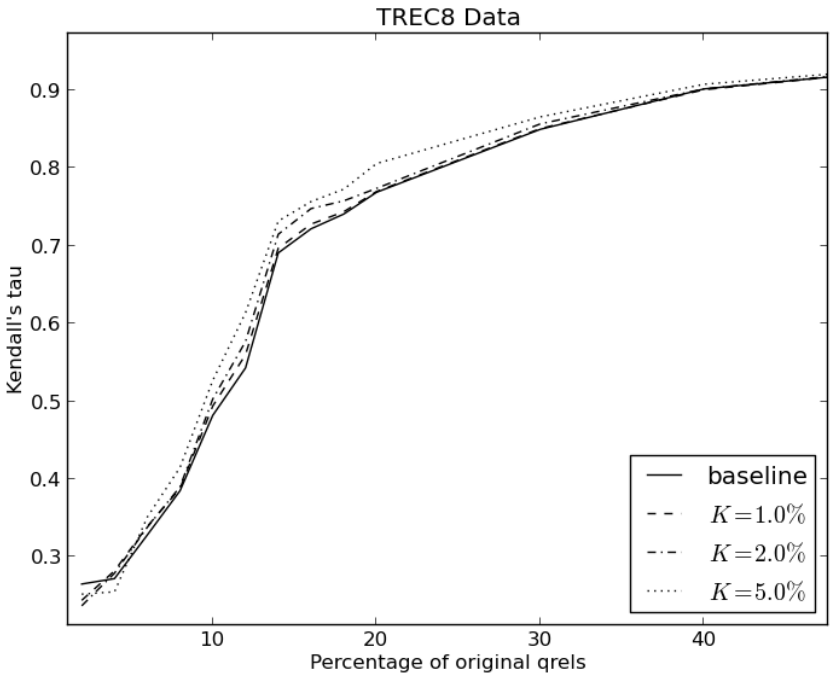


Figure C.4: Kendall's tau of system orderings focusing on the smaller percentages of the TREC data

depending on the initial choice of qrels, but in general there was a gain. Figure C.5 illustrates the impact of using different initial qrels for the TREC dataset.

C.5 Conclusions

We have compared the use of document similarity scores in two datasets, with the aim to compensate for the limited availability of qrels. The advantage of our approach

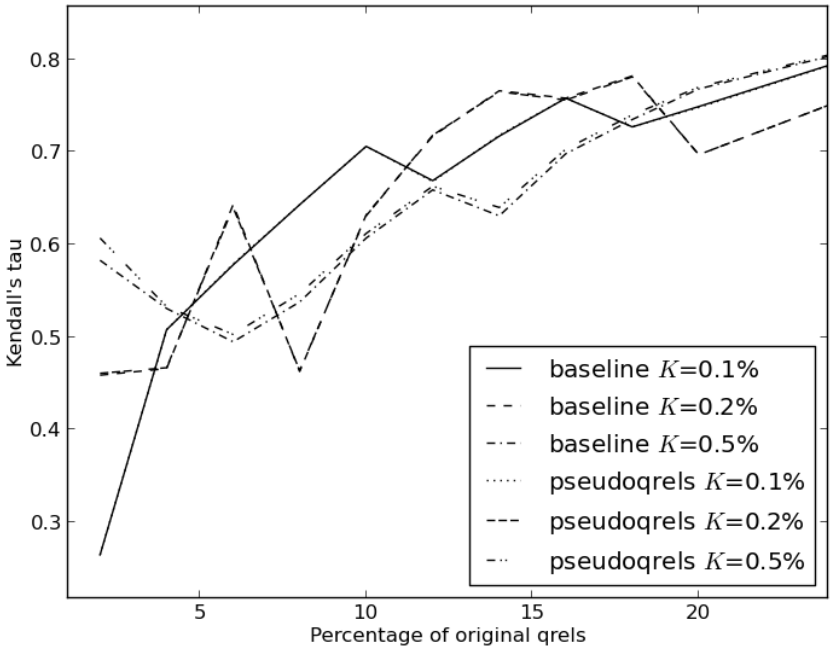


Figure C.5: Impact of using different initial qrels. In all cases, adding pseudo-qrels improved the results or remained practically the same.

against classification-based approaches such as those of prior work is that our method is applicable even when there are only positive relevance judgements.

The results are particularly encouraging when the number of available relevance judgements is very limited, and they suggest the use of distance-metrics extensions of relevance judgements as a quick and cheap evaluation step during the development stage of information retrieval systems when there are few and only positive relevance judgements. It can therefore be applied for the development of IR systems that search

for relevant clinical studies, even when the set of known available relevant documents is just the list of references of a sample clinical systematic review.

Further work includes a more comprehensive study of the thresholds that lead to the best evaluation setting, and the use of variants of distance metrics, other than straight cosine distance over a bag-of-words vector space model. Also, given that the measure of quality used in this study is based on the correlation of rankings with an automated evaluation metric, it is desirable to extend this study with real human judgements.

Finally, note that the present study expands the available qrels with positive judgements only. A further interesting line of research will include the automatic addition of negative judgements.

Bibliography

PewResearch, 2012. http://www.pewinternet.org/fact-sheets/health-fact-sheet/.

S. Abdou and J. Savoy. Searching in medline: Query expansion and manual indexing evaluation. *Information Processing & Management*, 44(2):781–789, 2008.

G. Amati. *Probabilistic Models for Information Retrieval Based on Divergence from Randomness*. PhD thesis, University of Glasgow, 2003.

G. Amati and C. Van Rijsbergen. Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

I. Amini, M. Sanderson, D. Martinez, and X. Li. Search for Clinical Records: RMIT at TREC 2011 Medical Track. In *Proceedings of Text Retrieval Conference*, 2011.

I. Amini, D. Martinez, and D. Molla. Overview of the ALTA 2012 Shared Task. In *Proceedings of ALTA 2012*, volume 7, pages 7–9, 2012a.

I. Amini, M. Sanderson, D. Martinez, and X. Li. Using Meta-data to search for Clinical Records: RMIT at TREC 2012 Medical Track. In *Proceedings of Text Retrieval Conference*, 2012b.

I. Amini, D. Martinez, X. Li, and M. Sanderson. Improving patient record search: A meta-data based approach. *Information Processing & Management*, 2015.

E. Apostolova, D. S. Channin, D. Demner-Fushman, J. Furst, S. Lytinen, and D. Raicu. Automatic segmentation of clinical texts. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 5905–5908. IEEE, 2009.

A. Aronson. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: the MetaMap Program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

A. R. Aronson and F.-M. Lang. An Overview of MetaMap: Historical Perspective and Recent Advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–36, 2010. ISSN 1527-974X. doi: 10.1136/jamia.2009.002733.

A. R. Aronson, O. Bodenreider, D. Demner-fushman, K. W. Fung, V. K. Lee, J. G. Mork, A. Névéol, L. Peters, and W. J. Rogers. From Indexing the Biomedical Literature to Coding Clinical Text : Experience with MTI and Machine Learning Approaches, 2007.

S. R. Beach. *Family problems and family violence: Reliable assessment and the ICD-11*. Springer Publishing Company, 2012.

S. Bedrick, T. Edinger, A. Cohen, and W. Hersh. Identifying Patients for Clinical Studies from Electronic Health Records: TREC 2012 Medical Records Track at OHSU. In *Proceedings of Text Retrieval Conference*, 2012.

E. V. Bernstam, J. R. Herskovic, Y. Aphinyanaphongs, C. F. Aliferis, M. G. Sriram, and W. R. Hersh. Using citation data to improve retrieval from medline. *Journal of the American Medical Informatics Association*, 13(1):96–105, 2006.

Bethesda. *UMLS Reference Manual [Internet]*. http://www.ncbi.nlm.nih.gov/books/NBK9679/, 2009.

S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

F. Boudin, J.-Y. Nie, and M. Dawes. Clinical information retrieval using document and pico structure. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 822–830. Association for Computational Linguistics, 2010.

C. Buckley. Why Current IR Engines Fail. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 584–585. ACM, 2004.

C. Buckley and E. Voorhees. Retrieval Evaluation with Incomplete Information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32. ACM, 2004.

S. Büttcher, C. L. Clarke, and G. V. Cormack. Domain-Specific Synonym Expansion and Validation for Biomedical Information Retrieval (MultiText Experiments for TREC 2004). In *Proceedings of Text REtrieval Conference (TREC)*, 2004.

S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. Reliable Information Retrieval Evaluation with Incomplete and Biased Judgements. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, page 63, New York, New York, USA, 2007. ISBN 9781595935977. doi: 10.1145/1277741.1277755.

J. Callan. Distributed information retrieval. *Advances in information retrieval*, pages 127–150, 2002.

B. Carterette and M. D. Smucker. Hypothesis testing with incomplete relevance judgments. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 643–652. ACM, 2007.

W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001a.

W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. Evaluation of Negation Phrases in Narrative Clinical Reports. In *Proceedings of the AMIA Symposium*, page 105. American Medical Informatics Association, 2001b.

D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, and D. S. Wishart. Polysearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic acids research*, 36 (suppl 2):W399–W405, 2008.

G. Chung. Sentence retrieval for abstracts of randomized controlled trials. *BMC Med Inform Decis Mak*, 9:10, 2009. doi: 10.1186/1472-6947-9-10.

C. W. Cleverdon. The significance of the cranfield tests on index languages. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. ACM, 1991.

J. Cogley, N. Stokes, J. Dunnion, and J. Carthy. UCD IIRG at TREC 2011 Medical Track. In *The Twentieth Text REtrieval Conference (TREC-20)*, 2011.

A. M. Cohen, W. R. Hersh, C. Dubay, and K. Spackman. Using co-occurrence network structure to extract synonymous gene and protein names from medline abstracts. *BMC bioinformatics*, 6(1):103, 2005.

M. Daoud, D. Kasperowicz, J. Miao, and J. Huang. York University at TREC 2011 : Medical Records Track. In *The Twentieth Text REtrieval Conference (TREC-20)*, 2011.

D. Demner-Fushman and J. Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103, 2007.

D. Demner-Fushman, S. Abhyankar, A. Jimeno-Yepes, R. Loane, B. Rance, F. Lang, N. Ide, E. Apostolova, and A. Aronson. A Knowledge-based Approach to Medical Records Retrieval. In *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*, 2011.

J. C. Denny, J. D. Smithers, R. A. Miller, and A. Spickard. Understanding medical school curriculum content using KnowledgeMap. *Journal of the American Medical Informatics Association*, 10(4):351–362, 2003.

A. Diaz, M. Ballesteros, J. Carrillo-de Albornoz, and L. Plaza. UCM at TREC-2012: Does negation influence the retrieval of medical reports? In *Proceedings of Text REtrieval Conference (TREC)*, 2012.

K. Dickersin, R. Scherer, and C. Lefebvre. Identifying Relevant Studies for Systematic Reviews. *BMJ (Clinical research ed.)*, 309(6964):1286–91, 1994. ISSN 0959-8138.

D. Dinh and L. Tamine. IRIT at TREC 2011: Evaluation of Query Expansion Techniques for Medical Record Retrieval. In *Proceedings of TREC*, 2011.

S. Gella and D. T. Long. Automatic sentence classifier for event based medicine: Shared task system description. In *Australasian Language Technology Workshop 2012 : ALTA Shared Task*, 2012.

L. Goeuriot, G. J. F. Jones, L. Kelly, J. Leveling, A. Hanbury, M. Henning, S. Salanter, and G. Zuccon. ShARe / CLEF eHealth Evaluation Lab 2013 , Task 3 : Information Retrieval to Address Patients Questions when Reading Clinical Reports. In *Online Working Notes of CLEF, CLEF (2013)*, pages 1–16, 2013.

L. Goeuriot, L. Kelly, G. J. Jones, H. Müller, and J. Zobel. Report on the SIGIR 2014 workshop on medical information retrieval (MedIR). In *ACM SIGIR Forum*, volume 48, pages 78–82. ACM, 2014a.

L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, G. Jones, and H. Mueller. ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. In *Proceedings of CLEF 2014*, pages 43–61, 2014b.

L. Goeuriot, G. J. Jones, L. Kelly, H. Müller, and J. Zobel. Medical information retrieval: introduction to the special issue. *Inf. Retr. Journal*, 19(1-2):1–5, 2016.

T. Goodwin, B. Rink, K. Roberts, and S. Harabagiu. Cohort Shepherd: Discovering Cohort Traits from Hospital Visits. In *Proceedings of TREC*, 2011.

M. D. Gordon and R. K. Lindsay. Toward discovery support systems: A replication, re-examination, and extension of swanson's work on literature-based discovery of a connection between raynaud's and fish oil. *Journal of the American Society for Information Science*, 47(2):116–128, 1996.

H. Gurulingappa, B. Müller, M. Hofmann-Apitius, and J. Fluck. A Semantic Platform for Information Retrieval from E-Health Records. In *Proceedings of TREC*, 2011.

D. Hanisch, K. Fundel, H. Mevissen, R. Zimmer, and J. Fluck. ProMiner: Rule-based Protein and Gene Entity Recognition. *BMC bioinformatics*, 6(Suppl 1):S14, 2005.

D. Harman and C. Buckley. The NRRC Reliable Information Access (RIA) Workshop. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 528–529. ACM, 2004.

V. Harmandas, M. Sanderson, and M. Dunlop. Image Retrieval by Hypertext Links. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 296–303. ACM, 1997.

R. B. Haynes, P. J. Devereaux, and G. H. Guyatt. Physicians' and Patients' Choices in Evidence Based Practice: Evidence Does Not Make Decisions, People Do. *British Medical Journal*, 324(7350):1350–1351, 2002.

W. Hersh, C. Buckley, T. Leone, and D. Hickam. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In *SIGIR94*, pages 192–201. Springer, 1994.

W. Hersh, S. Price, and L. Donohoe. Assessing thesaurus-based query expansion using the umls metathesaurus. In *Proceedings of the AMIA Symposium*, page 344. American Medical Informatics Association, 2000.

W. Hersh, A. Cohen, P. Roberts, and H. Rekapalli. TREC 2006 Genomics Track Overview. In *The Fifteenth Text Retrieval Conference*, pages 52–78, 2006.

W. Hersh, A. Cohen, P. Roberts, and H. Rekapalli. TREC 2007 Genomics Track Overview. In *The Sixteenth Text Retrieval Conference*, 2007.

W. R. Hersh. Report on the TREC 2004 genomics track. In *ACM SIGIR Forum*, volume 39, pages 21–24. ACM, 2005.

W. R. Hersh and R. T. Bhupatiraju. Trec genomics track overview. In *TREC*, volume 2003, pages 14–23, 2003.

K. Hirohata, N. Okazaki, S. Ananiadou, and M. Ishizuka. Identifying sections in scientific abstracts using conditional random fields. In *Proc. of 3rd International Joint Conference on Natural Language Processing*, pages 381–388, 2008.

K.-C. Huang, C. C.-H. Liu, S.-S. Yang, C.-C. Liao, F. Xiao, J.-M. Wong, and I.-J. Chiang. Classification of pico elements by text features systematically extracted from pubmed abstracts. In *Granular Computing (GrC), 2011 IEEE International Conference on*, pages 279–283. IEEE, 2011.

H. Jain, C. Thao, and H. Zhao. Enhancing electronic medical record retrieval through semantic query expansion. *Information Systems and e-Business Management*, 10 (2):165–181, June 2010. ISSN 1617-9846. doi: 10.1007/s10257-010-0133-5. URL http://link.springer.com/10.1007/s10257-010-0133-5.

V. Jalali and M. Borujerdi. The Effect of Using Domain Specific Ontologies in Query Expansion in Medical Field. In *International conference on innovations in information technology (IIT2008)*, pages 277–281. IEEE, 2008.

S. Karimi, D. Martinez, S. Ghodke, L. Zhang, H. Suominen, and L. Cavedon. Search for Medical Records: NICTA at TREC 2011 Medical Track. In *Proceedings of TREC*, 2011.

S. N. Kim, D. Martinez, L. Cavedon, and L. Yencken. Automatic classification of sentences to support evidence based medicine. *BMC bioinformatics*, 12:S5, 2011.

B. King, L. Wang, I. Provalov, and J. Zhou. Cengage Learning at TREC 2011 Medical Track. In *Proceedings of TREC*, 2011.

B. Koopman and G. Zuccon. Why assessing relevance in medical ir is demanding. In *ACM SIGIR Forum*, 2014.

B. Koopman, P. Bruza, L. Sitbon, and M. Lawley. Analysis of the Effect of Negation on Information Retrieval of Medical Data. In *Proceedings of 15th Australasian Document Computing Symposium (ADCS)*. University of Melbourne, 2010.

B. Koopman, P. Bruza, L. Sitbon, and M. Lawley. Evaluating Medical Information Retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1139–1140. ACM, 2011a.

B. Koopman, P. Bruza, L. Sitbon, and M. Lawley. Towards Semantic Search and Inference in Electronic Medical Records: an Approach Using Concept-based Information Retrieval. In *Proceedings of the First Australian Workshop on Artificial Intelligence in Health 2011*, pages 1–10. CSIRO Australian e-Health Research Centre, 2011b.

B. Koopman, P. Bruza, L. Sitbon, and M. Lawley. AEHRC & QUT at TREC 2011 Medical Track: a Concept-Based Information Retrieval Approach. In *Proceedings of Text Retrieval Conference*, 2011c.

B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley. Information retrieval as semantic inference: a graph inference model applied to medical search. *Information Retrieval Journal*, 19(1-2):6–37, 2016.

B. Koopman, J. Russell, and G. Zuccon. Task-oriented search for evidence-based medicine. *International Journal on Digital Libraries*, pages 1–13, 2017.

J. Lafferty, A. K. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc., 2001.

M. E. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information storage and retrieval*, 4(4):343–359, 1968.

N. Limsopatham, C. Macdonald, I. Ounis, G. McDonald, and M. Bouamrane. University of Glasgow at Medical Records Track: Experiments with Terrier. In *Proceedings of TREC*, 2011a.

N. Limsopatham, C. Macdonald, I. Ounis, G. McDonald, and M. Bouamrane. University of glasgow at medical records track 2011: Experiments with terrier. In *Proceedings of TREC*, 2011b.

Z. Liu and W. Chu. Knowledge-based Query Expansion to Support Scenario-specific Retrieval of Medical Free Text. *Information Retrieval*, 10(2):173–202, 2007.

H. J. Lowe and G. O. Barnett. MicroMeSH: a microcomputer system for searching and exploring the National Library of Medicine's Medical Subject Headings (MeSH) Vocabulary. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 717. American Medical Informatics Association, 1987.

Z. Lu, W. Kim, and W. Wilbur. Evaluation of Query Expansion Using MeSH in PubMed. *Information retrieval*, 12(1):69–80, 2009.

M. Lui. Feature stacking for sentence classification in evidence-based medicine. In *Australasian Language Technology Workshop 2012 : ALTA Shared Task*, 2012.

C. Macdonald, R. McCreadie, R. Santos, and I. Ounis. From Puppy to Maturity: Experiences in Developing Terrier. *Open Source Information Retrieval*, page 60, 2012.

L. Manchikanti, F. Falco, and J. A. Hirsch. Necessity and implications of ICD-10: Facts and Fallacies. *Pain Physician*, 14(5):E405–E425, 2011a.

L. Manchikanti, F. J. Falco, and J. A. Hirsch. Ready or not! Here comes ICD-10. *Journal of neurointerventional surgery*, pages neurintsurg–2011, 2011b.

C. D. Manning. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (MA), 2008. URL http://informationretrieval.org/.

D. Martinez, S. Karimi, L. Cavedon, and T. Baldwin. Facilitating Biomedical Systematic Reviews Using Ranked Text Retrieval and Classification. In *Australasian Document Computing Symposium ADCS*, 2008.

D. Martinez, A. Otegi, A. Soroa, and E. Agirre. Improving search over Electronic Health Records using UMLS-based query expansion through random walks. *Journal of biomedical informatics*, 51:100–106, 2014. ISSN 1532-0480. URL http://www.ncbi.nlm.nih.gov/pubmed/24768598.

A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

D. Molla. Experiments with clustering-based features for sentence classification in medical publications: Macquarie test's participation in the alta 2012 shared task. In *Australasian Language Technology Workshop 2012 : ALTA Shared Task*, 2012.

D. Mollá, D. Martinez, and I. Amini. Towards information retrieval evaluation with reduced and only positive judgements. In *Proceedings of the 18th Australasian Document Computing Symposium*, pages 109–112. ACM, 2013.

D. Mollá, I. Amini, and D. Martinez. Document Distance for the Automated Expansion of Relevance Judgements for Information Retrieval Evaluation. In *ACM SIGIR Workshop on Gathering Efficient Assessments of Relevance (GEAR)*, 2014.

E. C. Özlem Uzuner, Imre Solti. Extracting Medication Information from Clinical Text. *Journal of the American Medical Informatics Association*, 2012.

M. A. C. Pastor, Y. Wang, and H. Fang. Exploiting Domain Thesaurus for Medical Record Retrieval. In *Proceedings of Text REtrieval Conference (TREC)*, 2012.

J. Pestian, C. Brew, P. Matykiewicz, D. Hovermale, N. Johnson, K. Cohen, and W. Duch. A Shared Task Involving multi-label Classification of Clinical Free Text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104. Association for Computational Linguistics, 2007.

M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980a.

M. F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980b.

C. Puckett. *The Educational Annotation of ICD-9-CM*. Channel pub., 2011.

Y. Qi and P.-F. Laquerre. Retrieving Medical Records with sennamed: NEC Labs America. In *Proceedings of TREC 2012*, 2013.

D. Ravindran and S. Gauch. Exploiting hierarchical relationships in conceptual search. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 238–239. ACM, 2004.

C. Rijsbergen. Information retrieval. butterworth, 1979.

K. Roberts, M. Simpson, D. Demner-Fushman, E. Voorhees, and W. Hersh. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track. *Information Retrieval Journal*, 19(1-2):113–148, 2015a.

K. Roberts, M. S. Simpson, E. M. Voorhees, and W. R. Hersh. Overview of the TREC 2015 Clinical Decision Support Track. In *TREC*, 2015b.

F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, T. Werge, L. J. Jensen, and S. Brunak. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. *PLoS computational biology*, 7(8):e1002141, Aug. 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002141. URL http://dx.plos.org/10.1371/journal.pcbi.1002141.

D. L. Sackett, W. M. Rosenberg, J. M. Gray, R. B. Haynes, and W. S. Richardson. Evidence based medicine: what it is and what it isn't. *Bmj*, 312(7023):71–72, 1996.

T. Sakai and C.-y. Lin. Ranking Retrieval Systems without Relevance Assessments - Revisited. In *The Third International Workshop on Evaluating Information Access (EVIA)*, pages 25–33, 2010.

M. Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc, 2010.

M. Schuemie, D. Trieschnigg, and E. Meij. DutchHatTrick: Semantic Query Modeling, ConText, Section Detection, and Match Score Maximization, 2011.

W. Shen, J.-Y. Nie, X. Liu, and X. Liui. An Investigation of the Effectiveness of Concept-based Approach in Medical Information Retrieval GRIUM @ CLEF2014eHealthTask 3. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*, 2014.

I. Soboroff, C. Nicholas, and P. Cahan. Ranking Retrieval Systems without Relevance Judgments. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 66–73. ACM, 2001.

K. Spackman and K. Campbell. Compositional Concept Representation Using SNOMED: Towards Further Convergence of Clinical Terminologies. In *Proceedings of the AMIA Symposium*, page 740. American Medical Informatics Association, 1998.

P. Srinivasan. Query expansion and medline. *Information Processing & Management*, 32(4):431–443, 1996.

D. R. Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in biology and medicine*, 31(4):526–557, 1988.

B. Tinsley, A. Thomas, J. McCarthy, and M. Lazarus. Atigeo at TREC 2012 Medical Records Track: ICD-9 Code Description Injection to Enhance Electronic Medical Record Search Accuracy. In *Proceedings of Text Retrieval Conference*, 2012.

J. Urbano, M. Marrero, and D. Martín. On the measurement of Test Collection Reliability. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13*, page 393, New York, New York, USA, 2013. ACM Press.

E. Voorhees. Natural Language Processing and Information Retrieval. *Information Extraction*, pages 724–724, 1999.

E. Voorhees and W. Hersh. Overview of the TREC 2012 Medical Records Track. In *The tenth Text REtrieval Conference, Gaithersburg, MD. National Institute of Standards and Technology*, 2012.

E. Voorhees and R. Tong. Overview of the TREC 2011 Medical Records Track. In *The tenth Text REtrieval Conference, Gaithersburg, MD. National Institute of Standards and Technology*, 2011.

E. M. Voorhees. Query Expansion Using Lexical-Semantic Relations. In *SIGIR94*, pages 61–69, 1994.

E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36(5):697–716, 2000a.

E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36(5):697–716, 2000b.

E. M. Voorhees and D. Harman. Overview of trec 2001. In *Trec*, 2001.

W. Webber, P. Chandar, and B. Carterette. Alternative assessor disagreement and retrieval depth. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 125–134. ACM, 2012.

M. Weeber, H. Klein, A. R. Aronson, J. G. Mork, L. De Jong-van Den Berg, and R. Vos. Text-based discovery in biomedicine: the architecture of the dad-system. In *Proceedings of the AMIA Symposium*, page 903. American Medical Informatics Association, 2000.

A. Yeh. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics (COLING)*, pages 947–953, Saarbrücken, Germany, 2000.

H. Yu and E. Agichtein. Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, 19(suppl 1):i340–i349, 2003.

H. Yu, V. Hatzivassiloglou, C. Friedman, A. Rzhetsky, and W. J. Wilbur. Automatic extraction of gene and protein synonyms from medline and journal articles. In *Proceedings of the AMIA Symposium*, page 919. American Medical Informatics Association, 2002.

W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. Knowledge-intensive Conceptual Retrieval and Passage Extraction of Biomedical Literature. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 655–662. ACM, 2007.

D. Zhu and B. Carterette. Using Multiple External Collections for Query Expansion. In *Proceedings of TREC*, 2011.

D. Zhu and B. Carterette. Exploring Evidence Aggregation Methods and External Expansion Sources for Medical Record Search. In *Proceedings of TREC*, 2012.

J. Zobel. How Reliable are the Results of large-scale Information Retrieval Experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 307–314. ACM, 1998.

Q. Zou, W. W. Chu, C. A. Morioka, G. H. Leazer, and H. Kangarloo. IndexFinder: a method of extracting key concepts from clinical texts for indexing. In *AMIA*, 2003.