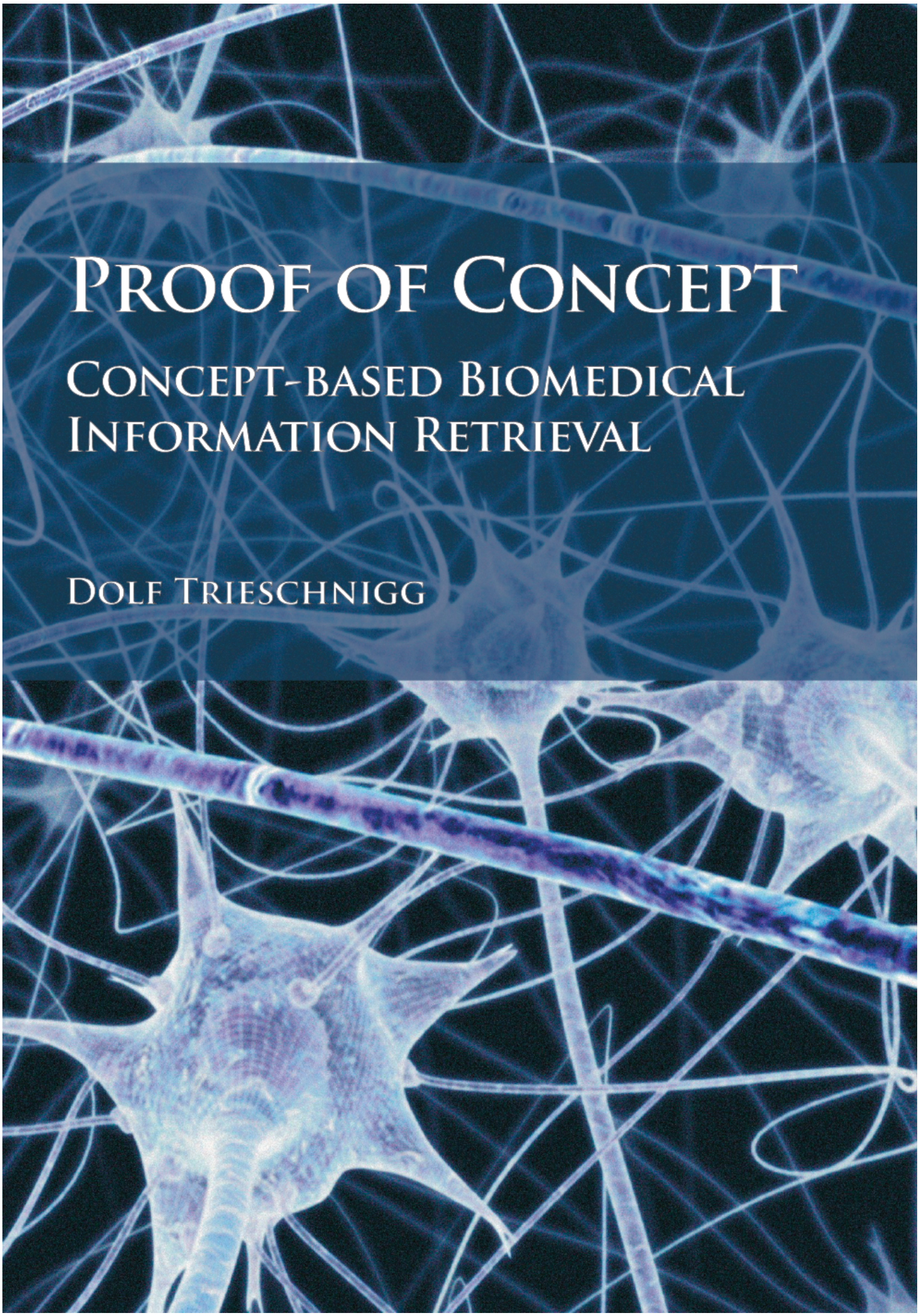

PROOF OF CONCEPT
CONCEPT-BASED BIOMEDICAL INFORMATION RETRIEVAL
DOLF TRIESCHNIGG

Proof of Concept

Concept-based Biomedical Information Retrieval

Dolf Trieschnigg

PhD dissertation committee:

Chairman and Secretary:

Prof. dr. ir. A. J. Mouthaan, University of Twente, NL

Promotores:

Prof. dr. F. M. G. de Jong, University of Twente, NL

Prof. dr. ir. W. Kraaij, Radboud University Nijmegen/TNO, NL

Members:

Prof. dr. H. J. van den Herik, Tilburg University, NL

Dr. ir. D. Hiemstra, University of Twente, NL

Prof. dr. T. W. C. Huibers, University of Twente, NL

Prof. dr. J. A. M. Leunissen, Wageningen Univeristy, NL

Dr. D. Rebholz-Schuhmann, European Bioinformatics Institute, UK

CTIT Ph.D. thesis Series No. 10-176, ISSN 1381-3617
University of Twente
Centre for Telematics and Information Technology (CTIT)
P.O. Box 217, 7500 AE Enschede, The Netherlands

SIKS Dissertation Series No. 2010-35
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Netherlands Bioinformatics Centre (NBIC)
This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

Human Media Interaction
The research reported in this thesis has been carried out at the Human Media Interaction research group of the University of Twente.

ISBN: 978-90-365-3064-4
ISSN: 1381-3617, No. 10-176
DOI: 10.3990/1.9789036530644

PROOF OF CONCEPT

CONCEPT-BASED BIOMEDICAL INFORMATION RETRIEVAL

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op woensdag 1 september 2010 om 15.00 uur

door

Rudolf Berend Trieschnigg

geboren op 20 april 1981
te Heino

Promotores:
Prof. dr. F. M. G. de Jong
Prof. dr. ir. W. Kraaij

ISBN: 978-90-365-3064-4

Acknowledgements

Finally, it is finished! I am very glad to write these acknowledgements, realising that it marks the end of a very hectic period. Despite that, I can look back on an enjoyable and valuable experience. I would like to thank the people who have enabled me to realise this thesis.

First of all, I would like to thank my supervisors Franciska de Jong and Wessel Kraaij. Franciska, thank you for offering me a PhD position and providing me the freedom to pursue my research. This thesis greatly benefited from your aid in writing. Wessel, thank you for the many interesting and motivating discussions we had in Delft and Rotterdam. I appreciate the time you made available in your busy schedule, even on days you were working from home. In Twente, I would like to thank Djoerd Hiemstra for always having his door open for lively discussion and for his comments on chapter 5 of this thesis. His enthusiasm motivated me a lot.

I would like to thank the NBIC BioRange programme, HMI and the Netherlands Genomics Initiative (NGI) for funding my research for the past years. I am grateful to TNO ICT for providing a workplace during my bi-weekly visits to Delft.

I am honoured that Jaap van den Herik, Djoerd Hiemstra, Theo Huibers, Jack Leunissen and Dietrich Rebholz-Schuhmann agreed to participate in the dissertation committee. I would especially like to thank Jaap van den Herik and Dietrich Rebholz-Schuhmann for their comments to improve this thesis.

A large part of the work reported in this thesis is the result of collaborations with colleagues across and outside the country. I enjoyed the collaborations with Edgar Meij and Maarten de Rijke (UvA), which resulted in a number of publications. I also appreciated the joint participations in the TREC Genomics benchmarks with Martijn Schuemie (ErasmusMC). Thank you for your help with cleaning up and concept-tagging the document collections, and for interesting discussions. Andra Waagmeester inspired me to apply for an EBI fellowship at the Netherlands Genomics Initiative. Thanks to this fellowship I was able to spend six months at the European Bioinformatics Institute in Cambridge (UK). I am very grateful to Dietrich Rebholz-Schuhmann for his hospitality at the Text Mining group of EBI. Beside the frequent table soccer matches with the group, I enjoyed the collaboration with Piotr Pęzik and Vivian Lee. Piotr, thanks for raising and discussing many questions and problems. Vivian, thanks for all your annotation work. Silvestras Kavaliauskas, thanks for making 'MeSH up' available as a webservice. I would like to thank Jetse Scholma from our own university for his assistance in analysing pairs of biomedical concepts.

I would like to thank all of my colleagues at HMI for creating a broad and interesting work environment. The sometimes absurd discussions during lunch, often provided rich food for thought. Some people I would like to mention in particular. I would like to thank Claudia Hauff for the many chats and discussions which contributed to this thesis. My

apologies for excessive use of the concepts [Mad cow disease] and [p53] on your whiteboard. Ingo Wassink turned out to be another willing victim for coffee breaks. Discussions about work and life (including sports related bruises and animal behaviour) made office life much more attractive. It is a pity you are not working at the UT anymore, but I will definitely see you around. Hendri, thanks for your support in my last-minute LaTeX, SVN, FTP and hard disk space requests. Charlotte, Alice and Ida (DB group), thank you for your administrative support. Many thanks to Lynn for proofreading and correcting my thesis.

I am very grateful to my family, family-in-law and friends for their support and their interest in the progress of my work. And for providing stress relieving activities, such as sailing, chopping lumber, digging ponds, mountain biking and playing poker games. Carolien, thanks for paving the PhD road in our family and for demonstrating that it is possible to write a thesis with more footnotes than pages. I could only rival you in the number of tables and equations. Remco, thanks for sharing experiences in PhD life, which I found very motivating. I am looking forward to your thesis.

Simon, thank you for providing your father an excellent deadline for finishing this thesis. A stroke to Teun and Siep for their purring support next to and sometimes on top of the keyboard. Last and foremost I want to thank Elske. Elske, thank you for supporting me through my 'delivery'. I am very lucky and proud to have you next to me.

Dolf Trieschnigg, July 2010

Contents

Chapter 1

Introduction

"A month in the laboratory can save an hour in the library."

Frank Westheimer[1]

This thesis will discuss the possibility to integrate domain-specific knowledge in biomedical information retrieval. The first chapter will introduce the field of biomedical information retrieval and the challenges related to its terminology. After that, the use of a concept-based representation for biomedical information retrieval will be motivated from a theoretical and a practical viewpoint. In section 1.5, three research themes and corresponding research questions will be described, followed by an overview of the chapters.

1.1 Biomedical IR

Recent decades have shown a fast growing interest in biomedical research, reflected by an exponential growth in scientific literature. MEDLINE, the primary bibliographic database for life sciences, contained more than 17 million article citations in 2009. In 2008, more than 600,000 new citations were added to the database (see Figure 1.1). Unsurprisingly, staying up-to-date and retrieving relevant information from this large repository of written scientific knowledge has become more challenging and more important. *Information retrieval* is defined as a field concerned with "the structure, analysis, organization, storage, searching, and retrieval of information" (Salton, 1968). Narrowing this definition, we define biomedical information retrieval as "the structure, analysis, organization, storage, searching, and retrieval of *biomedical* information". Biomedical IR is not only important for end-users, such as biologists, biochemists, and bioinformaticians searching directly for relevant literature but also plays an important role in more sophisticated *knowledge discovery*. During knowledge discovery, the available literature is automatically analysed to infer new knowledge or hypotheses. IR is required to reduce all the available literature to a large, but focused, set of documents which can be automatically analysed to find new relationships. Hence, biomedical knowledge discovery is strongly affected by and can greatly benefit from effective biomedical information retrieval systems.

[1] Late professor of Chemistry at Harvard University (citation from Lesk (2008))

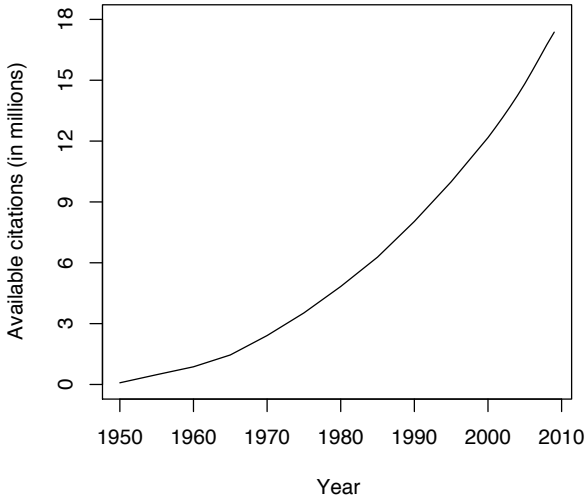


Figure 1.1: Number of available citations in MEDLINE.

1.2 Biomedical terminology

A major challenge for information retrieval in the life science domain is coping with its complex and inconsistent terminology (Krauthammer and Nenadic, 2004; Schuemie et al., 2005). The New Oxford American Dictionary (2005) defines *terminology* as: "the body of terms used with a particular technical application in a subject of study, theory, profession, etcetera". *Concepts* are defined as: "abstract ideas or general notions conceived in the mind". *Terms* are words or phrases used to refer to concepts. The terms 'mad cow disease', and 'BSE', for instance, refer to the concept [mad cow disease][2]. In the biomedical domain, the mapping between terms and concepts is particularly complex.

The difficulty of automatically handling biomedical terminology can be related to its complexity and inconsistency.

Complexity Biomedical terminology is inherently complex. Biomedical terms are often composed of several words or combine multiple terms. For example, the concept [nuclear factor kappa-light-chain-enhancer of activated B cells], also referred to as 'NF-κB'[3].

Inconsistency Biomedical terminology changes fast and new concepts and terms are frequently being introduced. Consider, for instance, the 2009 flu pandemic. The flu was caused by a novel strain of influenza, or to be more precise a variation of the 'Influenza A virus subtype H1N1'. Initially, it was referred to as 'Novel influenza A (H1N1)' or 'Novel influenza A/H1N1'. New terminology quickly appeared, such as '2009 H1N1 Flu', 'pig flu', 'Mexican flu', 'swine influenza' (abbreviated to 'SI'), 'North American influenza' and 'novel flu virus'.

[2]To distinguish between concepts and its terms throughout this thesis, concepts are enclosed in square brackets; terms are enclosed in 'single quotes'

[3]http://en.wikipedia.org/wiki/NF-κB

As a consequence many synonymous terms are encountered, which in turn can be ambiguous.

Synonymy As a result of inconsistent and complex terminology, many *synonyms* are encountered: multiple terms are used to refer to the same concept. These synonyms include spelling variation (for instance 'NF-κB' and 'NFkappaB'), symbols and abbreviations but also terms with totally different surface forms ('mad cow disease' and 'Bovine Spongiform Encephalopathy').

Ambiguity With so many terms (and in particular abbreviations) used to refer to concepts, biomedical terminology suffers from ambiguity: the same term is used to refer to different concepts. The polysemous term 'PSA', for instance, can refer to the concept [prostate specific antigen] but also to the concepts [puromycin-sensitive aminopeptidase], [psoriatric arthritis], [pig serum albumin] and many more.

The characteristics of biomedical terminology and its consequences for retrieval will be discussed in more detail in chapters 2 and 3 of this thesis.

From the above examples it is clear that the use of biomedical terminology causes a vocabulary mismatch problem for information retrieval: producers (authors) and consumers (searchers) of information use a different terminology to express the same, or similar concepts. It requires a considerable amount of *domain knowledge* to know what terms are used to express a concept. Or, perhaps more importantly which of these terms should not be used for searching because they are too ambiguous. Moreover, combining these terms effectively to find all relevant information on a particular topic can be difficult.

1.3 Early and contemporary biomedical IR

Early information retrieval, including biomedical IR, relied heavily on manual controlled vocabulary indexing: during this kind of indexing, expert indexers determine the most important concepts discussed in a document and assign appropriate index terms to the documents (Lancaster, 1969). To some extent, this type of indexing deals with the vocabulary mismatch problem described before: the representation used for indexing is independent from specific terminology used in the documents. One tough obstacle is, however, that the user has to formulate his[4] information need in terms of this controlled vocabulary, which can be difficult.

Modern retrieval systems commonly employ automatic word-based indexing, which uses all the words in a document as index terms in the retrieval system Manning et al. (2008). For end-users, this offers the possibility of formulating their queries in natural language. In contrast, additional effort is required to cope with a non-matching vocabulary. Lexical resources, such as domain-specific thesauri and controlled indexing vocabularies can be used to enhance text-based search and have been shown to be beneficial if implemented carefully Hersh et al. (2004). However, this type of conceptual knowledge is often incorporated in retrieval systems in an ad hoc fashion, mixed with a number of other approaches, or specifically designed for the task at hand. As a result, the added value of incorporating conceptual knowledge remains unclear.

[4]For brevity, we use "he" and "his" whenever "he or she" and "his or her" are meant.

1.4 Concept languages for biomedical IR

The main hypothesis of this thesis is that the effectiveness of biomedical IR can be improved by using a conceptual representation of documents and queries for indexing and searching.

Word-based IR suffers in particular from synonymous and ambiguous terminology. These characteristics can hurt retrieval performance in terms of both precision and recall. Recall is hurt when relevant documents use synonymous terms of terms in the query. Documents using terms that are synonyms of the terms in the query are not found. Precision is hurt by ambiguous terminology: ambiguous terms retrieve documents which use the term in a different sense than intended. To complicate IR even further, handling these characteristics will interfere with each other when they are handled in a word-based representation. Dealing with synonymy by expanding a query with synonymous terms, for example, can cause additional ambiguity problems. Expanding a query about the skin disorder 'atopic dermatitis' with its abbreviation 'AD' is likely to retrieve documents about Alzheimer's Disease as well.

A possible solution to the problems caused by these characteristics lies in carefully selecting the representation language. In theory, a conceptual representation is preferred over a word-based representation. Synonymous (including complex multi-word) terms are mapped to a single conceptual representation. Ambiguous terms are mapped onto the conceptual representation which corresponds to the context in which they appear. IR then simply reduces to matching the conceptual representations of documents to queries.

In practice however, a concept-based representation also has its limitations in improving the effectiveness of IR. These limitations are caused by the choice of conceptual representation language, how it is used for representing queries and documents and how the conceptual representation is obtained.

Firstly, limitations are introduced by the choice of the concept vocabulary. In this thesis, we will investigate the usefulness of two terminological resources as concept representation vocabularies. They both have their own advantages and disadvantages for this purpose. A small controlled vocabulary, for example, will not contain all fine-grained concepts (Hersh et al., 1994b). A large thesaurus might define concepts that are too specific for searching.

Secondly, limitations are introduced by the use of the concept vocabulary to represent documents and queries. For instance, when the topics in documents have not been exhaustively described in its concept-based representation, a query expressed in such a representation language will not retrieve all relevant documents (van Rijsbergen, 1979).

Thirdly, how the concept-based representations are obtained limits the effectiveness of such a representation. The concept-based representations can be based on manual labour, for example performed by a human indexer assigning concepts to documents, or by a user selecting concepts for searching. Such a manual approach can provide high quality representations, but is laborious and not user-friendly. A conceptual representation can also be generated automatically, but such a process can be error-prone, subsequently affecting retrieval effectiveness based on such a representation (Lam et al., 1999).

1.5 Research themes

The main subject of this thesis is dealing with terminology in biomedical information retrieval. We distinguish three research themes (RT) in this thesis.

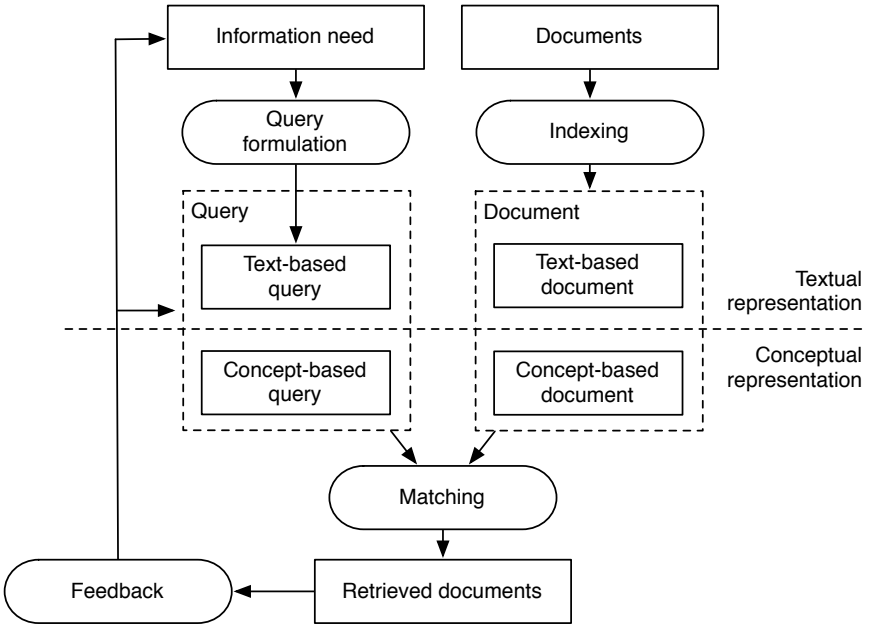


Figure 1.2: Separated text and concept representations in the IR processes. Adapted from Croft (1993).

RT1: Robust word-based retrieval

The first research theme in this thesis is concerned with making word-based retrieval more robust. Variations on word-based retrieval will be investigated to deal with one challenge of biomedical terminology: spelling variation. In chapter 3, we will investigate how choices in text preprocessing affect retrieval effectiveness in the biomedical domain. A combination of effective text preprocessing methods is proposed and used in subsequent chapters for creating word-based representations.

We will answer the following research question (RQ).

RQ1: *How can the effectiveness of word-based biomedical information retrieval be improved using document preprocessing heuristics?*

RT2: Concept-based retrieval

The second research theme in this thesis is concept-based retrieval. To investigate the added value of a concept-based representation, the word-based and concept-based representations are strictly separated. This separation is illustrated in Figure 1.2: A user has an *information need* which is converted into a (textual) query through a process of *query formulation*. The collection of *documents* is indexed to obtain a representation for the retrieval system. We assume that both the query and documents can be represented in terms of words and concepts. During the *matching* process, either or both representations are compared to obtain a set or list of *retrieved documents*. Through a *feedback* process the information need or query representation might be updated.

In chapter 4, the added value of a concept-based representation for biomedical IR will be investigated. We will investigate the following five topics.

RT2a: How documents are represented in a concept-based representation.
RT2b: To what extent such a document representation can be obtained automatically.
RT2c: To what extent a text-based query can be automatically mapped onto a concept-based representation and how this affects retrieval performance.
RT2d: To what extent a concept-based representation is effective in representing information needs.
RT2e: How the relationship between text and concepts can be used to determine the relatedness of concepts.

We will propose and investigate two approaches to obtain a concept-based representation from text automatically and will demonstrate their usefulness for improving word-based retrieval and predicting concept relatedness.

We will answer the following research question.

RQ2: *What is the added value of a concept-based representation based on terminological resources for biomedical IR?*

RT3: A framework for concept-based retrieval

The approach of strictly separating a word and concept-based representation is quite unsophisticated: it might not be as effective as some of the ad hoc approaches to integration of concept-based information which use a combined representation.

In chapter 5, we will propose a framework for a more tight integration between a word and concept-based representation. The framework aids in analysing the integration of a concept-based representation in IR. We will demonstrate the usefulness of such a framework by implementing a selection of translation and retrieval models and evaluating their effectiveness.

We will answer the following research question.

RQ3: *Is it possible to cast the integration of knowledge from terminological resources in biomedical IR into a retrieval framework?*

1.6 Thesis overview

The overview of this thesis is as follows.

Chapter 2 will provide a general background to this work. It introduces biomedical information retrieval, discusses its terminological challenges and summarises related work.

In chapter 3, text or more precisely, word-based biomedical IR will be investigated. In particular, document preprocessing heuristics will be compared which try to cope with spelling variations encountered in biomedical terminology. RT1 will be examined in this chapter.

In chapter 4, a concept-based approach to biomedical IR will be investigated. It focusses on the characteristics of a concept-based representation, on the mapping between textual and conceptual representations of both queries and concepts and lastly the determination of concept relatedness. RT2 will be examined in this chapter.

In chapter 5, a framework will be presented in which textual and conceptual representations can be more tightly integrated. RT3 will be examined in this chapter.

Finally, in chapter 6 we will answer the research questions, summarise our contributions and indicate directions for future work.

Chapter 2

Background

"Biologists would rather share their toothbrush than a gene name."

Michael Ashburner[1]

The goal of this chapter is to serve as a background for chapters to follow for researchers from both the biomedical and the IR community[2]. It introduces retrieval terminology to readers with a biomedical background and the biomedical domain to readers with an IR background. In sections 2.1 and 2.2 a brief introduction is provided to information retrieval, with an emphasis on the biomedical domain and its terminological challenges. Then, a high level overview of approaches to cope with these challenges is discussed (section 2.3). Finally, an overview of experiments and experiences in biomedical IR is provided, with a particular focus on the TREC Genomics evaluation benchmark (section 2.4).

2.1 Information retrieval

Most readers will be familiar with web search engines such as Google and Yahoo. These are *information retrieval* (IR) systems for the Web: based on a few keywords provided by the user, these systems try to present the most relevant web pages. In this section, a brief introduction is presented into information retrieval.

Traditionally, IR research has been concerned with retrieval of textual information, but in the last few decades its focus has broadened to different types of information, such as audio, video, and even entities. This thesis is focused on the disclosure of biomedical literature. The term *document* is used to refer to the unit of retrieved information. This may be a citation consisting of a title and an abstract, a complete journal article or a selected passage from such a publication.

A typical information retrieval setting consists of a *user*, a *collection* of documents and an *IR system*. The user has an information need, formulates a *query*, and submits it to the retrieval system. In response, the system presents a selection of documents from the

[1] Professor of biology in the Department of Genetics at University of Cambridge, UK (quote from Pearson 2001)

[2] Primary sources of information for this chapter are van Rijsbergen (1979); Baeza-Yates and Ribeiro-Neto (1999); Kraaij (2004); Manning et al. (2008); Zhai (2008), and Hersh (2009).

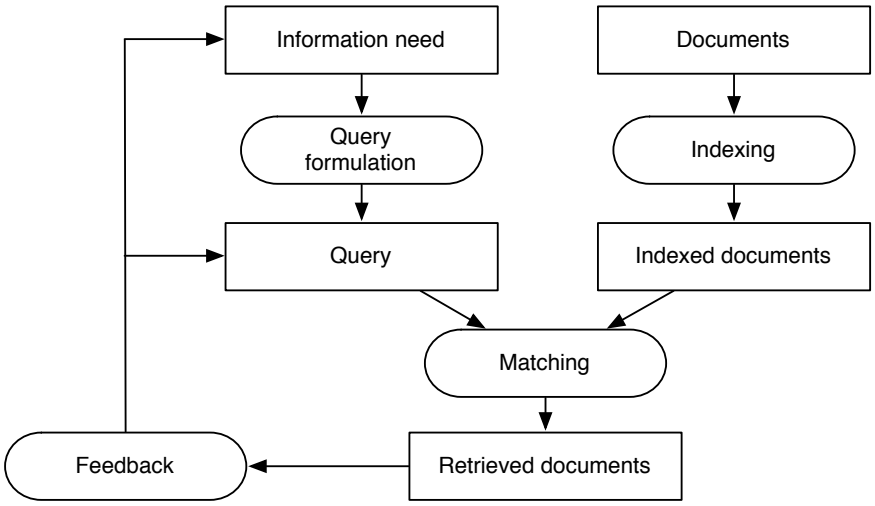


Figure 2.1: Information retrieval processes (from Croft 1993).

collection. It is up to the user to decide which of these documents are *relevant*, that is, which documents contribute to answering his information need and whether his information need has been met. If not, the user may wish to reformulate the query and resubmit it to the system. Alternatively, the system may allow the user to give *relevance feedback*, that is, letting the user indicate which retrieved documents are relevant or not. Subsequently, this information can be used by the system to retrieve additional relevant documents, or to reorder the documents in such a way that the most relevant documents are presented first.

For small IR problems, such as finding a particular paper on your desk, simply browsing through all available information can be quite effective. For larger collections, however, such a linear search soon becomes unfeasible. Before retrieval can take place, a structure has to be built which allows fast and effective retrieval.

An IR system commonly distinguishes between *indexing*, *query formulation*, and *matching* processes, visualised in Figure 2.1. The indexing process is carried out once before querying, or incrementally as new documents are added to the collection, resulting in an index structure which allows fast lookup. The user is involved in the process of formulating a query to represent his information need. The retrieval system matches this query to the indexed documents and returns a set or ranked list of retrieved documents. In subsection 2.1.1, the indexing process will be described in more detail. After that, the query formulation and matching process will be discussed in subsection 2.1.2. In subsection 2.1.3 a brief introduction will be provided to the retrieval model used throughout this thesis, based on statistical language models.

2.1.1 Indexing

Indexing is the process of assigning *index terms* to documents. The set of index terms assigned to a document form the document's index description and should give a topical description of the document. An index term could, for example, be a single keyword such as 'cancer', or a fine-grained phrase such as 'male breast cancer', indicating that documents assigned with that term discuss that topic to some extent. The set of indexing terms used to index a collection forms the *index language* or *index vocabulary*. The choice of an indexing

vocabulary strongly influences the characteristics of the retrieval system.

The index should strike a balance between *exhaustivity* and *specificity*. The exhaustivity of indexing is defined as the number of different topics indexed (van Rijsbergen, 1979); the number of index terms assigned to a document can be used as an indicator of its index description's exhaustivity. The specificity of the index language is its ability to describe topics precisely (Cleverdon et al., 1966; Spärck Jones, 1972; van Rijsbergen, 1979); the number of documents to which an index term is assigned can be used as an indicator of the term's specificity. For example, indexing a document with the term 'cancer' when it only remotely discusses this topic would be part of an exhaustive description of the document. In contrast, the specificity of the index term decreases since (a binary) assignment of the term cannot discriminate documents discussing the topic in detail from documents only marginally mentioning it.

The index vocabulary can either be *controlled* or *uncontrolled*, indicating whether the terms in it are manually maintained or not. A second, closely related, distinction is whether the actual indexing is carried out *automatically* or *manually*. Automatic indexing is often combined with an uncontrolled vocabulary: the vocabulary is then determined by, for example, the words encountered in the documents. Manual indexing is often combined with a controlled vocabulary; maintaining the vocabulary is then combined with manually indexing the documents.

These two indexing approaches will now be described and compared.

Manual indexing using a controlled vocabulary

Manual, controlled vocabulary indexing has its roots in library science, where for centuries librarians manually categorised their books to allow lookup. In this scenario, a human indexer manually selects appropriate index terms for each publication. With new topics appearing, new index terms are also added to the index vocabulary. Often the terms in these controlled vocabularies are organised in some form of hierarchy. The hierarchical relationships can, for example, indicate meronymy (part-of relationships) or hyponymy (is-a relationships) between connected terms. Assembling and organising such a controlled vocabulary can be regarded as a *categorisation* task: depending on the collection to be organised, appropriate categories are determined and arranged. Indexing new documents with appropriate terms from the vocabulary can be viewed as a *classification* task: the vocabulary does not (directly) change as a result of the indexing process (Jacob, 2004).

Automatic indexing using an uncontrolled vocabulary

Around the 1960s, an alternative to manual, controlled indexing was first presented (Luhn, 1957). Rather than using the terms from a carefully crafted, controlled vocabulary, Luhn suggested the use of words found in the text for *free-text* indexing, which turned out to be an effective method. The development of the computer further fuelled research into automatic *full-text* indexing, which uses the complete document text for extracting index terms. This preprocessing step of automatically obtaining index terms from documents is discussed in more detail in chapter 3. For the time being, index words can be regarded as an uninterrupted sequence of letters or digits encountered in free text.

A basic indexing approach discards word order and keeps track of the documents in which a particular index term can be found. Additionally, the positions of the terms in the

document can be stored in the index to enable phrase or proximity searches of index term combinations (searching, for example, for documents with the index terms 'protein' and 'binding' next to each other). Such a structure allows for more complex *post-coordinate* matching: index terms can be combined at search time. In contrast, in a *pre-coordinated* index, more complex subjects are indexed with a single term. For example, a document can be indexed with the single index term 'breast cancer' rather than with two index terms 'breast' and 'cancer'.

Pros and cons of manual indexing with a controlled vocabulary

There are a number of differences between manual, controlled vocabulary indexing and automatic uncontrolled indexing and they both have their advantages and disadvantages.

The first advantage of using manual, controlled vocabulary indexing is normalisation. The human indexer has to read and understand the document and has to select the most appropriate index terms. Variations in language use in different documents on the same topic (consider, for example, the language in a highly technical document versus the introduction to a topic) are normalised by indexing them with the same term. Synonymous terminology, that is different textual expressions with the same meaning, can be indexed using the same term. Moreover, ambiguous terminology, that is the same word with different meanings, can be indexed in an unambiguous manner. In subsection 2.2.3, it is explained how important this normalisation is for the biomedical domain. A second advantage is that some form of abstraction can take place, by, for example, indexing a document about both rats and mice with the more general index term 'rodents'. Thirdly, a controlled vocabulary often relates indexing terms to each other by structuring them in a tree-like hierarchy. Depending on the type of relationships (for example, is-part-of or is-a relationships), this makes broadening or narrowing a search easier, by picking parent or child terms for searching.

There are a number of drawbacks to indexing this way. Firstly, it is labour intensive and therefore expensive to carry out manual indexing. Secondly, indexing and consistency errors can be made. A text can be incorrectly interpreted by a human indexer, resulting in incorrect indexing terms. Different indexers might not agree on the indexing terms used for a particular document and an indexer might use different terms when indexing a document a second time. Thirdly, there is the issue of flexibility and maintainability of a controlled vocabulary over time. New documents might address topics which are not covered by the vocabulary, requiring new or more specific index terms to be added to the language. These changes to the vocabulary might require older documents to be re-indexed, which becomes an infeasible job with a large and growing collection.

Pros and cons of automatic indexing

Automatic, uncontrolled indexing also has a number of advantages and disadvantages. We will mention four of them. Firstly, automatic indexing is cheap in comparison to controlled vocabulary indexing, especially with current computing and storage capabilities. Secondly, uncontrolled indexing is usually more exhaustive than controlled vocabulary indexing. More terms are assigned to a document which allows them to be found more easily. Thirdly, there is no longer an issue with consistency: every document is indexed using exactly the same process. Hence, indexing a document twice results in the same index terms. Fourthly,

an automatic index is easier to maintain: new terms are automatically added to the index vocabulary, when new terms are encountered during indexing of new documents.

There is also a number of disadvantages to automatic indexing using an uncontrolled vocabulary. We will mention three. Firstly, the selection process of indexing terms is limited: all words are used as indexing terms, requiring *weighting* to determine the relative importance of terms both within a document and between documents. The word 'cancer' in a document is more important than the word 'the'; a document containing 'cancer' once is probably not as important as another mentioning it five times. Secondly, depending on which automatic indexing unit is used, potentially valuable dependency information is lost during indexing. For example, word combinations may lose their informativeness when separated (for example, 'division' and 'cell' separately are far less informative than 'cell division'). Thirdly, without any additional processing, no abstraction or normalisation is available: the index descriptor is limited to what is literally mentioned in the text. Summarising, the interpretation, abstraction and normalisation which takes place during manual indexing is not available for automatic full-text indexing.

2.1.2 Query formulation and matching

During the searching process, the user faces a query formulation problem: his information need has to be formulated as a query to the system. In the case of full-text indexing, the query can be formulated in free text. In the case of a controlled vocabulary index, the user has to select suitable terms, perhaps semi-automatically, from the vocabulary to search with. The retrieval model determines how the query is matched against the document representations. In the next block, the Boolean retrieval model will be discussed, which is frequently used in combination with controlled vocabulary indexing. In the subsequent block, ranked retrieval models will be discussed, which are commonly used in combination with free text indexing.

Exact match retrieval: the Boolean model

The Boolean model is the first model used for information retrieval. Based on Boolean operators, such as AND, OR, and NOT, query terms can be combined to precisely describe which documents should be retrieved. For instance, the query "(cancer OR neoplasms) AND NOT stomach" would return documents indexed with 'cancer' or 'neoplasms' (or both), but would filter out documents indexed with the term 'stomach'. The basic Boolean model is an *exact match* retrieval model: it only retrieves documents that match the given query exactly. In contrast, *partial match* retrieval systems do not require all query terms to be present in matching documents.

Advantages of the strict Boolean model are its implementation efficiency and the amount of control the query language gives the user to retrieve (or not to retrieve) documents. The control of building complex queries is also a disadvantage, however: naive users find it difficult to build good queries. A second major disadvantage is that it is not trivial to incorporate term weighting and relevance feedback in a theoretically sound way.

Ranked retrieval models

Ranked retrieval models try to retrieve the most relevant documents first in response to a query. Often this is combined with partial matching: documents not containing all query terms may, for example, still be relevant, but be returned at a lower rank so that the user is still able to find them. Ranking is particularly useful when documents are exhaustively indexed, as in the case of free text indexing. Since more documents will match a query, ranking is beneficial to present the most relevant documents first.

Many IR systems treat documents and queries during retrieval as *bags-of-words*: determining the (relative) relevance of documents does not take into account the order of words. More complex representations incorporating term dependencies have been shown to perform only slightly better at best and they tend to suffer from data sparseness (see subsection 2.3.1).

Empirically effective models in essence combine three important components (Zhai, 2008). Firstly, a *term frequency* (*TF*) component which indicates the local importance of a term in a document: a document containing a term often is more likely to be about that term. Secondly, an *inverse document frequency* (*IDF*) component, which indicates the global importance of a term: terms occurring in many documents are less important for searching. Thirdly, some form of *document length normalisation*: a longer document containing a particular term the same number of times as a shorter document is likely to be less relevant. Different retrieval models have been proposed in the past, varying from high-dimensional vector calculations to models based on probability theory and formal logic. Discussing these models in detail is outside the scope of this thesis. Overviews can be found in, for example, Baeza-Yates and Ribeiro-Neto (1999); Manning et al. (2008), and Zhai (2008).

2.1.3 Language Model IR

Retrieval models based on statistical language models (LM IR) were introduced in the late 1990s after successful applications in speech recognition and machine translation. LM IR has been appreciated for its sound statistical foundations in combination with its simplicity and strong performance in retrieval evaluations (Ponte and Croft, 1998; Berger and Lafferty, 1999; Hiemstra and Kraaij, 1999; Miller et al., 1999). Central to LM IR are *language models*, which are probability distributions over language use, or, more precisely, over word sequences.

A general language model of English could, for example, assign a probability to the sequence of words 'Cancer is caused by smoking', a smaller probability to 'smoking is caused by cancer' (since it is less likely to be discussed) and an even smaller probability to 'caused is cancer smoking by'.

The most commonly used language models for IR are based on single terms rather than sequences of terms. In these unigram language models, the words are assumed to occur independently (term independence). The models are defined as multinomial probability distribution over single words. For example, the probability of observing the sequence of words 'colon cancer' in a fragment of English is assumed to be the product of the word probabilities: $P(\text{'colon', 'cancer'}) = P(\text{'colon'})P(\text{'cancer'})$. Moreover, the sum of the word probabilities over all possible words (in the index vocabulary V) equals 1: $\sum_{w \in V} P(w) = 1$.

The documents in a collection can be represented by *document language models*. These language models can be used to assign a probability to a certain sequence of terms. For

example, a document LM representing a document discussing the relationship between cancer and smoking might assign a higher probability to 'cancer is caused by smoking' than the LM of a document about a totally different topic.

One of the earliest LM retrieval models is based on *query likelihood*: documents, or rather their language models, are ranked according to the probability of generating the query, that is, the probability of drawing the query terms from the document language model. Formally, documents are ranked according to $P(Q|\theta_D)$, where Q is the query and θ_D is the document language model. The sequence of query terms q_1 to q_n in the query is assumed to be independently sampled from the document language model. The likelihood of sampling the query from the document can thus be calculated as follows.

$$P(Q|\theta_D) = P(q_1, \ldots, q_n|\theta_D) = \prod_{i=1..n} P(q_i|\theta_D) \tag{2.1}$$

Document language model estimation

The parameters of the document language model, the values of $P(w|\theta_D)$, are commonly based on the relative frequencies of words in the document, *smoothed* with probabilities from a background model. Smoothing makes the document language models more robust for retrieval, especially when the documents are small. Moreover, smoothing "explains" the non-informative words in the query. In this case smoothing has an IDF function, that is it decreases the importance of more common terms in the query (Zhai and Lafferty, 2004; Zhai, 2008). Several smoothing methods exist, such as Jelinek-Mercer smoothing, additive smoothing, Dirichlet prior smoothing, smoothing using absolute discounting and Good-Turing smoothing (Jelinek and Mercer, 1980; Katz, 1987; Chen and Goodman, 1998).

Formally, the parameters of the document language model (adopting Jelinek-Mercer smoothing) are estimated as follows.

$$P(w|\theta_D) = (1-\lambda)P(w|\hat{\theta}_D) + \lambda P(w|\hat{\theta}_{\mathcal{C}}) \tag{2.2}$$

$$P(w|\hat{\theta}_D) = \frac{f(w,D)}{|D|} \tag{2.3}$$

$$P(w|\hat{\theta}_{\mathcal{C}}) = \frac{\sum_{D\in\mathcal{C}} f(w,D)}{\sum_{D\in\mathcal{C}} |D|} \tag{2.4}$$

$P(w|\hat{\theta}_D)$ is the probability of the term w in the document language model based on a maximum likelihood estimate, that is, the relative frequency of the word in the document ($f(w,D)$ is the term frequency of the word, the number of times a word appears in a document, and $|D|$ is the length of the document). $P(w|\hat{\theta}_{\mathcal{C}})$ is the background or collection model which assigns probabilities to terms based on a large set of documents $\mathcal{C}$. The amount of smoothing is controlled by the parameter λ.

Probabilistic distance retrieval models

Besides ranking based on query likelihood, a second, more flexible approach to LM IR is to define a *query language model* and to rank documents by comparing its language models

to this query language model. The initial parameters of the query language model are commonly based on the relative frequencies of words in the query. Subsequently, a more precise query language model can be based on (pseudo) relevance feedback (Lavrenko and Croft, 2001; Zhai and Lafferty, 2001).

Formally, the query language model based on the initial query is estimated as follows.

$$P(w|\theta_Q) = \frac{f(w,Q)}{|Q|} \tag{2.5}$$

Where $f(w,Q)$ is the term frequency of the word w in the query and $|Q|$ is the query length, that is, the total number of words in the query.

Different but related measures, such as Kullback-Leibler (KL) divergence and Cross Entropy Reduction (CER), have been proposed for comparing the language models (Kraaij, 2004; Zhai and Lafferty, 2006). As ranking functions, they both essentially calculate the negated cross entropy ($-H(\theta_Q, \theta_D)$) of the query language model with respect to the document language model plus a query dependent constant. The retrieval status value (RSV), the score used to rank a document, is calculated as follows.

$$\begin{aligned} RSV_{KL}(D,Q) &= -D(\theta_Q||\theta_D) = -\sum_{w\in V} P(w|\theta_Q)\log\frac{P(v|\theta_Q)}{P(w|\theta_D)} \\ &= \sum_{w\in V} P(w|\theta_Q)\log P(w|\theta_D) \left[-\sum_{w\in V} P(w|\theta_Q)\log P(w|\theta_Q)\right] \\ &= -H(\theta_Q,\theta_D)\,[+H(\theta_Q)] \end{aligned} \tag{2.6}$$

$$\begin{aligned} RSV_{CER}(D,Q) &= D(\theta_C||\theta_Q) - D(\theta_Q||\theta_D) = \sum_{w\in V} P(w|\theta_Q)\log\frac{P(w|\theta_D)}{P(w|\theta_C)} \\ &= \sum_{w\in V} P(w|\theta_Q)\log P(w|\theta_D) \left[-\sum_{w\in V} P(w|\theta_Q)\log P(w|\theta_C)\right] \\ &= -H(\theta_Q,\theta_D)\,[+H(\theta_Q,\theta_C)] \end{aligned} \tag{2.7}$$

The query dependent constant, enclosed by square brackets in the previous equations, can be left out for ranking purposes. For comparing scores across different queries, for example, in the case of topic detection and clustering, the constant does play an important role (Kraaij, 2004).

A more comprehensive discussion of language model IR can be found in Zhai (2008).

2.1.4 Evaluation

An important theme of information retrieval research is to find out whether the systems perform well in practice. Retrieval *effectiveness* indicates to what extent the retrieval system retrieves relevant rather than non-relevant documents. Retrieval effectiveness is often determined in a laboratory setting. In the Cranfield (Cleverdon, 1967) and Text REtrieval Conference (TREC) tradition (Voorhees and Harman, 2005), a test collection consisting of a document collection, a set of user topics and relevance judgements is assembled and reused for evaluating retrieval systems. A typical benchmark collection is constructed in

the following way. Firstly, a task and a document collection is selected. For example, an ad hoc search task: find all documents discussing a particular topic, enabling the user to write an article about it. The document collection consists of a fixed set of documents, for example a set of news articles over a period of time, or a set of scientific articles. Secondly, a set of queries is chosen, for example by asking a number of domain specialists to write down their information needs. Thirdly, relevance judgements are gathered to determine which documents in the collection are relevant for each query. Since, it is not feasible to determine the relevance of each and every document for a large collection of documents, a *pooling* method is commonly employed (Spärck Jones and Van Rijsbergen, 1975). A pool of documents is created by selecting the top-ranked documents from a number of different IR systems. This pool is subsequently judged on its relevance. Despite the incompleteness of this set, these pooled relevance judgements can be used reliably to compare the system performance (Zobel, 1998; Buckley and Voorhees, 2004).

For the calculation of retrieval effectiveness, documents are considered relevant or non-relevant for a particular topic. This is obviously debatable, but makes evaluation more straightforward.

A distinction can be made between set-based and rank-based effectiveness measures. Set-based measures indicate the quality of a set of retrieved documents. Rank-based measures also take into account the rank at which documents are retrieved. The latter is necessary for ranked retrieval systems which try to order the documents in decreasing probability of relevance. The metrics are averaged over a set of topics to compare the performance across systems.

The most important set and rank-based metrics will be described in the next two blocks. The last two blocks of this subsection describe significance testing and IR evaluation outside the lab.

Set-based metrics

The primary set-based metrics are *precision* and *recall*. The precision of a set of retrieved documents is the fraction of retrieved documents which are relevant to the query. The recall of a search is the fraction of relevant documents in the collection retrieved by the system. The metrics are defined as follows (van Rijsbergen, 1979).

$$\begin{aligned} precision &= \frac{r}{n} \\ recall &= \frac{r}{R} \end{aligned} \qquad \begin{aligned} &r : \text{number of relevant retrieved documents} \\ &n : \text{number of retrieved documents} \\ &R : \text{total number of relevant documents} \end{aligned} \tag{2.8}$$

For example, when the collection contains 20 relevant documents, and the set of 100 documents retrieved by the system contains 15 of them, the recall is $\frac{15}{20} = 0.75$ and the precision is $\frac{15}{100} = 0.15$.

Usually a trade-off can be observed between precision and recall: the precision of a search can be increased at the cost of recall and vice versa. For instance, a retrieval system which would simply return all documents in response to a query would achieve a recall of 1 at the lowest possible precision. The system can increase precision by returning fewer documents, however at the risk of lowering recall by missing relevant documents.

Precision and recall can be combined into a single *F-measure*, which is defined as the weighted harmonic mean of precision and recall. The parameter β indicates the relative importance of recall over precision.

$$F_\beta = \frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall} \tag{2.9}$$

Rank-based metrics

The rank-based retrieval measures such as *rank precision* and *average precision* are based on precision and recall, but also take rank into account. Rank precision (precision at rank X, P@X) is used to indicate the precision of the highest ranked documents. P@10 for example, indicates the precision of the first 10 retrieved documents. Average precision (AP) is a single value which takes into account both precision and recall. It is calculated by averaging the rank precision of the relevant documents; the rank precision of relevant documents not retrieved by the system is assumed to be 0.

The AP is calculated as follows.

$$AP = \frac{\sum_{i=1}^{n} precision(i) \times rel(i)}{R} \tag{2.10}$$

Where n is the number of retrieved documents; R is the total number of relevant documents; $precision(i)$ is the precision of the retrieved documents at rank i, and $rel(i)$ is a binary function which indicates whether the document retrieved at rank i is relevant (1) or not relevant (1).

For example, when a system finds 3 of 4 relevant documents at rank 1, 4, and 10, the average precision for this topic is: $\frac{1/1 + 2/4 + 3/10}{4} = 9/20$. When averaging the AP over a collection of topics, this gives the *mean average precision* or MAP, commonly used to express the effectiveness of a retrieval system on a particular benchmark collection.

Significance testing

An important aspect of comparing the retrieval effectiveness of two systems is determining whether the differences are significant. A higher average performance score (MAP or average rank precision) might suggest that one system is better than another, but a significance test should point out how likely it is that this difference was encountered by chance. Different significance tests are used for this purpose, such as the Student's paired t-test, Wilcoxon signed rank test, and the so-called sign test (Fisher, 1935; Hull, 1993; Smucker et al., 2007). The tests differ in the assumptions they make about the data. A paired t-test, for example, assumes that the differences between the two populations of performance scores follow a normal distribution, an assumption which can be easily violated by the performance scores of a system over a set of topics. As a result, incorrect conclusions may be drawn from a significance test: an insignificant difference can be judged as significant (type-I error), or vice versa (type-II error). Throughout this thesis the sign test is used. The sign test makes only few assumptions about the data and is accurate (few type-I errors), at the cost of sensitivity, however (more type-II errors).

Evaluations outside the lab

IR evaluation is not limited to determining retrieval effectiveness. Additionally, the speed of indexing and retrieval, and the size of the index can be evaluated. Outside this laboratory setting, user studies can be carried out to determine the user satisfaction of a system. A drawback of these studies is that they are costly and cannot be quickly repeated.

2.2 Biomedical IR

Biomedicine covers a large number of disciplines including (human and veterinary) medicine and biosciences, such as (bio)chemistry, biology, molecular biology, biomedical engineering, botanics, and microbiology. It deals with a broad range of biological and medical topics investigated from different viewpoints and at different levels of detail.

The results of biomedical research are primarily disseminated through written publications, such as books and periodicals. In 2009, MEDLINE, the bibliographic database maintained by the U.S. National Library of Medicine (NLM) contained more than 17 million references to biomedical journal articles[3] and has shown an exponential growth in the number of published publications since the 1950s. In 2008, over 600,000 new citations were added to the repository. The full texts of these publications are also becoming more freely available through open-access publishers such as BioMed Central[4]. Accessing these vast amounts of literature has become increasingly difficult, demanding effective biomedical information retrieval systems.

In the following subsections, the history and modern-day practice of biomedical IR will be discussed, followed by a discussion of challenges related to its terminology and resources to cope with these challenges. Finally, the evaluation of biomedical IR will be discussed.

2.2.1 Early biomedical indexing

Making biomedical literature accessible was first attempted more than a century ago when two early controlled vocabulary indices, the Index-Catalogue and Index Medicus were created (Coletti and Bleich, 2001; Greenberg and Gallagher, 2009).

The *Index-Catalogue of the Library of the Surgeon-General's Office, United States Army*, Index-Catalogue in short, was intended to be a complete index of biomedical literature, covering books, journal articles, and theses. The index was published in series of revolving alphabetical volumes: first the 'A'-volume would appear, containing all index terms starting with an A and corresponding publications, followed by the next alphabetical volume. Its construction was incredibly labour intensive: the first series of volumes finally finished after 15 years in 1895. Obviously, this index suffered greatly from the slow production process and the large backlog of publications not yet indexed.

Therefore, an additional publication was made available to stay up-to-date with recent publications. John Shaw Bilings started in 1879 with a service called Index Medicus: the publication would present a selection of recently published journal articles, theses, and books arranged by subject. In 1926, the Index Medicus was merged with a similar service called the *Quarterly Cumulative Index to Current Literature*.

[3] `http://www.nlm.nih.gov/bsd/revup/revup_pub.html#med_update`, accessed 4th of August 2009.
[4] `http://www.biomedcentral.com`, accessed 4th of August 2009.

In 1950, it was decided to discontinue the Index-Catalogue. The Index-Catalogue had such a long backlog that it had lost its usefulness: it could take up to a decade until a new citation would appear in print. The Index Medicus was more successful, however: in 1960, a renewed Index Medicus appeared using a "freshly revised and expanded list of standardised subject headings" (Coletti and Bleich, 2001) called *Medical Subject Headings* (MeSH). This controlled vocabulary is updated yearly and still in use today (see subsection 2.2.4).

The invention of the computer triggered the development of one of the first biomedical bibliographic retrieval systems called MEDLARS (Medical Literature Analysis and Retrieval System), which became available in 1964 (Lancaster, 1969). The system was in fact a computerised Index Medicus. The search system used punched cards for submitting queries to the system, required up to 3 months of training to operate and had a turnaround time for a search request of around 4 to 6 weeks (Coletti and Bleich, 2001). The system was superseded by an online system in 1971, MEDLARS ONLINE, shortened to MEDLINE. MEDLINE allowed queries to be issued over a telecommunication line. The service still required users to take two weeks training, including an introduction on how to use MeSH. Searches were often *mediated*, that is the actual information consumer discussed his information need with a trained librarian, the latter actually formulating and issuing the queries. Since the mid 1990s, MEDLINE has been accessible on the internet as a subset of PubMed[5]. PubMed also includes in-process citations and citations of journal articles before they are officially added to MEDLINE.

2.2.2 Modern-day biomedical IR: serving knowledge discovery

For many users, PubMed is still the entry-point when searching for biomedical literature. But biomedical IR is more than finding related literature for end-users (Shatkay and Feldman, 2003; Krallinger and Valencia, 2005; Shatkay, 2005). Hersh (2009) described IR as one of the first steps in a knowledge acquisition funnel depicted in Figure 2.2. Information retrieval forms the entry point for knowledge acquisition: it reduces the entire volume of available literature to a smaller, focused set of publications. A retrieval system can, for example, retrieve all publications about a particular gene. This initial process may still result in a large number of related publications. In a following information extraction step, facts can be extracted from this set of documents. For example, a named entity recognition process can be used to find (other) genes or proteins mentioned in the texts. The co-occurrence of the gene of interest with other genes and protein names in a text might indicate a (known, hypothesised or denied) relationship between the two. Additionally, automatic analysis of the verbs connecting the two genes might give insight into the type of relationship. At the lower end of the funnel, there is the output of what Hearst (1999) refers to as true *text mining*: finding novel information "nuggets", that is, finding or hypothesising knowledge which is not explicitly mentioned in the text. A textbook example of this kind of knowledge discovery are Swanson's experiments (Swanson, 1986). Based on a co-occurrence analysis of literature available at that time, he hypothesised that fish oil could be a treatment for Raynaud's disease which was experimentally confirmed later.

Concluding, biomedical IR is not only important for end-users but also an essential step in more sophisticated knowledge acquisition.

[5] `http://www.pubmed.gov`

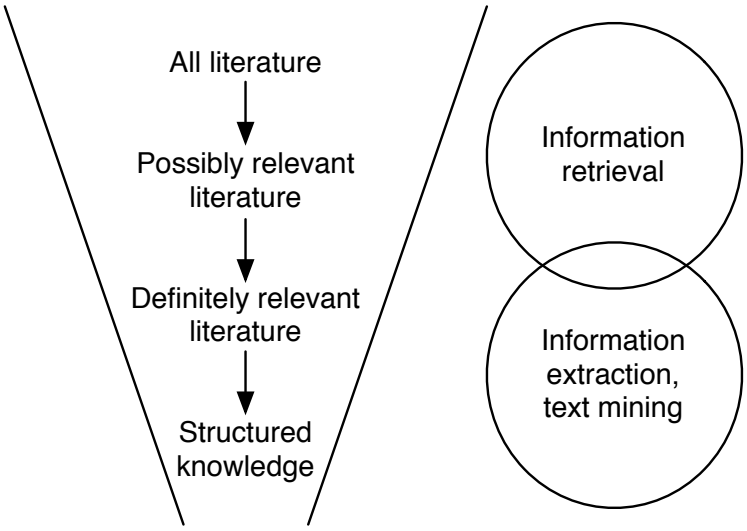


Figure 2.2: Funnel of knowledge acquisition and use, from Hersh 2009, p. 14.

2.2.3 Terminological challenges

One major challenge of working with biomedical literature is the variation and ambiguity of its terminology. Biological entities, such as diseases, genes, and organisms, are referred to in many different ways in texts. Automatically processing biomedical text suffers from lexical ambiguity (homonymy and polysemy) and synonymy (Krovetz, 1997; McCray, 1998; Nenadic et al., 2005; Hersh, 2009).

Homonymy refers to strings with different meanings. An example of a homonym is the abbreviation 'PSA' which can refer to 'prostate specific antigen', 'puromycin-sensitive aminopeptidase', 'psoriatric arthritis', 'pig serum albumin', or one of many more meanings found in the literature (Schijvenaars et al., 2005). Tuason et al. (2004) observed a considerable ambiguity across gene names from different organisms: between 1.87% and 20.3% of the names used for genes in one database also occurred in a database covering a different organism. Chen et al. (2005) measured a similar ambiguity of gene terms across 21 species: 15% of the investigated terms were used for genes in different organisms.

Polysemy refers to words which have multiple but related meanings (Manning and Schütze, 1999). The difference between polysemy and homonymy can be subtle and depends on the notion of relatedness used. For example, 'P450' can be regarded as a polyseme, since it is used to refer to many different Drosophila genes which belong to the same family of genes.

Synonymy refers to multiple words which have the same (or similar) meaning (Manning and Schütze, 1999). For example, 'Bovine Spongiform Encephalopathy', 'BSE', and 'mad cow disease' all have the same meaning.

The following causes for lexical ambiguity and synonymy can be indicated (Krauthammer and Nenadic, 2004; Nenadic et al., 2005).

Complexity of terminology Biomedical terminology is inherently complex. Multi-word terms are often used to indicate specific concepts. Nenadic et al. (2005) note that more than 85% of the terms encountered in the Genia corpus (consisting of 2000 abstracts) consist of more than one word. Rather than using these long forms throughout a

document, short forms are introduced throughout the text. These abbreviations often have different meanings in different contexts, such as 'PSA' mentioned before.

Lack of naming conventions There is a lack of naming conventions in biomedicine, causing great variations in names and spellings used. General English words or phrases are often used to indicate genes, such as 'hedgehog', 'bazooka' and even 'white'. Different abbreviations may be in use for the same term: the gene [prion protein] is abbreviated as both 'PRNP' ('PRioN Protein') and 'PRIP' ('PRIon Protein')[6]. The gene's product, the actual prion protein, is also referred to as 'prnp'. Chen et al. (2005) reported that authors frequently (75%) use terms other than the official gene symbol or full gene name in their publications.

Due to the compound nature of terms, spelling variations are frequently encountered. Superscript, hyphens ('-'), slashes, parentheses, brackets, numbers and additional letters are used to indicate variations of gene and gene product names. Rather than using 'PrnP', one might write 'Prn-P'. Krauthammer and Nenadic (2004) noted that even if naming conventions were adhered to, "there are still a huge number of documents containing "legacy" and ad hoc terms".

The lack of naming conventions is also illustrated by change in terminology (Krauthammer and Nenadic, 2004). Developments in biomedicine, such as newly discovered genes, treatments, and new types of diseases, result in a fast changing terminology. It is difficult to keep up with the latest terminology. For example, the flu causing the 2009 flu pandemic was first referred to as 'H1N1 influenza', which was quickly replaced by new terms such as 'pandemic H1N1/09 virus', 'pig flu', 'swine flu', and 'novel H1N1 virus'.

In section 2.3 we will discuss how retrieval systems cope with these terminological challenges.

2.2.4 Terminological resources

Several terminological resources are available to cope with the lexical ambiguity and synonymy present in biomedical terminology. They vary both in coverage and purpose. MeSH (described later), for example, has quite a broad coverage of the biomedical domain, but does not cover the gene names as well as Entrez Gene (a database with gene information). In general, they conveniently group the (synonymous) terms used to refer to a particular biomedical concept. One drawback they all have, however, is that as a result will always be behind the current terminology and they will remain incomplete.

In the following four blocks, frequently used terminological resources will be discussed: UMLS, SNOMED CT, MeSH and biological databases. MeSH will be covered in more detail, since it is used extensively throughout this thesis.

UMLS

The goal of the Unified Medical Language System (UMLS) is "to facilitate interoperable computer programs processing biomedical texts by integrating and distributing key terminology, classification, and coding standards" (McCray and Miller, 1998).

[6] http://www.genenames.org/data/hgnc_data.php?hgnc_id=9449

The Metathesaurus is the primary component of the UMLS. It is a large multi-lingual biomedical vocabulary, combining several resources containing biomedical and health related concepts in a uniform format. The Metathesaurus is organised by concepts which group alternative names and views from the different resources. Also the relationships between concepts is maintained from its originating resources. Since the information is composed from several resources, the Metathesaurus does not provide a single consistent view of the world.

The resources in the Metathesaurus include "many different thesauri, classifications, code sets, and lists of controlled terms used in patient care, health service billing, public health statistics, indexing and cataloging biomedical literature, and/or basic, clinical, and health services research"[7]. SNOMED CT and MeSH are parts of the Metathesaurus and are discussed in the following two blocks.

SNOMED CT

Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) is a multi-lingual, controlled vocabulary focused on medical terminology and covering most areas of clinical information. SNOMED CT is maintained and developed by the International Health Terminology Standards Development Organisation, an international organisation funded by national governments. Its purpose is "to provide a consistent way of indexing, storing, retrieving and aggregating clinical data from structured, computerised clinical records"[8]. The 2008 release of SNOMED CT consists of more than 311,000 hierarchically organised concepts.

MeSH

The Medical Subject Headings thesaurus is the controlled vocabulary maintained and developed by the United States' National Library of Medicine. It has been in existence since 1960 (see subsection 2.2.1) and it is used for "indexing, cataloging, and searching for biomedical and health-related information and documents"[9]. It consists of a large number of *descriptors*, also known as *main headings*, arranged hierarchically, which describe biomedical topics at different levels of granularity. Additionally, a larger thesaurus of *Supplementary Concept Records* is provided, which primarily lists chemicals and drugs.

Citations in MEDLINE are indexed with main headings, optionally combined with one or more *topic classifiers* or *subheadings*. Figure 2.3 shows part of a MEDLINE citation, indexed with MeSH terms. Each line starting with 'MH' contains a main heading, optionally followed by subheadings (separated by a '/'). For example, the main heading 'Encephalopathy, Bovine Spongiform' is assigned with two subheadings 'epidemiology' and 'transmission'. An asterisk ('*') is used to point out important MeSH terms. On average, around 9 MeSH descriptors are assigned to a MEDLINE citation (see Figure 2.4).

The thesaurus is updated on a yearly basis. Additional terms are suggested by experienced indexers as they are encountered in newly published literature. The 2008 thesaurus consists of 24,767 descriptors (with a Supplementary Concept Records thesaurus containing

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

[8] SNOMED Clinical Terms Overview 2008 (July 7, 2009) - Presentation by Kent Spackman, IHTSDO Chief Terminologist.

[9] `http://www.nlm.nih.gov/mesh/`

```
PMID -     9307349
DP   -     1997 Sep 24
TI   -     The risk of bovine spongiform encephalopathy ('mad cow disease') to human health.
AB   -     Some human cases of the transmissible neurodegenerative disorder Creutzfeldt-Jakob
disease recently seen in Great Britain are thought to have resulted from eating beef infected with
the agent of bovine spongiform encephalopathy. Reasons for and against this presumption are
explained, and the question of a similar situation occurring in countries other than Britain-in
particular, the United States-is discussed in terms of the existence of scrapie (in sheep) or
unrecognized bovine spongiform encephalopathy (in cattle), the practice of recycling nonedible
sheep and cattle tissue for animal nutrition, and precautionary measures already taken or under
consideration by government agencies
AD   -     Laboratory of Central Nervous System Studies, National Institute of Neurological
Disorders and Stroke, National Institutes of Health, Bethesda, Md 20892, USA. pwb@codon.nih.gov
FAU  -     Brown, P
AU   -     Brown P
LA   -     eng
PT   -     Journal Article
PT   -     Review
JT   -     JAMA : the journal of the American Medical Association
MH   -     Animal Feed
MH   -     Animals
MH   -     Cattle
MH   -     Creutzfeldt-Jakob Syndrome/epidemiology/transmission
MH   -     Encephalopathy, Bovine Spongiform/*epidemiology/transmission
MH   -     Europe/epidemiology
MH   -     Humans
MH   -     Meat
MH   -     Risk Assessment
MH   -     Sheep
MH   -     United States/epidemiology
SO   -     JAMA. 1997 Sep 24;278(12):1008-11.
```

Figure 2.3: Partial MEDLINE entry with assigned MeSH terms.

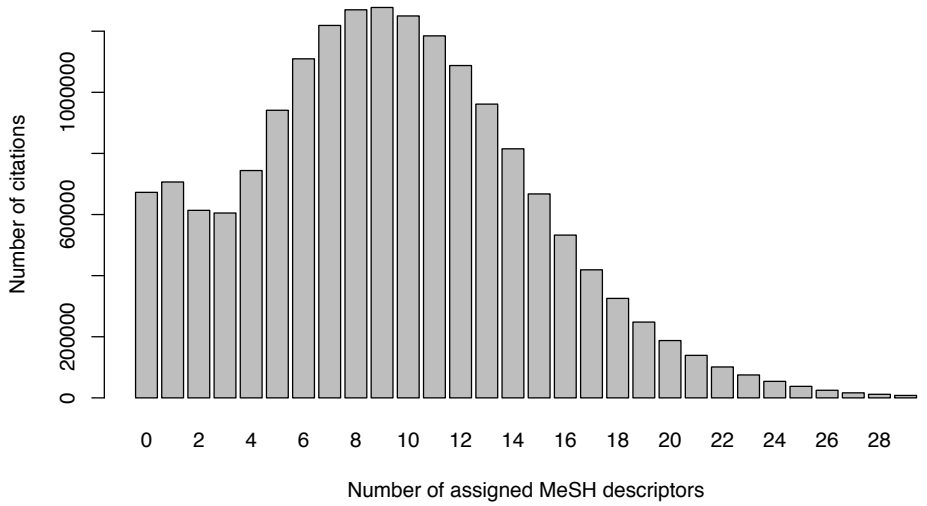


Figure 2.4: Histogram of number of MeSH descriptors assigned to each document (based on the MEDLINE 2008 baseline distribution).

181,069 entries); in contrast, the 1963 edition consisted of only 5,700 descriptors. Between 2006 and 2009, the thesaurus has grown by between 410 and 885 descriptors per year.

The organising principle behind MeSH is "to conceptually partition the literature" (Nelson et al., 2001). Each MeSH descriptor should "meaningfully distinguish literature". As an example, Nelson et al. (2001) argued that despite the fact that [DNA fingerprints] and [DNA fingerprinting] have a distinct meaning, the literature does not make this distinction sufficiently. In contrast, [Radiography] and [Radiographs], which differ in meaning in a similar way, are separated in two descriptors because they are described "sufficiently distinctly" in the literature. During indexing, (if available) the full-text of the article is used to determine the most appropriate terms. The most specific descriptor which covers the topic is used to index topics. Such a pragmatic approach has its advantages and disadvantages. On the one hand, no superfluous terms are added to the thesaurus and as a result the thesaurus concisely represents the current state of the literature. On the other hand, the approach can lead to inconsistencies when at a later time becomes evident that more specific descriptors are needed. Since the earlier citations are not reindexed with newly introduced descriptors, unpredictable behaviour may occur if the searcher is not aware of these changes. Searching with a new term will in such cases only retrieve newly indexed documents. The MeSH thesaurus includes notes to aid searchers. For example, the heading [Information Storage and Retrieval] notes that one should use [Information Systems] to search citations between 1982 and 1990. For untrained users these updates may pose a hurdle to using the right terms.

The types of parent-child relationships in MeSH are not strictly defined, but are often is-a and part-of relationships. An "aboutness" organising principle is used: if a search for one descriptor should also return documents with a second descriptor, then this second descriptor should be a child of the first (Nelson et al., 2001).

The thesaurus is structured as a directed acyclic graph (DAG), with a single root node counting the following 16 general categories: [Anatomy], [Organisms], [Diseases], [Chemicals and Drugs], [Analytical, Diagnostic and Therapeutic Techniques and Equipment], [Psychiatry and Psychology], [Phenomena and Processes], [Disciplines and Occupations], [Anthropology, Education, Sociology and Social Phenomena], [Technology, Industry, Agriculture], [Humanities], [Information Science], [Named Groups], [Health Care], [Publication Characteristics], [Geographicals]. The structure of the MeSH hierarchy is unbalanced: 65% of the descriptors are found in a group of three categories ([Chemicals and Drugs], [Diseases], and [Organisms]).

A single descriptor can be found at several locations in the structure. In fact, half of the 24,766 descriptors have two or more locations in the MeSH tree. An extreme example is the [WAGR Syndrome] descriptor which can be found at 19 different locations. The descriptor is used to denote a rare genetic syndrome which affects different parts of the body and can therefore be found as a child of different descriptors below the [Diseases] descriptor, such as [Neoplasms], [Nervous System Diseases], and [Eye Diseases].

What the set of children is for a descriptor depends on the location of this descriptor in the hierarchy. For example, the MeSH descriptor [Pain] occurs at five positions in the structure. For one location, the MeSH descriptor has 9 child descriptors, including [Back Pain], [Facial Pain], and [Headache]. For a second location, it has two child descriptors including [Arthralgia] and [Pain Threshold].

The MeSH organising principle in combination with "MeSH explosion", that is, query

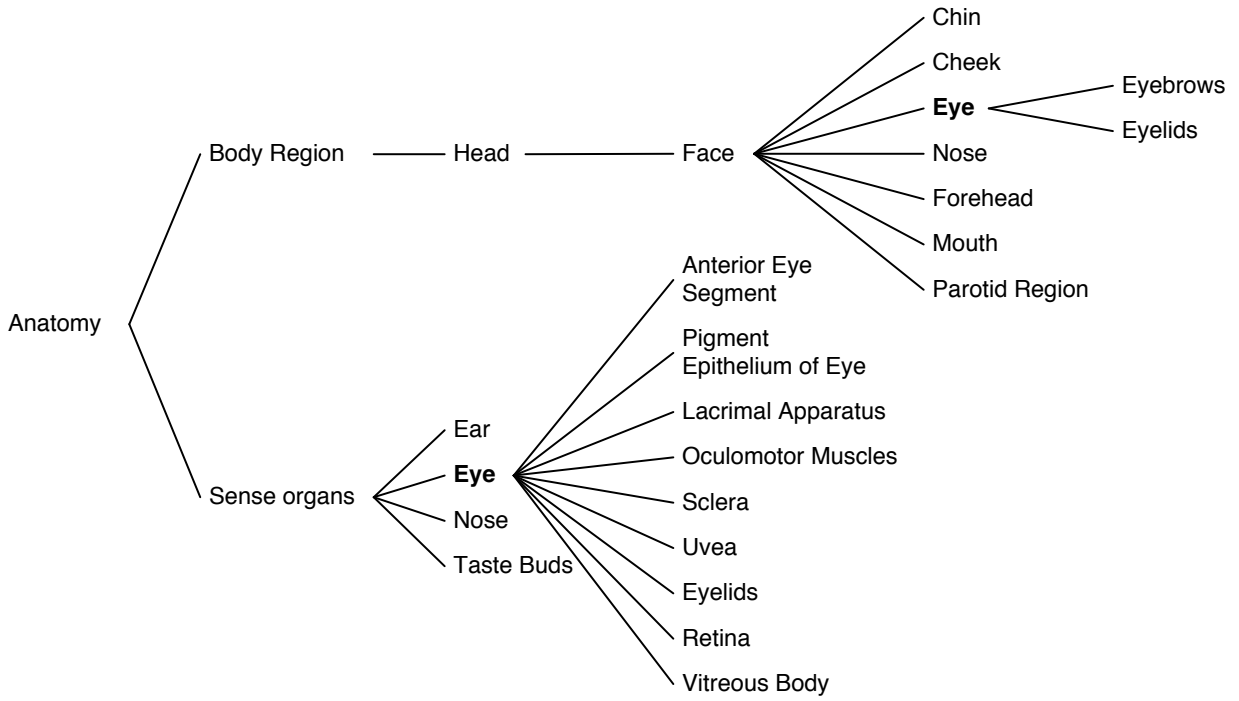


Figure 2.5: The MeSH descriptor [Eye] at two positions in the MeSH hierarchy with its siblings, ancestors, and children.

expansion with all child descriptors when only a single MeSH term is entered, can be a strong recall-enhancing device. However, it can lead to unexpected behaviour when a user is not aware of the special structure of MeSH. For example, consider a user searching for the MeSH descriptor [Sense Organs] which is automatically expanded with its children [Ear], [Eye], [Nose], and [Taste Buds]. In this part of the tree-structure the descriptor [Eye] has children relevant in the context of [Sense Organs] and the query is consequently expanded with descriptors such as [Retina] and [Sclera]. Following a different path in the MeSH hierarchy, the [Eye] descriptor also has the child descriptor [Eyebrows]. As a result, searching for the single descriptor [Eye] would also include documents indexed with [Eyebrows] (see Figure 2.5). So if a user truly wants to narrow down the results of the initial query [Sense Organs], he should issue a Boolean query 'Sense Organs AND Eye', which at a first seems counterintuitive since the latter is a child of the first.

Concluding, despite being incomplete, the MeSH thesaurus is a valuable resource for biomedical IR. It provides a high-level entry point to literature and its structure allows sophisticated queries and searching. However, for a novice user, the thesaurus structure and inconsistencies introduced by incremental changes, can mean that using the thesaurus can be difficult and give unexpected results.

Biological databases

The last two decades have been indicated as the start of the "omics-era": research areas such as genomics (studying the genome of organisms), proteomics (investigating proteins), and metabolomics (studying the chemicals in cellular processes) have received increased research interest. These research endeavours have led to the development of biological

databases in which the acquired knowledge is collected and linked. Some of these can be used as terminological resources as well.

The mission of UniProt is to "provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information."[10] The database consists of entries from Swiss-Prot and TrEMBL; the first is manually annotated and reviewed, the latter contains automatic annotations which have not been reviewed. The resource can be used for finding gene/protein synonyms.

The HGNC (HUGO Gene Nomenclature Committee)[11] is responsible for approving and storing human gene names. The HGNC database contains around 28,000 entries, primarily protein-coding genes submitted through the Human Genome Project. It stores the HGNC approved human gene names and symbols but also the name and symbol aliases used.

The National Center for Biotechnology Information' Entrez Gene provides access to gene information, with a focus of genomes that have been completely sequenced (Maglott et al., 2007). Also Entrez Gene is commonly used as a source of gene nomenclature.

As a last example of a terminological resource we mention ADAM[12] (Zhou et al., 2006a). This is a database of abbreviations of biomedical terms automatically extracted from MEDLINE citations. It covers frequently used single word abbreviations and their definitions (or long-forms). Zhou et al. reported a precision of 97.4% and noted that over a third of the found abbreviations are not reported in the UMLS or Stanford Abbreviation Database. The work clearly illustrates the incompleteness of manually maintained databases and the abundant use of abbreviations in biomedical literature.

2.2.5 Evaluation of biomedical IR

Two well-known IR evaluation initiatives in the biomedical (and health) domain are the OHSUMED and TREC Genomics evaluations. Following the Cranfield tradition, they provide a fixed document collection, information needs, and relevance judgements to carry out laboratory retrieval experiments.

OHSUMED

The OHSUMED test collection uses a clinically-oriented subset of 348,566 MEDLINE citations published between 1987 and 1991 (Hersh et al., 1994a)[13]. Novice physicians formulated 106 queries and provided information about their patient and information needs. The searches were carried out by two medical librarians and two physicians experienced in searching MEDLINE. The retrieved references were judged on relevance by another group of physicians.

TREC Genomics

The Genomics track of the Text REtrieval Conference was organised between 2003 and 2007 as a benchmark with the goal to investigate and improve biomedical and in particular

[10] http://www.uniprot.org
[11] http://www.hugo-international.org
[12] http://arrowsmith.psych.uic.edu
[13] http://ir.ohsu.edu/ohsumed/

genomics information retrieval (Hersh and Bhupatiraju, 2003; Hersh et al., 2004, 2005, 2006).

The track's task gradually evolved from ad hoc document retrieval in 2003 to full-text question answering in 2007. In the first three years a subset of MEDLINE citations was used for the evaluation; in the last two years a smaller collection of 162,259 full-text articles provided by Highwire Press was used.

The 2003 evaluation consisted of both a document retrieval and categorisation task[14]. The document retrieval task concerned finding information about a particular gene and was formulated as follows: "for gene X, find all MEDLINE references that focus on the basic biology of the gene or its protein products from the designated organism. Basic biology includes isolation, structure, genetics, and function of genes/proteins in normal and disease states". The use case is a "biological researcher or graduate student (i.e., someone who already has considerable general domain knowledge) who is confronted with the need to learn about a new scientific area quickly". GeneRIFs (Gene Reference into Function) were used as "pseudo-relevance judgements". GeneRIFs can be found in Entrez Gene and are short phrases stating the functions of a gene and link to the corresponding publication. As suspected by the organisers, the GeneRIFs relevance judgements showed to be incomplete. The 2003 collection has therefore not been reused often. From 2004 onwards, a pooling method (see subsection 2.1.4) was employed to obtain relevance judgements.

The 2004 and 2005 test collections used a considerably larger subset of MEDLINE as a document collection, consisting of 4,591,008 citations, most of them published between 1994 and 2004. The task remained ad hoc document retrieval and the topics were based on interviews with biologists. How the topics were provided was different, however. The 2004 query collection provides a *title*, *need*, and *context* section. The title provides a brief statement of the information need, the need section is more verbose. Finally, the context section provides additional contextual information, primarily used by the judges to make relevance judgements. Based on an analysis of the topics of the 2004 task, the 2005 topic set was categorised in five *generic topic types*, which indicate what the documents should describe: 1. standard methods or protocols; 2. the role of a gene involved in a disease; 3. the role of a gene in a biological process; 4. interactions between genes in the function of an organ or disease; 5. mutations of a gene and its impact.

The 2006 and 2007 evaluations matured to passage retrieval tasks. Rather than using a collection of citations, a smaller collection of 162,259 full-text journal articles was used. The 2006 topics were derived from the 2005 topics, some of them changed substantially however. Topics were provided as questions following four patterns. 1. What is the role of *gene* in *disease*? 2. What effect does *gene* have on *biological process*? 3. How do *genes* interact in *organ function*? 4. How does a mutation in *gene* influence *biological process*? The systems were to return passages from the full-text, indicated by the offset in the article and the number of characters which should be included in the passage. The retrieval performance was measured in terms of mean average precision (MAP) at the document, passage, and aspect level. Document MAP was calculated as explained in subsection 2.1.4. The passage MAP was based on the character-based overlap between passages returned by the system and the passages marked as relevant. During construction of the relevance judgements, the judges were asked to categorise the passages found for a topic into aspects. The aspect MAP indicated to which extent all aspects are retrieved by the system. The 2007 topics still

[14]The categorisation task will not be discussed here, see Hersh and Bhupatiraju (2003) for more information.

Table 2.1: TREC Genomics benchmark collections.

Year	Task	Topics	Doc. collection	Coll. size
2003	Ad hoc (GeneRIF)	50	MEDLINE (2002-2003)	525,938
2004	Ad hoc (TNC)	50	MEDLINE (1994-2004)	4,591,008
2005	Ad hoc (topic templates)	50	MEDLINE (1994-2004)	4,591,008
2006	QA (topic templates)	28	Highwire Press	162,259
2007	QA (typed list)	36	Highwire Press	162,259

required passages to be retrieved, but in this case the topics were typed list questions: the retrieved passages should contain named entities of the requested type. The types ranged from genes and proteins to drugs, antibodies, and molecular functions.

Table 2.1 summarises some of the features of the TREC Genomics test sets. The sets of queries can be found in Appendix A.

2.3 Coping with terminology

In the previous sections, the terminological challenges in biomedical IR have been discussed. Lexical ambiguity and homonymy aggravate the vocabulary mismatch problem of word-based retrieval systems (Furnas et al., 1987): documents are indexed with terms different from those that users actually use to find them. A large range of approaches has been proposed in the past to cope with these phenomena.

In the following three subsections a broad overview will be provided of extensions to automatic, unigram (single "word"), bag-of-words indexing and searching. Firstly, methods will be discussed which do not change the terms in the original query, but impose a structure on it or on the documents it matches. Secondly, query expansion methods will be discussed which do actually add terms to the query issued to the system, with various sources to choose terms from. Thirdly, we will look at methods which impose a meta-structure at the collection level, this can for example be achieved by grouping related documents. It should be noted that this categorisation is not strict; methods have been proposed which are in fact a combination of these approaches.

2.3.1 Incorporating term dependencies

Imagine a user interested in 'cell division'. Neither of the words 'cell' or 'division' are informative on their own, but the phrase indicates something quite specific. By incorporating the dependence between the two terms the performance of a bag-of-words retrieval system might be improved.

When an index with positional posting lists is used, that is, an index which stores the positions of terms in the document, matching can be restricted to documents containing the query terms in a particular proximity and optionally in a particular order. In the case of searching for a phrase, the proximity between the terms should be one and the order of the terms should be kept. Alternatively, a larger matching window could be used (for example a proximity of 10) without taking word order into account. The latter approach would, for

example, also match 'division of the cell' appearing in a document. One drawback of this approach is that merging the posting lists of two (or more) terms during searching can be slow, especially when many phrases are combined.

Term proximity can be directly offered to the user by integrating it in the query language: the user can indicate which phrases should be searched for. Term dependencies can also be integrated "under the hood": the system analyses the query and determines plausible term dependencies and automatically takes them into account during ranking. As with boolean operators, offering a proximity operator gives the user explicit control. Nevertheless, it requires the user's understanding of both the operator and the collection being searched.

Automatically detecting term dependencies and incorporating them in a theoretically sound and empirically effective manner in the retrieval model has shown to be possible but challenging (Xu and Croft, 1996; Mitra et al., 1998; Metzler and Croft, 2005; Bai et al., 2007, 2005). Most models rely on term statistics, which in the case of single words (or word stems) can be estimated on a large document collection with reasonable correctness. Estimating parameters of more complex models, such as word bigrams or trigrams, often suffers from data sparseness: less information is available for determining the importance of word combinations than for single words.

2.3.2 Query reweighing and expansion

After initial query formulation, there is a large range of techniques to update the query, automatic updating, manual updating, and combinations thereof. Query expansion, adding terms to the user's original query, is primarily a recall enhancing device: by adding more related terms, more related, and hopefully relevant, documents can be matched and retrieved.

A common problem of query expansion is *query drift*: expansion of a query can lead to the overemphasis of a particular aspect of the query. The retrieved documents may therefore drift towards a particular aspect. For example, consider the query 'What is the role of PrnP in mad cow disease?', which contains the aspects 'PrnP' and 'mad cow disease'. Expanding the query with many synonyms of either aspect might lead to neglect of the other aspect in the retrieved documents. Structured queries can be used to group synonyms to prevent this from happening.

In general, a distinction can be made between query expansion using external sources, expansion based on analysis of the collection, and expansion taking into account the user's context.

Expansion using external resources

Terminological resources such as MeSH, UMLS, and ADAM, can provide a valuable source of synonyms for the terminology found in the query. An important issue here is mapping the query to the appropriate entries in the resource. The original query might contain ambiguous terminology, such as the abbreviation 'AD' which is often used to denote 'Alzheimer disease', but is also used for 'atopic dermatitis'. Expanding such an abbreviation with synonyms of the incorrect sense will inevitably lead to degraded retrieval performance. Even when the query is unambiguous, expanding it with ambiguous terms might lead to undesirable query drift. When the original query contains 'atopic dermatitis', expansion with 'AD' might

not be appropriate either since it is more commonly used to indicate 'Alzheimer Disease'. Even manual expansion of queries using controlled vocabularies does not always lead to improvements in retrieval effectiveness (Voorhees, 1994).

In the biomedical domain mapping free text to biomedical entities has received considerable attention. The MetaMap program, for example (Aronson, 2001), maps free text to concepts in the UMLS thesaurus. A second well-known example is the search interface of PubMed, which automatically detects MeSH terms in the query and expands the search with both these MeSH terms and text phrases with MeSH term synonyms[15]. This Automatic Term Mapping can be viewed as a combination of an integration of external resource and integrating term dependencies.

Krauthammer and Nenadic (2004) distinguish between term recognition, term classification, and term mapping. Term recognition is the process of finding entities in text; term classification indicates the type of entity (for example disease or gene) and term mapping is the process of linking the text to a unique identifier in a biological database. The BioCreAtIvE challenges (Morgan et al., 2008) have shown that gene and protein recognition and classification still perform worse than recognising named entities in news text. The top-performing system at BioCreAtIvE has a recall and precision of 0.833 and 0.789, respectively. Named entity recognition in English texts is considered an easier task, where numbers around 0.89 on both precision and recall are reported (Tjong Kim Sang and De Meulder, 2003).

It is actually debatable whether ambiguous terminology in queries indeed is an issue for IR (Sanderson, 2000). From experiments in an artificial setting, Sanderson (1994) concluded that word sense disambiguation is only of use in a retrieval context if the disambiguation is very accurate or if the queries are short. Often, explicit disambiguation is not necessary, since the ambiguous words are disambiguated by the context provided by the other words in the query (the query collocation effect). However, for highly ambiguous abbreviations in the biomedical domain, Stokes et al. (2007) argue that this query collocation effect is not strong enough. In these cases retrieval may require a (manual) mapping to an unambiguous conceptual representation.

When the terminological resource includes structure, this can also be exploited for query expansion. In the case of MeSH for example, the terms do not only provide synonyms, but children in the MeSH structure can be used for MeSH term "explosion".

It should be noted that terminological resources are often not made with word-based IR in mind. Many MeSH terms are not likely encountered in documents, such as 'Encephalopathy, Bovine Spongiform' (rather than 'Bovine Spongiform Encephalopathy'). They therefore may require additional preprocessing before they can be used.

Expansion based on the collection

The collection itself can also be used as a source for expansion terms, where a distinction can be made between *global* and *local* analysis (Xu and Croft, 1996).

During a global analysis, the collection as a whole is used to determine terms for expansion (Spärck Jones and Jackson, 1970; Salton, 1971; Jing and Croft, 1994). The association hypothesis suggests: "if an index term is good at discriminating relevant from irrelevant documents then any closely associated index term is also likely to be good at

[15] `http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#AutomaticTermMapping`

this" (van Rijsbergen, 1979). The collection can, for example, be used to build a similarity thesaurus which groups frequently co-occurring words or phrases (Jing and Croft, 1994; Qiu, 1995). In contrast to an ordinary thesaurus-based approach, the vocabulary is entirely based on terminology actually used. Such an automatically built thesaurus may contain errors since co-occurrence of terms is not a guarantee for actual similarity between terms.

Local analysis uses the top-ranked documents obtained using the original query to reweigh the query terms and add expansion terms (Rocchio, 1971). Local analysis is often combined with *relevance feedback*, that is, feedback from the user indicating which of the initially retrieved documents are relevant (or not). When these judgements are not available, the top-ranked documents can be assumed to be relevant, a technique called *blind* or *pseudo-relevance feedback*. Manning et al. (2008) noted that relevance feedback is only beneficial when relevant documents are similar to each other. Obviously, it requires the initial query to find relevant documents. A number of situations can be mentioned in which relevance feedback does not work in general (Manning et al., 2008): in the case where the collection contains subsets of relevant documents using a different language; in the case that the relevant documents are "inherently disjunctive": the query combines different topics[16]; and when the query is very general.

The user can also give feedback on the term rather than the document level. Rather than picking relevant documents, the system can suggest query terms either after the first retrieval run or even while the user is formulating his query (White and Marchionini, 2007).

Using context

A third possibility to cope with ambiguous terminology is by incorporating more information about the user's context, such as the user's interests and previous information needs. Korfhage (1984), for example, experimented with building user profiles and argued that having such a profile could easily "exclude a large portion of the document collection from consideration". The eventual query processed by the system would be a combination of the profile and the query issued to the system. The profile can be specified by the user or automatically derived from past queries, browsing history, and locally stored documents and e-mail (Dumais et al., 2003; Teevan et al., 2005; Chirita et al., 2007).

2.3.3 Adding (meta-)structure

Lastly, adopting some kind of structure on top of the documents can be used to circumvent ambiguity. By grouping documents based on all their words, the different senses of words can be distributed over different clusters of documents.

The clustering hypothesis supports this view: "closely associated documents tend to be relevant to the same requests" (van Rijsbergen, 1979). Rather than retrieving single documents, groups of associated documents can be retrieved, or can be used to update the query (Kurland and Lee, 2009).

Latent Semantic Indexing (Deerwester et al., 1990) can also be regarded as a way to group closely associated documents. The hidden semantic structure of the collection is obtained by carrying out dimensionality reduction on the term-document matrix. Documents

[16] Manning et al. mention as an example 'Pop stars who once worked at Burger King'

and queries are then represented as latent concepts. A drawback is that the approach is computationally demanding and the resulting index is difficult to comprehend.

2.4 Experiences in concept-based biomedical IR

In this subsection a number of previous experiments in health and biomedical IR will be highlighted, to illustrate the developments in automatic vs. manual indexing and the attempts to integrate terminological resources in IR.

Table 2.2 summarises the work described. In the respective columns is indicated which resources were integrated, how the mapping to the entries this resource was accomplished, what additional operations were used to integrate it into the retrieval model, the test collection the approach was tested on, and the primary conclusions regarding the use of terminological resources.

We first discuss the related research one by one. After that, the approaches are categorised and compared.

Related research

Salton (1972) compared conventional biomedical indexing used in MEDLARS to the automatic indexing method used by the SMART ranked retrieval system. Based on an experiment with a document set of 450 documents and a query set of 30 queries, Salton concluded that "no technical justification exists for maintaining controlled, manual indexing in operational retrieval environments". Relevance feedback especially turned out to improve the retrieval effectiveness of the SMART retrieval system.

More than two decades later, Hersh and Hickam (1995) summarised work done in the SAPHIRE (Semantic and Probabilistic Heuristic Information Retrieval Environment) project, focusing on bibliographic search in the clinical domain. Similarly, they conclude that "studies suggest that the incremental benefit of human indexing as measured by retrieval performance is small". Novice users achieved better search results using the SAPHIRE free-text search than by using Boolean queries. Trained librarians, however, performed better using Boolean queries. Initial approaches with the UMLS Metathesaurus were less successful, however: except for some individual cases, automatic "concept-based" indexing using UMLS terms performed poorly in comparison to automatic single word indexing (Hersh et al., 1994b). One major reason was that not all information needs could be expressed in terms of UMLS concepts. A second attempt to improve text-based retrieval using the UMLS Metathesaurus led to the same conclusion (Hersh et al., 2000): "thesaurus-based query expansion causes a decline in retrieval performance generally but improves it in specific instances".

Reference	Resource(s)	Mapping	Integration	Model	Collection	Conclusions
Salton (1972)	MeSH	manual	n.a.	Boolean	450 docs, 30 queries	Compared automatic word-based indexing to controlled vocabulary indexing and concluded that there is "no technical justification for [indexing with a] controlled vocabulary".
Hersh et al. (1994b)	UMLS	string matching	n.a.	TF.IDF	HERSH[1]	Compared automatic concept-based indexing and search to MEDLINE search: One fourth of the queries could not be accurately represented in terms of UMLS concepts.
Hersh and Hickam (1995)	MeSH, UMLS	automatic, manual	n.a.	TF.IDF	HERSH[1]	"the incremental benefit of human indexing is [...] small." In the aggregate, concept-based automatic indexing appeared not to offer any benefit over the use of single words.
Srinivasan (1996b)	MeSH	feedback, statistical thesaurus	n.a.	VSM	HERSH[1]	Recommended query expansion based on pseudo-relevance feedback for adding MeSH search terms. "MeSH terms are important for retrieval."
Aronson and Rindflesch (1997)	UMLS	string matching	structured queries, phrases, weighting	prob. model	HERSH[1]	"Query expansion [...] is an effective method of enhancing retrieval effectiveness and compares favourably to document feedback."
Hersh et al. (2000)	UMLS	manual	n.a.	VSM	OHSUMED	No improved performance was observed when queries were expanded using thesaurus relationships.
Hersh et al. (2003)	multiple[2]	semi-automatic	phrases, weighting	VSM	TREC '04	"Query expansion using information extracted from online databases failed to improve MAP."
Guo et al. (2004)	UMLS (filtered)	string matching	structured queries, weighting	TF.IDF	TREC '04	Marginal increases were observed in both precision and recall.
Kraaij et al. (2004)	MeSH	feedback	fusion	LM	TREC '04	Marginal performance increases were observed.
Büttcher et al. (2004)	AcroMed, euGenes, LocusLink	string matching	structured queries, expansion validation	prob. model	TREC '04	"automatic acronym expansion [...] almost always improves the performance" Generation of lexical variants and validation of expansion terms through feedback proved to be useful.
Ruiz (2005)	UMLS	string matching	n.a.	VSM	TREC '05	Query expansion with UMLS could not improve over an unexpanded baseline.
Ando et al. (2005)	multiple[3]	string matching	weighting	prob. model	TREC '05	Query expansion with synonyms shows inconclusive changes in performance.
Pirkola (2005)	Entrez Gene, MeSH	manual	structured queries proximity	prob. model	TREC '05	The manually expanded queries perform worse than the automatic run.
Zhou et al. (2006b)	UMLS	string matching	topic signatures	LM	TREC '04, '05	Topic signatures based on combinations UMLS concepts were useful to improve retrieval.

Reference	Resource(s)	Mapping	Integration	Model	Collection	Conclusions
Si et al. (2006)	AcroMed, LocusLink, UMLS	string matching	structured queries weighting	LM	TREC '06	"Query expansion based on external biomedical resources is an effective technique"
Zhou and Yu (2006); Zhou et al. (2007)	MeSH, Entrez Gene, ADAM	string matching from template	concept-based	prob. model	TREC '06	"Any available type of domain-specific knowledge improved the performance in passage retrieval."
Dorff et al. (2006)	HUGO, MeSH	manual	n.a.	prob. model	TREC '06	"synonym expansion[...] is marginally beneficial."
Camous (2007)	MeSH	feedback	fusion	prob. model	TREC '06	MeSH-based feedback only marginally improved retrieval performance.
Stokes et al. (2009)	multiple[4]	template, string matching, ATM	structured queries, weighting, phrases	prob. model	TREC '06	The choice of ranking algorithm is most important "Phrase-based querying is essential" Expansion based on ontologies is more effective than based on corpus statistics. The choice of knowledge source is not so important.
Lu et al. (2009)	MeSH	string matching	fusion	Boolean, TF.IDF	TREC '06, '07	Expansion with MeSH can improve effectiveness.
Abdou and Savoy (2008)	MeSH	feedback	fusion	multiple	TREC '06	"retrieval performance can improve from 2.4% to 13.5%, depending on the underling IR model."

[1] The HERSH corpus consists of 2,334 MEDLINE citations and 75 queries from a set originally composed by Haynes et al. (1990)
[2] including LocusLink, SwissProt and Gene Ontology
[3] LocusLink, Gene Ontology, MeSH, SwissProt
[4] ADAM, Entrez Gene, GO, Hugo, Metathesaurus, MeSH, OMIM, SNOMED CT, UMLS and UniProt

Table 2.2: Experiments in using terminological resources in biomedical IR.

Aronson and Rindflesch (1997) also used the UMLS Metathesaurus for expanding queries on the same collection but came to a different conclusion. They concluded that MetaMap, a program to map text to UMLS concepts, could be used effectively to expand queries with UMLS concepts. The differences with Hersh et al. (2000) might be explained by differences in the mapping process and the style of integration. Firstly, Aronson and Rindflesch used MetaMap for mapping the query to UMLS concepts; MetaMap might perform better at mapping the documents and queries to UMLS than the approach used by Hersh et al. Secondly, the integration of the obtained concepts is different: Aronson and Rindflesch used structured queries, incorporating phrases, and reweighing expansion terms in an ad hoc fashion, whereas Hersh et al. use an unstructured query model.

Braun (2008) used the UMLS Metathesaurus to manually translate templated information needs for finding medical literature. She concluded that an automatic translation mechanism was not sufficiently effective in comparison with a manual approach.

Srinivasan (1996a) and Lam et al. (1999) positively confirmed the use of MeSH terms for improving retrieval effectiveness. Pseudo-relevant documents were used to gather MeSH terms to expand the original textual query. A second round of (text-based) pseudo-relevance feedback was shown to give the best retrieval performance.

During the TREC Genomics workshops and following work, it became clear that query expansion using a controlled vocabulary could work beneficially if implemented carefully (Hersh et al., 2007). A large number of variables have to be taken into account: the choice of controlled vocabulary, the strategy to do lookup and select expansion terms, the integration into the original query, the retrieval model etcetera. Given the complexity and variety of approaches used, it is difficult to make general conclusions on which approaches do or do not work.

Zhou et al. (2006b) proposed the use of "topic signatures", based on major and minor UMLS concepts detected in text. By searching and indexing using combinations of automatically detected UMLS terms, improved retrieval effectiveness was reported over a text-based baseline.

Camous (2007) used a similar approach to Srinivasan's (1996a) to obtain a MeSH-query, but in contrast merged the two ranked document lists, that is, one from the original textual query and one from the MeSH-query, to determine the final ranking. Kraaij et al. (2004) pursued a similar approach, using a separate text and MeSH-index, but reported that merging only gave marginal improvements.

Stokes et al. (2007) investigated handling abbreviations in the biomedical domain, and in particular on the TREC Genomics collections. They reported that for highly ambiguous abbreviations the query collocation effect is not strong enough to prevent retrieving documents with an incorrect sense. They suggested handling abbreviations at indexing time rather than querying time.

Stokes et al. (2009) also investigated the use of different query expansion sources for the TREC Genomics 2006 task. They concluded that the most important component for good performance is not the controlled vocabulary, but rather the retrieval model. The retrieval model they used preferred documents covering all query aspects rather than documents providing a single query aspects in more depth. Moreover, a normalisation technique was used which ensured expansion terms were not too influential.

Lu et al. (2009) investigated the added value of MeSH term-expansion for actual PubMed users (which retrieves documents in descending publication order). They concluded that

MeSH term-expansion did indeed improve recall, but that actual users would not notice much of a difference because the precision on the first result pages (rank precision at 5, 10, and 20) remained the same.

Ide et al. (2007) described ESSIE, a highly complex biomedical search system developed by National Library of Medicine. It combines several approaches, such as phrase-based search, query expansion using a controlled vocabulary, and pseudo-relevance feedback. To cope with spelling variations, a large search expansion tree is built. Despite all of the enhancements and an outstanding performance on an interactive run, a fully automatic evaluation showed performance similar to text-only retrieval systems.

Abdou and Savoy (2008) investigated different retrieval models including query likelihood, Okapi, and different TF.IDF weighting schemes and the usefulness of MeSH for expansion. They reported on improvements between 3% and 14% mean average precision, depending on the retrieval model.

Approaches to integration

The following two major approaches to integrating terminological resources can be distinguished.

Integration on the query side The first group of approaches primarily focusses on optimising the query. A standard word-based representation, sometimes in combination with an already available concept-based representation, is used for the documents. The next steps are commonly followed. In a first step, the text-based representation is in some way mapped to a conceptual representation. Commonly used approaches are: 1. String matching or dictionary lookup, optionally taking into account that the topics use templates (Aronson and Rindflesch, 1997; Büttcher et al., 2004; Ruiz, 2005). 2. Retrieval feedback, which uses the available conceptual document representations of feedback documents for mapping the query to concepts (Srinivasan, 1996b; Kraaij et al., 2004; Camous, 2007). 3. Manual mapping of the query text to the entries in the terminological resource (Salton, 1972; Hersh et al., 2000; Pirkola, 2005; Braun, 2008). After that, the mapping to concepts is used to update the textual query. Several variations are reported. Firstly, often structured queries are used to group synonyms and related terms (Hersh et al., 2003; Pirkola, 2005; Si et al., 2006). Optionally, the expanded query terms receive a different weight than the original query terms (Aronson and Rindflesch, 1997; Hersh et al., 2003; Ide et al., 2007). Secondly, an option is to use phrase or proximity operators to group multi-word terms (Hersh et al., 2003). A few attempts were made to use only the mapped conceptual representation for searching, but these approaches performed poorly.

Integration on both the query and document side The second group of approaches also uses the terminological resource to extend the document representation. Zhou et al. (2006b), for example, explicitly mapped the textual representation of the documents to "topic signatures", based on concepts in the UMLS. By means of pseudo-relevance feedback a topic signature representation is also obtained for the query. An early attempt to map both queries and documents to conceptual representations by Hersh and Hickam (1995), showed that matching only in this conceptual representation failed to improve over word-based retrieval.

Quite a few researchers report on using special lexical analysis or term normalisation to improve the textual representation of the queries and documents (for a more comprehensive discussion see chapter 3).

It is difficult to explain the reported differences. The retrieval systems are compared as black boxes and often use similar steps and resources during the process. We indicate five possible causes to explain the reported differences.

Baseline Improving a weak baseline is obviously easier than competing with a strong one. (Unintentional) suboptimal parameter settings may heavily impact retrieval performance. Training, or in the worst case over-fitting the parameters of the proposed method may lead to the false conclusion that the new method indeed is better.

Retrieval model Some of the reported retrieval models are designed for the task at hand. The systems used by Stokes et al. (2009) and Zhou et al. (2007), for example, explicitly take into account that the 2006 topics ask for two relatively disjoint aspects and their retrieval model takes into account that both should appear in matching documents. As a result, these models are less sensitive to query drift when one particular aspect has many synonyms which are included during expansion.

Terminological resource Not only the choice, but also the preprocessing of the resource may affect the observed performance. Ando et al. (2005), for example, reported on filtering very ambiguous terms and pruning parts of the terms in the terminological resource. Depending on the method of integration, expanding with ambiguous terms may lead to undesired query drift.

Choice of integration Strongly related to the previous two causes is the way in which the mapped concepts are integrated into the retrieval model. Some approaches use too long and strict phrases for matching, resulting in poor performance. Others seek the solution in reweighing expansion terms in an ad hoc fashion. Choosing the right weighing scheme may affect how well the approach performs.

Choice of mapping process The choice of mapping process plays an important role in the effectiveness of integrating terminological resources. Strict string matching may only map to concepts found explicitly in the query; in contrast, a feedback approach will also map to indirectly related concepts. Again, this choice is likely to interact with the other components of the retrieval model. One can imagine that unweighted synonym expansion is more likely to work for strict matching than for the feedback approach.

These experiences show that coping with the terminological issues of biomedical IR is far from trivial. On the one hand, free text indexing and retrieval can be more straightforward and even more effective than controlled vocabulary search. On the other hand, controlled vocabulary search may offer more control and flexibility to professional searchers. Integrating "conceptual knowledge" has shown contradictory effects on retrieval effectiveness, most likely caused by differences in experimental settings. The choice of concept vocabulary, retrieval model, type of integration, and details such as document preprocessing may explain these differences. Therefore, we have taken up the challenge to investigate the robust and principled integration of conceptual knowledge in IR.

2.5 Chapter summary

In this chapter, the relevant background to this thesis was outlined. A basic introduction into information retrieval was provided for readers with a biomedical background; an introduction into information retrieval in the biomedical domain was provided for readers with an IR background.

The need for biomedical IR systems was highlighted, both in the context of end-users but also as a required component for biomedical text mining. As a major challenge for biomedical IR the ambiguity and complexity of biomedical terminology was examined. Several high-level approaches to cope with these challenges were discussed, followed by a discussion of earlier experimental work in the biomedical domain. Terminological resources, such as thesauri and controlled vocabularies, contain domain knowledge which can be used to reduce vocabulary mismatch problems. It is evident that integrating these terminological resources still remains a challenge in biomedical IR. Integration of terminological resources has the potential to be beneficial, but successful integration is far from trivial. The added value of integration is not always clear and is often blurred by other techniques applied at the same time, such as term weighting, structured queries, and adapted retrieval models.

To determine the added value of the integration of terminological resources first a solid text-only baseline needs to be established. To achieve this, different heuristics to cope with the spelling variations in biomedical terminology will be examined in the following chapter. In chapter 4 different facets of translating textual representations into conceptual representations will be examined, including document classification, query classification, and determining the relatedness between concepts. In chapter 5 a translation-based retrieval framework will be proposed in which terminological resources are integrated in a transparent way.

Chapter 3

Word-based Biomedical IR

"The word of man is the most durable of all material."

Arthur Schopenhauer

Parts of chapter work have been published in Trieschnigg, Kraaij, and de Jong (2007).

In order to obtain indexing terms for automatic full text retrieval, documents have to be preprocessed. *Document preprocessing* determines how the text is converted to index terms and hence determines the document representation used by the retrieval system. In the previous chapter we explained that one of the terminological challenges of biomedical IR is spelling variation. If these variations are not handled, mismatches may occur when query terms are spelled differently from the equivalent terms extracted from the documents. To some extent, document preprocessing can reduce the mismatch of spelling variations by normalising word variants to the same index terms. In this chapter we will compare sixteen different preprocessing heuristics that we expect to improve word-based biomedical IR. The goal is to achieve a word-based baseline which handles spelling variations well. In subsequent chapters, word-based retrieval based on this representation will be further improved with information from terminological resources. We will answer RQ1 posed in chapter 1.

RQ1: *How can the effectiveness of word-based biomedical information retrieval be improved using document preprocessing heuristics?*

Figure 3.1 shows the approach investigated in this chapter schematically. For both the queries and the documents only a text-based representation is used for matching. For now, the concept-based representation introduced in chapter 1 is not taken into account.

The structure of this chapter will be as follows. First, in section 3.1, the preprocessing steps for automatically obtaining index terms from biomedical text will be discussed. In section 3.2, four questions regarding document preprocessing will be raised. In section 3.3, we will describe the experimental setup for answering these questions, followed by the results of these experiments in section 3.4. The chapter will be concluded by a discussion and a conclusion in sections 3.5 and 3.6, respectively.

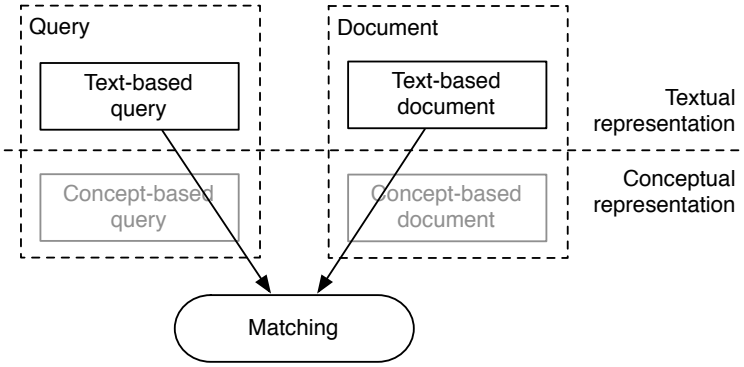


Figure 3.1: Using text-only representations for matching queries to documents.

3.1 Steps in document preprocessing

Automatic full text indexing requires a process to determine the index terms for a document automatically. The same process is used to determine the search terms in a query posed in natural language. These terms are subsequently used to match documents to queries.

Typically, the following four steps can be distinguished in this process (Baeza-Yates and Ribeiro-Neto, 1999; Manning et al., 2008).

Decoding During decoding, the original digital representation in, for example, HTML or PDF is converted to plain text.

Tokenization The plain text is split into words during tokenization. Tokenization handles character case, punctuation, and the like.

Stop-word removal Optionally, non-informative words such as 'the', 'for', and 'from' are removed during stop-word removal.

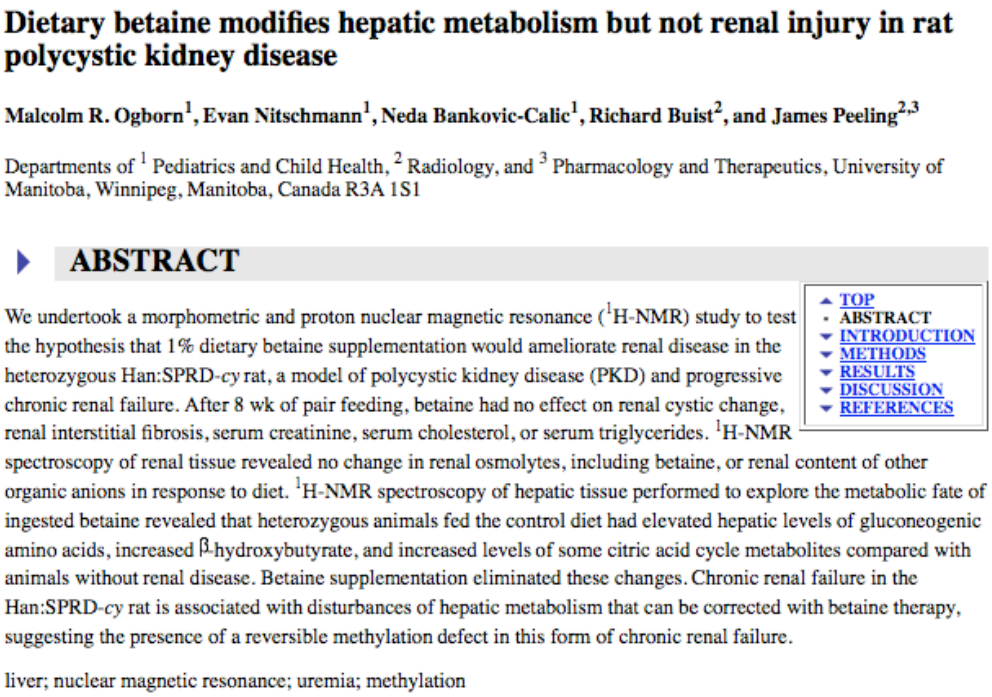
Stemming or lemmatisation A second option is to stem or lemmatise the words found in the text. Both approaches conflate words to a root form, for instance 'cells' to 'cell', to reduce vocabulary size and improve matching.

Figure 3.2 shows an example document in HTML before decoding. Figure 3.3 illustrates the intermediate output of the individual processing steps. In subsections 3.1.1 to 3.1.4 these preprocessing steps will be discussed in more detail.

3.1.1 Document decoding

Textual documents are available in various digital formats, varying from HTML to PDF and Microsoft Word. To obtain a plain text representation for these documents, the files need to be *decoded*. During this process, information such as layout and formatting is discarded. It is also possible to decide to remove particular document elements, such as figures or tables and their captions.

When preprocessing the MEDLINE citation database for searching, document decoding is not an issue: the database provides bibliographical entries in plain text, conveniently grouped into fields such as 'title', 'abstract', and 'authors' (see Figure 2.3 on page 24).



Dietary betaine modifies hepatic metabolism but not renal injury in rat polycystic kidney disease

Malcolm R. Ogborn[1], Evan Nitschmann[1], Neda Bankovic-Calic[1], Richard Buist[2], and James Peeling[2,3]

Departments of [1] Pediatrics and Child Health, [2] Radiology, and [3] Pharmacology and Therapeutics, University of Manitoba, Winnipeg, Manitoba, Canada R3A 1S1

ABSTRACT

TOP
ABSTRACT
INTRODUCTION
METHODS
RESULTS
DISCUSSION
REFERENCES

We undertook a morphometric and proton nuclear magnetic resonance (^{1}H-NMR) study to test the hypothesis that 1% dietary betaine supplementation would ameliorate renal disease in the heterozygous Han:SPRD-*cy* rat, a model of polycystic kidney disease (PKD) and progressive chronic renal failure. After 8 wk of pair feeding, betaine had no effect on renal cystic change, renal interstitial fibrosis, serum creatinine, serum cholesterol, or serum triglycerides. ^{1}H-NMR spectroscopy of renal tissue revealed no change in renal osmolytes, including betaine, or renal content of other organic anions in response to diet. ^{1}H-NMR spectroscopy of hepatic tissue performed to explore the metabolic fate of ingested betaine revealed that heterozygous animals fed the control diet had elevated hepatic levels of gluconeogenic amino acids, increased β-hydroxybutyrate, and increased levels of some citric acid cycle metabolites compared with animals without renal disease. Betaine supplementation eliminated these changes. Chronic renal failure in the Han:SPRD-*cy* rat is associated with disturbances of hepatic metabolism that can be corrected with betaine therapy, suggesting the presence of a reversible methylation defect in this form of chronic renal failure.

liver; nuclear magnetic resonance; uremia; methylation

Figure 3.2: Screenshot of the abstract of an article in HTML.

The full text of biomedical articles, however, is not commonly available in a standardised plain text format. The trend towards open access publishing in biomedicine has resulted in the growing availability of full-text journal publications in PDF and HTML formats. Extracting the text from PDF documents is technically possible, but since PDF is primarily intended as a displaying format, this extraction often results in (small) errors. The HTML format allows more straightforward extraction of the text. The markup can be used to locate and handle certain interesting information. However, the HTML markup used by various publishers is not consistent, requiring considerable effort to take full advantage of this markup. Small inconsistencies in encoding can interfere with a consistent plain text distillation process. The Beta-character can, for example, be represented in HTML as an image, or encoded using different Unicode values (both the official β-character and the similar looking German Eszett 'ß' are used). The hyphen ('-') is encoded using even more values.

Figure 3.3(a) shows the result of converting the first lines of the article in Figure 3.2. Author names and affiliations have been removed, as well as the small navigation table. Note that the superscript ^{1}H-NMR has been converted to '1H-NMR'.

3.1.2 Tokenization

After decoding, a *tokenization* or *lexical analysis* process is required to convert the stream of characters into a stream of "words" or "tokens". Tokenization strongly influences the index vocabulary, since the index terms are primarily based on these tokens.

Some of the properties of biomedical text which require consideration during the design

```
Dietary betaine modifies hepatic metabolism but not renal injury in rat polycystic kidney disease.
We undertook a morphometric and proton nuclear magnetic resonance (1H-NMR) study to test the
hypothesis that 1% dietary betaine supplementation would ameliorate renal disease in the
heterozygous Han:SPRD-cy rat, a model of polycystic kidney disease (PKD) and progressive chronic
renal failure.  After 8 wk of pair feeding, betaine had no effect on renal cystic change, renal
interstitial fibrosis, serum creatinine, serum cholesterol, or serum triglycerides.  1H-NMR
spectroscopy of renal tissue revealed no change in renal osmolytes, including betaine, or renal
content of other organic anions in response to diet.
```

(a) After decoding.

```
dietary betaine modifies hepatic metabolism but not renal injury in rat polycystic kidney disease
we undertook a morphometric and proton nuclear magnetic resonance 1hnmr study to test the
hypothesis that 1 dietary betaine supplementation would ameliorate renal disease in the
heterozygous hansprdcy rat a model of polycystic kidney disease pkd and progressive chronic renal
failure after 8 wk of pair feeding betaine had no effect on renal cystic change renal interstitial
fibrosis serum creatinine serum cholesterol or serum triglycerides 1hnmr spectroscopy of renal
tissue revealed no change in renal osmolytes including betaine or renal content of other organic
anions in response to diet
```

(b) After tokenization.

```
dietari betain modifi hepat metabol renal injuri rat polycyst kidnei diseas undertook morphometr
proton nuclear magnet reson 1hnmr studi test hypothesi 1 dietari betain supplement amelior renal
diseas heterozyg hansprdci rat model polycyst kidnei diseas pkd progress chronic renal failur 8 wk
pair feed betain effect renal cystic chang renal interstiti fibrosi serum creatinin serum
cholesterol serum triglycerid 1hnmr spectroscopi renal tissu reveal chang renal osmolyt includ
betain renal content organ anion respons diet
```

(c) After stop-word removal and stemming.

```
 8 renal        4 betain   3 diseas        3 serum         2 1hnmr          2 chang
 2 dietari      2 kidnei   2 polycyst      2 rat           1 1              1 8
 1 amelior      1 anion    1 cholesterol   1 chronic       1 content        1 creatinin
 1 cystic       1 diet     1 effect        1 failur        1 feed           1 fibrosi
 1 hansprdci    1 hepat    1 heterozyg     1 hypothesi     1 includ         1 injuri
 1 interstiti   1 magnet   1 metabol       1 model         1 modifi         1 morphometr
 1 nuclear      1 organ    1 osmolyt       1 pair          1 pkd            1 progress
 1 proton       1 reson    1 respons       1 reveal        1 spectroscopi   1 studi
 1 supplement   1 test     1 tissu         1 triglycerid   1 undertook      1 wk
```

(d) Frequencies and index terms for the example article.

Figure 3.3: The intermediate output of typical document preprocessing steps.

of the tokenization process will be discussed next.

Case In general English texts, start of sentences, most proper names, and abbreviations make use of uppercase letters. In gene and protein symbol names, upper and lowercase letters are often mixed as in 'PrPSc proteins' which refers to isoforms of prion proteins ('PrP proteins'). Using case sensitive tokenization can be beneficial for searching these named entities, but since capitalisation is not consistent, it might do more harm than good (Jiang and Zhai, 2007).

Multi-word terms Biomedical terms often consist of multiple words and are frequently complex noun phrases which combine multiple terms. Nenadic et al. (2005) noted that in a collection of MEDLINE citations, 85% of the terms consisted of more than one word. Splitting terms such as 'Tumor Necrosis Factor-alpha' into multiple independent tokens may result in nondescript index terms. This can be partially solved by using a positional index and allowing phrase or proximity queries (Carpenter, 2004; Manning et al., 2008), but this does present the new challenge of weighing the phrases in the retrieval model. In fact, during the TREC Genomics 2004 benchmarking, Carpenter (2004) noticed that phrase-based searching performed worse than word-based searching. That performance might be explained by a poor phrase recognition system (with a precision and recall between 60% and 80%), using phrases that were too long and specific (which are very infrequently used in the collection) or incorrect inclusion of the phrases in the retrieval model. Alternatively, a multi-word term may be identified and indexed as a single token, but this might be too stringent since it will no longer match separate words.

Complex multi-word terms are often abbreviated using improvised abbreviations, which mix uppercase letters, lowercase letters, and digits to reflect the originating terms. The 'solute carrier family 40 (iron-regulated transporter), member 1' gene, responsible for encoding the ferroportin1 protein, is abbreviated as 'SLC40A1'. Tokenising 'SLC40A1' as 'slc40a1' is very precise. However, separating the token 'slc' also allows matching to the abbreviation 'SLC', used for the (related) gene family of solute carriers. How these "compound" abbreviations are handled is expected to influence the precision and recall of searches using these terms.

Numbers Numbers are typically not valuable in an index without their surrounding context. They are used to indicate quantities, dates, and database identifiers. Depending on the context, it may be desirable to store numbers in the index as well.

Different number systems are used (interchangeably) in gene and protein names to indicate sub-families, members and other variants. Arabic numerals are used ('p42'); the Greek alphabet is used ('ABCB5beta', but also 'ABCB5b' and 'ABCB 5 β') and Roman numerals are used ('ApoL-III'). During lexical analysis, these numbers can be normalised and stored with their surrounding context (Büttcher et al., 2004; Huang et al., 2006; Jiang and Zhai, 2007).

Again, how combinations of letters and digits in a single word are to be treated depends on the underlying retrieval system. Pirkola and Leppänen (2003) split these words into separate tokens, but used a proximity operator to make sure the tokens appeared close together in matching documents. Tomlinson (2003) reported on

better performance by keeping sequences of letters and digits as single tokens using a commercial TF.IDF-based system.

Hyphens and other special characters Hyphens are used to attach prefixes ('anti-depression'), suffixes ('amyloid-like'), as a breakpoint between syllables in printed text (for example if the text has been decoded from PDF), for compounds ('Creutzfeldt-Jakob'), and as hanging hyphens ('alpha- and beta-isomorphs'). In the case of in-term hyphens, lexical analysis can remove the hyphen and treat the word combination as a single token. Such an approach is adequate when the in-term hyphens are consistently used across documents and queries. In many cases treating a word with multiple hyphens as a single token is undesirable. For example, tokenising 'N-Acetyl-Muramyl-L-Alanyl-D-Glutamic-alpha-Amide' as 'nacetylmuramyllalanyldglutamicalphaamide' is not useful. A tokenization heuristic can be used to decide how to handle hyphens.

Parentheses (or round brackets) are commonly used to provide supplementary information, for example to introduce an abbreviation ('Transmissible spongiform encephalopathies (TSEs)'), or as indicator of an optional plural noun ('enzyme(s)'). In biomedical text parentheses are also frequently a part of gene symbols such as 'PrP(C)' and 'GST-Ub(K63A)'.

Huang et al. (2006) used the term *breakpoints* to indicate positions at which biomedical terms can be split into parts. These breakpoints can be indicated by explicit characters in words such as hyphens, slashes, and parentheses, but also implicitly by change in case and the alteration between letters and digits. These parts are subsequently replaced by variants and combinations are used as tokens. Huang et al. (2006) used this method to expand queries. However, the approach was not compared to a baseline without expansion. Büttcher et al. (2004) used a similar approach for query expansion, resulting in a large number of query variants. 'Lsp1alpha' (Larval serum protein 1 alpha) for example, was expanded to: 'lsp-1-alpha', 'lsp-1-a', 'lsp-1alpha', 'lsp-1a', 'lsp1- alpha', 'lsp1-a', 'lsp1alpha', 'lsp1a' (Büttcher et al., 2004). This approach of finding breakpoints and normalising them was further investigated by Jiang and Zhai (2007) for tokenising both queries and documents. This will be discussed later.

Figure 3.3(b) shows the result of a basic lexical analysis: non-alphanumeric are removed and sequences of alphanumeric characters are treated as tokens. Notice that 'Han:SPRD-cy' is tokenised as 'hansprdcy'.

Character n-gramming

As an alternative to word-based tokenization, the text can be transformed into *character n-grams*. Character n-grams are fixed length sequences of characters found in the text. For example, the phrase 'cell division' can be tokenised as the word overlapping 5-grams 'cell ', 'ell d', 'll di', 'l div', ' divi', etcetera. Character n-gramming has been successfully used for languages without explicit word separators such as Chinese, Japanese, and Korean, but also for cross-lingual IR of European languages (McNamee, 2008). The many breakpoints and multiword terms in biomedical IR can be considered as a similar word separation problem. Therefore, n-grams might be a valuable representation for biomedical IR as well.

3.1.3 Stop-word removal

During stop-word removal, frequently used, uninformative words are filtered from the tokens. Words such as 'a', 'the' and 'are' occur in almost every document: storing these words as index terms would make up for a large part of the index size (Manning and Schütze, 1999). Moreover, these uninformative words are only rarely used for searching. Often, a fixed list of stop-words is used. This is sometimes called a stoplist or negative dictionary. Frequently occurring domain-specific stop-words, such as such as 'disease', 'biology' in the case of TREC Genomics, can be added to such a list (Urbain et al., 2006). Alternatively, terms encountered frequently or infrequently can be pruned from the index.

Query-specific stop-word removal can also be employed to remove non-informative terms from the queries. By removing words such as 'find', 'related' and 'documents' from the query, retrieval performance is likely to be improved.

3.1.4 Stemming and lemmatisation

Stemming and lemmatisation are forms of conflation: they remove word endings to obtain the same root form. 'diseases' and 'disease' can for example be stemmed to 'diseas'; searching for 'diseases' will subsequently also match documents which only contain 'disease'. Conflation is a recall-enhancing operation: the same query term returns more documents containing the actual search word or a similar word. There is the risk, however, to conflating words with an unrelated meaning to the same stem. For example, when 'universe' and 'university' are both stemmed to 'univers' using a Porter stemmer.

The difference between stemming and lemmatisation is that the first does not use any contextual information and stems each word on its own. In contrast, lemmatisation tries to determine the lexeme of a word. This requires information about the part of speech and grammar of a language. Different rule-based stemming methods can be used such as a Porter (1980) and Lovins (1968) stemmer.

To overcome some of the limitations, additional rules and heuristics can be used to prevent incorrect stemming. Zhou and Yu (2006) for example, did not apply stemming in cases where the word looked like a gene name, preventing gene names such as 'IDEE' from being stemmed to a different gene 'IDE'. Similarly, Urbain et al. (2006) only applied stemming when the word was not an acronym.

Stemming can also be carried out online, by storing the full word forms in the index and expanding query words with word forms with the same stem. This also allows weighted stemming, that is, assigning less importance to expansion terms, but this has shown not to be as effective as simply not weighing them (Kraaij, 2004).

Figure 3.3(c) shows the result of stop-word removal and applying a Porter stemmer. Note that words such as 'in', 'but' and 'not' have been removed and word endings have been changed, such as 'kidney' to 'kidnei'. Finally, a selection of the tokens can be used actually as index terms. Figure 3.3(d) shows the index terms sorted by descending frequency in the document.

3.2 Research questions

In the previous section, the automatic preprocessing of text to obtain index terms was explained. The various peculiarities of biomedical terminology that need to be taken into consideration when choosing a particular preprocessing approach were described.

We expect that small changes in preprocessing heuristics have a strong impact on retrieval performance. Moreover, we expect that preprocessing heuristics can strongly improve a baseline tokenization strategy.

We will investigate a number of document preprocessing heuristics in the context of language model IR (introduced in subsection 2.1.3). We limit this investigation to single word-based retrieval. No phrase or proximity operators are used, since inclusion of these operators, even manually, is far from trivial (Carpenter, 2004; Pirkola, 2005).

The following four research questions are posed.

RQ1.1: *What is the impact of stop-word removal?*

Ideally, stop-word removal should not influence retrieval at all: the retrieval model should take into account that these words are uninformative and should be ignored when matching a query to a document. However, the inclusion of a stop-word in a query might also signal that the term is more important for this information need.

RQ1.2: *What is the impact of stemming?*

In most cases, stemming is expected to improve retrieval effectiveness. It is expected that stemming may be especially beneficial in the biomedical domain: many domain-specific terms are used both as nouns and as verbs; on the one hand, stemming is expected to enhance the recall of a single stem, at only a small cost to precision. On the other hand, gene symbol names might suffer from stemming, as removing the last characters can easily conflate different genes to a single index term.

RQ1.3: *What is the impact of using different* breakpoints *to find word parts and how should these word parts be normalised?*

Since biomedical terms often consist of multiple words, or abbreviations consisting of multiple parts, using breakpoints and normalising the word parts determined with these breakpoints is expected to have a large impact on retrieval effectiveness.

RQ1.4: *How does word-based tokenization compare to character n-gramming?*

As an alternative to finding breakpoints and normalising the found word parts, character n-gramming can be used to determine index terms. This has been shown to be an effective tokenization approach for languages without explicit word boundaries and large morphological variation (McNamee, 2008). Similarly, character n-gramming might also be useful for biomedical IR, where boundaries between terms are equally unclear.

3.3 Experimental setup

In this section the experimental setup for comparing different lexical analysis methods will be described. In subsection 3.3.1 the test collections will be introduced. In subsections 3.3.2 and 3.3.3 the retrieval model and evaluation measures will be described, respectively. Finally, in subsection 3.3.4 the tested heuristics will be outlined.

3.3.1 Test collection

The TREC Genomics document collections, topics and relevance judgements, used for the TREC Genomics benchmarks between 2004 and 2007, were used for the evaluation (Hersh et al., 2004, 2005, 2006, 2007). The 2004 and 2005 topic sets were used to search a collection of 4,591,008 MEDLINE citations, referred to as the "2004 document collection". The 2006 and 2007 topics sets were used to query a collection of 162,259 full-text journal articles, referred to as the "2006 document collection". See section 2.2.5 for a detailed description of these collections and topics.

A purely ad hoc document retrieval task was investigated: based on a short textual description of an information need, the most relevant documents had to be retrieved. It should be noted that originally the TREC 2006 and 2007 tasks were passage retrieval tasks: in these cases, the documents in which relevant passages have been located are assumed to be relevant. Document retrieval can be considered as a first step to passage retrieval in these cases.

The four (2004 to 2007) topic sets were used to obtain the following two sets of queries.

original The complete and unaltered original topic description was used as a query to the retrieval system.

manual Queries were obtained by taking the topic description and by manually removing query specific words and phrases, such as "Find articles about".

For the 2004 topic descriptions, which consists of a Title, Need and Context section, both the Title and Need were used for obtaining queries (see appendix A for a list of topics).

The 2006 document collection consisting of HTML documents was split into sections using several different manually created templates to support the differences in document formatting. The text in these sections was converted into plain text; tables, figures, headers, and footers of the webpages were discarded.

3.3.2 Retrieval model

The Kullback-Leibler divergence retrieval model as implemented in the Lemur toolkit was used for the experiments[1]. The model ranks documents by descending KL-divergence between query and document language models. The parameters of the document language models were based on a maximum likelihood estimate linearly smoothed (Jelinek-Mercer smoothing) with the collection language model (Jelinek and Mercer, 1980).

For the experiments with pseudo-relevance feedback, "relevance models" proposed by Lavrenko and Croft (2001) were used (described in their paper as method 2). These

[1] `http://www.lemurproject.org`

relevance models are updated query language models, based on an interpolation between the original query language model and a language model sampled from pseudo-relevant documents. The parameters of the latter language model are estimated as a joint probability of observing a word with the terms in the query. The feedback mechanism requires setting of three parameters: 1) the number of feedback documents, 2) the number of feedback terms, and 3) a weight controlling the interpolation between original and update query model.

In our experiments, the optimal settings of the smoothing parameter and the parameters for pseudo-relevance feedback were based on a sweep over a range of values. In particular, the amount of background smoothing, was varied between 0.05 and 0.95 with a step size of 0.05; the number of feedback terms and documents was varied between 5 and 25 with a step size of 5; the weight of the original query in the feedback query were varied between 0.1 and 0.9 with a step size of 0.1. The results based on parameters yielding the highest Mean Average Precision will be reported, similar to Lafferty and Zhai (2001); Zhai and Lafferty (2004); Metzler and Croft (2005). This parameter sweep is computationally expensive, but provides an upper bound on the retrieval performance using the described heuristics.

3.3.3 Evaluation measures

As indicators of retrieval effectiveness, mean average precision (MAP) and rank precision (P@10) were used. MAP is a good summary measure which emphasises early precision, but also strongly takes into account recall. P@10 gives an indication of how the system performs for end-users who are primarily interested in the first set of retrieved documents. Additionally, the impact of tokenization on index size was analysed.

3.3.4 Evaluated tokenization heuristics

The following 16 tokenization heuristics were tested.

base This heuristic lowercases the input text and keeps uninterrupted sequences of either characters [a-z] or digits [0-9] as tokens.

basestop This heuristic extends *base* with stop-word removal. The PubMed stop-word list[2] was used.

basestem This heuristic extends *base* with Porter stemming.

breakpoint Ten combinations of breakpoint sets and breakpoint normalisation heuristics were tested[3]. The tokenization was carried out as follows.

1. Six preprocessing steps suggested by Jiang and Zhai (2007) were applied. These heuristics replaced a number of special characters by spaces[4] and removed brackets and punctuation in the text followed or preceded by whitespace.

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

[3] Three breakpoint sets were combined with four normalisation heuristics; two combinations of breakpoint set and normalisation led to the same tokenization, resulting in $3 \times 4 - 2$ combinations.

[4] these characters are obviously not in the breakpoint set

Table 3.1: Special characters used in breakpoints, as defined by Jiang and Zhai (2007).

	characters
character set 1	( ) [ ] - _
character set 2	. : ; , ’ +

2. Non-alphanumeric characters not in the breakpoint set were replaced by spaces.
3. The text was split into (character) strings based on whitespace.
4. In each string breakpoints were identified using the breakpoint set.
5. The parts in the breakpointed string were lowercased.
6. The parts in the breakpointed string were turned into tokens using the normalisation heuristic.

The following breakpoint sets defined by Jiang and Zhai (2007) were used for the identification of *breakpoints*. Table 3.1 lists two character sets which were used in the breakpoint sets.

breakpoint set 1 Characters in character set 1 were treated as breakpoints.

breakpoint set 2 Characters in character set 1+2 were treated as breakpoints.

breakpoint set 3 Characters in character set 1+2 were treated as breakpoints. Additionally, hidden breakpoints (transitions between sequences of letters immediately followed by digits or vice versa) were identified.

The following four breakpoint normalisation heuristics were investigated.

split A string divided by breakpoints was split and the parts were used as tokens.

join The parts were joined into a single token.

join + split (js) The parts were tokenised to both a single joint token and its separate tokens.

join + split + extend (jse) additional tokens were generated by moving a sliding window over the separated parts and generating the join of two sequential parts as additional tokens as well.

Table 3.2 illustrates the output of the different normalisation heuristics.

ngram Three heuristics were tested which use word spanning character n-grams, with n set to 4, 5 and 6. First a stream of tokens separated by spaces was obtained by first applying *join* breakpoint normalisation using breakpoint set 2. N-gram tokens were obtained by sliding a window of n characters over the stream of tokens, one character at a time. For instance, the stream ‘mad cow’ would be tokenised as 4-grams ‘mad ’, ‘ad c’, ‘d co’ and ‘ cow’).

The same tokenization heuristic was used for tokenising the documents and the queries.

Table 3.2: Normalising the breakpoints of the artificial word `a*b*c*d` (`*` indicates breakpoints); individual tokens are separated by spaces.

Normalisation	Tokens
join	`abcd`
split	`a b c d`
js	`abcd a b c d`
jse	`abcd a b c d ab bc cd`

3.4 Results

In the following two sections the impact of the evaluated tokenization heuristics on index size and retrieval effectiveness will be discussed.

3.4.1 Index size

Table 3.3 lists the impact of different tokenization strategies on index size. *Base*, *basestop*, *basestem* refer to the baseline, tokenization with stopword removal, and tokenization with stemming, respectively. *Split*, *join*, *js*, and *jse* refer to the for investigated breakpoint normalisation strategies. The number suffix indicates the breakpoint set used. In case the combination of normalisation method and breakpoint set resulted in the same tokenization both numbers are mentioned. *split 1/2* refers, for instance, to the tokenization using *split* breakpoint normalisation using either breakpoint set 1 or 2. *ngram4* to *ngram6* refer to the character ngramming tokenization with ngrams of size 4 to size 6, respectively.

The *vocabulary size* is the number of unique index terms used to index all the documents in the collection. The 2004 baseline index (*base* in Table 3.3) has over 1 million unique terms in its vocabulary and each document contains 164 tokens on average. The 2006 vocabulary is twice as large and its documents are on average 34 times larger. The increased vocabulary size, visualised in Figure 3.4 is caused by numbers, author names, and noisy terms caused by errors in the HTML to plain text decoding process.

As expected, removing stop-words reduced the size of the index vocabulary only marginally. The average document length was reduced strongly however (34% and 31% for 2004 and 2006, respectively). Stemming reduced the vocabulary size with 19% and 8%, respectively.

We expected that splitting words on breakpoints would decrease the vocabulary size and increase average document length. This was only the case when all types of breakpoints were considered. The fact that splitting with breakpoint set 1 (*split1/2*) has an even larger vocabulary size is caused by the handling of sequences of characters and digits. *Base* separated 'p' and '53' in 'p53', whereas *split* only did this when hidden breakpoints (breakpoint set 3) are considered. Sequences of both letters and digits were therefore added to the vocabulary. As a side effect, the document length actually slightly decreased compared to the baseline. Splitting with breakpoint set 1 (*split3*) showed a small decrease in vocabulary size and increase in document length.

Joining word parts at breakpoints as new index terms (*join*, *js* and *jse*) lead to large increases in vocabulary size (up to 312% and 313% for *jse3*), and increases in document

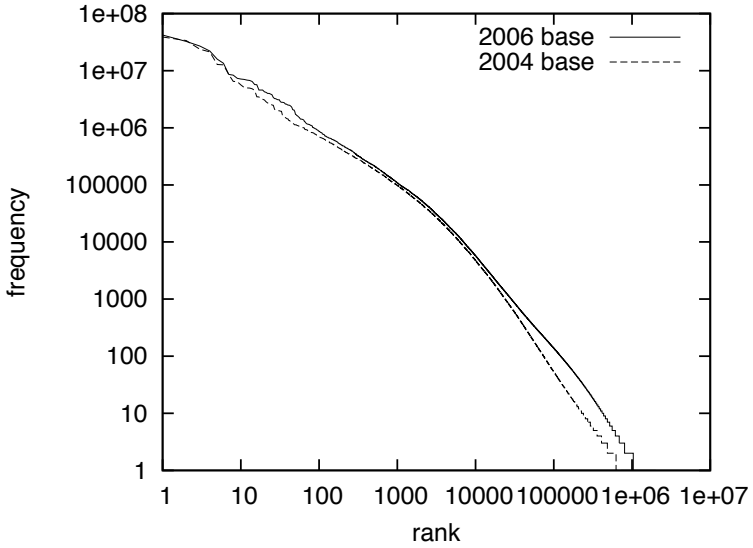


Figure 3.4: Zipf' curve of 2004 and 2006 term frequencies (using the baseline tokenization heuristic): collection frequencies are plotted against descending frequency rank.

size (up to 13% and 22% for *jse3*).

The n-gram tokenization approach produces long documents, with a strongly growing vocabulary with the size of n.

3.4.2 Retrieval effectiveness

Table 3.4 to Table 3.8 summarise the retrieval effectiveness of the tested tokenization heuristics in terms of mean average precision (MAP) and precision at rank 10 (P@10). The left part of the table lists the results obtained by using the original queries, the right part shows queries from which query specific stop-words had been manually removed. As one would expect, all scores of the manually crafted queries are lower than their original counterpart. We will discuss the results as answers to the research questions posed in section 3.2.

RQ1.1: *What is the impact of stop-word removal?*

Table 3.4 lists the impact of stop-word removal and stemming compared to the baseline tokenization (*base*). Removing stop-words (*basestop*) showed to be effective on all collections, especially when the original queries were used (up to 12.6% increase in MAP). The manual queries showed only small changes. The experiments on the full-text collection (2006 and 2007) benefitted more from stop-word removal than the experiments using the citation collection (2004 and 2005).

RQ1.2: *What is the impact of stemming?*

For most query sets, Porter stemming (*basestem*) also showed positive improvements in retrieval effectiveness. Interestingly, the manual queries benefitted more from stemming than the original queries. The improvements can be attributed to word types which appear in a similar meaning as verb and noun. Example words found in the query were 'activation'

Table 3.3: Index statistics for different tokenizations.

	2004		2006	
tokenization	Token types	Avg. length	Token types	Avg. length
base	1,039,502	163.7	2,008,168	5,641.8
basestop	1,039,369	107.7	2,008,035	3,902.9
basestem	845,445	163.7	1,838,611	5,641.8
split 1/2	1,320,097	162.2	2,928,946	5,482.7
split3	998,976	165.7	1,892,240	5,785.6
join1	3,230,655	156.3	4,705,310	5,255.5
join 2/3	3,550,579	154.5	6,366,735	5,157.9
js1	3,358,405	167.4	5,091,554	5,688.3
js2	3,688,914	168.8	6,879,627	5,758.3
js3	3,701,130	174.3	6,858,425	6,236
jse1	3,729,135	173.3	5,537,655	5,915.5
jse2	4,169,376	176.5	8,110,532	6,083.2
jse3	4,283,066	185.6	8,302,990	6,863.8
ngram4	564,347	1,029.2	806,270	32,311.2
ngram5	3,329,192	1,027.5	5,251,299	32,018.4
ngram6	13,649,057	1,025.7	21,751,322	31,727.4

(stemmed to 'activ'), 'synthesize' ('synthes'), 'mutations' ('mutat'), and 'toxicities' ('toxic'). In few cases stemming hurt retrieval, for example when the stemmed word became too general ('infection' as part of 'HIV type 1 infection' was stemmed to 'infect').

RQ1.3: *What is the impact of using different* breakpoints *to find word parts and how should these word parts be normalised?*

The breakpoint normalisation schemes affected retrieval performance rather differently (see Table 3.5 and Table 3.6). However, most of the differences to the baseline were not significant. The largest and most significant changes were found on the 2004 document collection. On the 2006 collection, the impact of the breakpoint normalisation was small and the changes were mostly insignificant. For the 2007 queries, this can be simply explained by the fact that the queries did not contain that many breakpoints (28 out of 36 queries did not contain any breakpoints). For the set of 2006 queries, breakpoint normalisation decreased retrieval effectiveness. This is explained by the fact that the 2006 topics asked about the relationship between two aspects which should both be present in the retrieved documents. The optimal smoothing values for the manual queries (see appendix B.1) confirmed this. The low smoothing values indicate that almost perfect coordinate level matching performs best: all query terms should be present in the document to achieve best performance. The breakpoint normalisation could have overemphasised one aspect, therefore deteriorating the results.

The *join* method performed worse than the baseline on all collections and queries on both MAP and P@10. The tokenization of particular query terms often caused the deterioration, such as 'presenilin-1' (tokenised as the 'presenilin1') and '4-GABAA' (tokenised as '4gabaa'). In only a few cases, the join-approach improved the baseline, for example when gene symbols were kept as a single token (genes 'L1' and 'L2', and virus 'HPV11').

The conservative approach of splitting words at breakpoints (*split*) showed small improvements for the citation collection when a conservative set of breakpoints was used (*split1*). On the full-text collection splitting on breakpoints often deteriorated the performance. Most of the differences were not significant however.

Joining and splitting (*js*) words especially improved performance on the 2004 collection. On the 2006 collection, the improvements with this approach were smaller and more unpredictable.

The most beneficial heuristic for the 2004 collection combined joining, splitting and expanding (*jse*): Large increases (between 7 and 17% MAP) in retrieval effectiveness were observed. For the 2006 and 2007 query sets however, performance again showed only small positive and negative changes.

RQ1.4: *How does word-based tokenization compare to character n-gramming?*

Table 3.7 lists the results of using character n-grams as an index and search representation. Irrespective of the size of the window, the approach showed large (up to 56% MAP) and significant losses in retrieval effectiveness on all collections and query sets and is clearly not a good option for biomedical IR based on language models. Only a few topics showed improvement when using ngrams. In these cases, the improvement was caused by the phrase-search effect of using word spanning ngrams (such as ngrams found in the phrase 'time course' or 'glyphosate tolerance').

Table 3.4: Retrieval effectiveness of tokenization with stemming and stop-word removal, in comparison to the baseline tokenization. Percentages indicate differences to the baseline. [1], [2] and [3] indicate significant differences to the baseline at confidence levels 0.05, 0.01 and 0.001 respectively, determined with a paired sign test. The highest value of each column is printed in boldface.

(a) 2004 queries

	2004 original		2004 manual	
	MAP	P@10	MAP	P@10
base	0.2950	0.5540	0.3032	0.5660
basestop	0.2967 +0.6%	0.5680 +2.5%	0.3008 -0.8%	0.5700 +0.7%
basestem	**0.3139** +6.4%	**0.5740** +3.6%	**0.3232** +6.6%	**0.5720** +1.1%

(b) 2005 queries

	2005 original		2005 manual	
	MAP	P@10	MAP	P@10
base	0.1819	0.3041	0.2196	**0.3735**
basestop	**0.1942** [3] +6.8%	**0.3327** +9.4%	0.2213 +0.7%	0.3653 -2.2%
basestem	0.1866 +2.6%	0.3184 +4.7%	**0.2255** [1] +2.7%	0.3612 -3.3%

(c) 2006 queries

	2006 original		2006 manual	
	MAP	P@10	MAP	P@10
base	0.3565	0.4500	0.4245	0.4769
basestop	**0.3920** [3] +9.9%	**0.4615** +2.6%	0.4270 +0.6%	0.4769
basestem	0.3463 -2.9%	0.4462 -0.9%	**0.4455** +4.9%	**0.4846** +1.6%

(d) 2007 queries

	2007 original		2007 manual	
	MAP	P@10	MAP	P@10
base	0.2309	0.4194	0.2745	0.4750
basestop	**0.2599** [3] +12.6%	**0.4583** +9.3%	0.2672 [1] -2.7%	0.4639 -2.3%
basestem	0.2363 [1] +2.3%	0.4139 -1.3%	**0.2933** [2] +6.8%	**0.4778** +0.6%

Table 3.5: Retrieval effectiveness of tokenization using breakpoint normalisation (2004 collection). See Table 3.4 for legend.

(a) 2004 queries

	2004 original				2004 manual			
	MAP		P@10		MAP		P@10	
base	0.2950		0.5540		0.3032		**0.5660**	
join1	0.2950		0.5300	-4.3%	0.2919	-3.7%	0.5560	-1.8%
join2	0.2945	-0.2%	0.5260	-5.1%	0.3012 [1]	-0.6%	0.5520	-2.5%
split1	0.3274	+11.0%	0.5600	+1.1%	0.3274	+8.0%	0.5600	-1.1%
split3	0.2799	-5.1%	0.5320	-4.0%	0.2941	-3.0%	0.5600	-1.1%
js1	0.3119	+5.7%	0.5400	-2.5%	0.3174	+4.7%	0.5560	-1.8%
js2	0.3118	+5.7%	0.5400	-2.5%	0.3171	+4.6%	0.5560	-1.8%
js3	0.3309	+12.2%	0.5620	+1.4%	0.3378	+11.4%	0.5420	-4.2%
jse1	0.3026 [2]	+2.6%	0.5440	-1.8%	0.3096	+2.1%	0.5540	-2.1%
jse2	0.3017 [1]	+2.3%	0.5320	-4.0%	0.3093	+2.0%	0.5540	-2.1%
jse3	**0.3454** [2]	+17.1%	**0.5640**	+1.8%	**0.3524**	+16.2%	0.5580	-1.4%

(b) 2005 queries

	2005 original				2005 manual			
	MAP		P@10		MAP		P@10	
base	0.1819		0.3041		0.2196		0.3735	
join1	0.1776 [1]	-2.4%	0.2939	-3.4%	0.2079	-5.3%	0.3469	-7.1%
join2	0.1737 [1]	-4.5%	0.2980	-2.0%	0.2027	-7.7%	0.3429	-8.2%
split1	0.1938	+6.5%	0.3000	-1.3%	0.2291	+4.3%	0.3653	-2.2%
split3	0.1845	+1.4%	0.3204	+5.4%	0.2237	+1.9%	0.3653	-2.2%
js1	0.1943	+6.8%	0.2939	-3.4%	0.2308	+5.1%	0.3714	-0.5%
js2	0.1926	+5.9%	0.2959	-2.7%	0.2291	+4.3%	0.3735	-0.0%
js3	**0.2022**	+11.1%	**0.3571**	+17.4%	0.2320	+5.6%	0.3898	+4.4%
jse1	0.1840	+1.1%	0.2837	-6.7%	0.2275	+3.6%	0.3571	-4.4%
jse2	0.1810	-0.5%	0.2837	-6.7%	0.2250	+2.4%	0.3735	
jse3	0.1954	+7.4%	0.3388	+11.4%	**0.2362**	+7.5%	**0.3980**	+6.6%

Table 3.6: Retrieval effectiveness of tokenization using breakpoint normalisation (2006 collection). See Table 3.4 for legend.

(a) 2006 queries

	2006 original				2006 manual			
	MAP		P@10		MAP		P@10	
base	0.3565		0.4500		0.4245		0.4769	
join1	0.3508	-1.6%	0.4308	-4.3%	0.3787	-10.8%	0.4615	-3.2%
join2	0.3490	-2.1%	0.4462	-0.9%	0.3793	-10.6%	0.4654	-2.4%
split1	0.3511	-1.5%	0.4500	-0.0%	0.3924	-7.6%	0.4769	
split3	0.3575	+0.3%	0.4308	-4.3%	**0.4284**	+0.9%	0.4731	-0.8%
js1	0.3493	-2.0%	0.4462	-0.9%	0.3919	-7.7%	0.4577	-4.0%
js2	0.3462	-2.9%	0.4385	-2.6%	0.3923	-7.6%	0.4692	-1.6%
js3	**0.3878**	+8.8%	**0.4731**	+5.1%	0.4248	+0.1%	**0.5077**	+6.5%
jse1	0.3390	-4.9%	0.4385	-2.6%	0.3912	-7.8%	0.4692	-1.6%
jse2	0.3342	-6.3%	0.4385	-2.6%	0.3853	-9.2%	0.4731	-0.8%
jse3	0.3665	+2.8%	0.4423	-1.7%	0.3979	-6.3%	0.4769	

(b) 2007 queries

	2007 original				2007 manual			
	MAP		P@10		MAP		P@10	
base	0.2309		0.4194		0.2745		0.4750	
join1	0.2281	-1.2%	0.4139	-1.3%	0.2631	-4.2%	0.4611	-2.9%
join2	0.2234	-3.3%	0.4111	-2.0%	0.2552	-7.0%	0.4528	-4.7%
split1	**0.2394**	+3.7%	**0.4306**	+2.6%	**0.2815**	+2.5%	0.4833	+1.8%
split3	0.2204 [3]	-4.6%	0.4056	-3.3%	0.2612	-4.8%	0.4722	-0.6%
js1	0.2392	+3.6%	0.4194		0.2806	+2.2%	0.4861	+2.3%
js2	0.2346	+1.6%	0.4167	-0.7%	0.2773	+1.0%	0.4806	+1.2%
js3	0.2330 [3]	+0.9%	0.4139	-1.3%	0.2762 [1]	+0.6%	**0.4889**	+2.9%
jse1	0.2364	+2.4%	0.4167	-0.7%	0.2777	+1.2%	0.4750	
jse2	0.2261	-2.1%	0.3861	-7.9%	0.2695	-1.8%	0.4556	-4.1%
jse3	0.2193 [1]	-5.0%	0.3778	-9.9%	0.2640 [1]	-3.8%	0.4528	-4.7%

Table 3.7: Retrieval effectiveness of tokenization based on character n-gramming. See Table 3.4 for legend.

(a) 2004 queries

	2004 original		2004 manual	
	MAP	P@10	MAP	P@10
base	**0.2950**	**0.5540**	**0.3032**	**0.5660**
ngram4	0.1768 [3] -40.1%	0.3940 -28.9%	0.1960 [3] -35.3%	0.4240 -25.1%
ngram5	0.1611 [3] -45.4%	0.3500 -36.8%	0.1712 [3] -43.5%	0.3780 -33.2%
ngram6	0.1332 [3] -54.8%	0.2920 [1] -47.3%	0.1388 [3] -54.2%	0.3020 -46.6%

(b) 2005 queries

	2005 original		2005 manual	
	MAP	P@10	MAP	P@10
base	**0.1819**	**0.3041**	**0.2196**	**0.3735**
ngram4	0.0796 [3] -56.3%	0.1980 -34.9%	0.1445 [2] -34.2%	0.2714 -27.3%
ngram5	0.0768 [3] -57.8%	0.1592 -47.7%	0.1151 [3] -47.6%	0.2204 -41.0%
ngram6	0.0746 [3] -59.0%	0.1531 -49.7%	0.0914 [3] -58.4%	0.1633 -56.3%

(c) 2006 queries

	2006 original		2006 manual	
	MAP	P@10	MAP	P@10
base	**0.3565**	**0.4500**	**0.4245**	**0.4769**
ngram4	0.2763 [1] -22.5%	0.3731 -17.1%	0.2838 [3] -33.1%	0.3654 -23.4%
ngram5	0.1801 [2] -49.5%	0.2808 -37.6%	0.2420 [3] -43.0%	0.3192 -33.1%
ngram6	0.1998 [3] -44.0%	0.2731 -39.3%	0.2012 [3] -52.6%	0.2654 -44.4%

(d) 2007 queries

	2007 original		2007 manual	
	MAP	P@10	MAP	P@10
base	**0.2309**	**0.4194**	**0.2745**	**0.4750**
ngram4	0.1588 [1] -31.2%	0.3389 -19.2%	0.1700 [2] -38.1%	0.3694 -22.2%
ngram5	0.1696 -26.6%	0.3389 -19.2%	0.1760 [1] -35.9%	0.3583 -24.6%
ngram6	0.1622 -29.8%	0.3167 -24.5%	0.1671 [1] -39.1%	0.3250 -31.6%

Table 3.8: Retrieval effectiveness based on combining heuristics and adding relevance feedback. See Table 3.4 for legend.

(a) 2004 queries

	2004 original				2004 manual			
	MAP		P@10		MAP		P@10	
base	0.2950		0.5540		0.3032		0.5660	
combined	0.3570	+21.0%	0.5960	+7.6%	0.3596 [1]	+18.6%	0.5960	+5.3%
base+fb	0.3174 [1]	+7.6%	0.5540		0.3281 [2]	+8.2%	0.5920	+4.6%
combined+fb	**0.4060** [3]	+37.6%	**0.6180**	+11.6%	**0.4041** [3]	+33.3%	**0.6000**	+6.0%

(b) 2005 queries

	2005 original				2005 manual			
	MAP		P@10		MAP		P@10	
base	0.1819		0.3041		0.2196		0.3735	
combined	0.2115	+16.2%	0.3469	+14.1%	0.2355	+7.2%	**0.3878**	+3.8%
base+fb	0.1870	+2.8%	0.2959	-2.7%	0.2323	+5.8%	0.3776	+1.1%
combined+fb	**0.2261**	+24.2%	**0.3796**	+24.8%	**0.2407**	+9.6%	0.3755	+0.5%

(c) 2006 queries

	2006 original				2006 manual			
	MAP		P@10		MAP		P@10	
base	0.3565		0.4500		0.4245		0.4769	
combined	**0.4340** [2]	+21.7%	**0.5192**	+15.4%	0.4349	+2.5%	**0.5192**	+8.9%
base+fb	0.3816	+7.0%	0.4500	-0.0%	**0.4467**	+5.2%	0.5077	+6.5%
combined+fb	0.4280	+20.0%	0.4846	+7.7%	0.4265	+0.5%	0.4846	+1.6%

(d) 2007 queries

	2007 original				2007 manual			
	MAP		P@10		MAP		P@10	
base	0.2309		0.4194		0.2745		0.4750	
combined	0.2724 [3]	+17.9%	0.4556	+8.6%	0.2798	+1.9%	0.4583	-3.5%
base+fb	0.2482	+7.5%	0.4306	+2.6%	0.2853	+3.9%	0.4806	+1.2%
combined+fb	**0.2985** [3]	+29.2%	**0.4889**	+16.6%	**0.3017**	+9.9%	**0.5000**	+5.3%

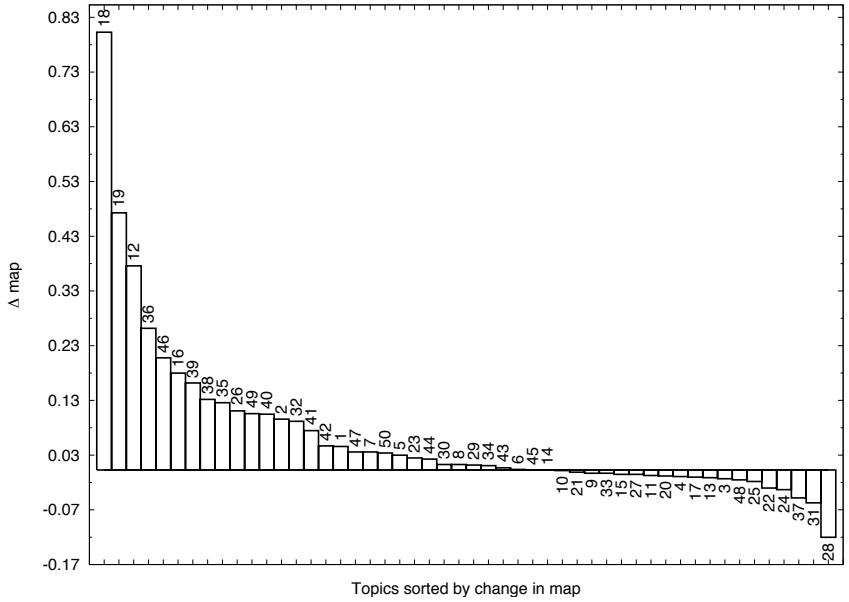


Figure 3.5: Per-topic change in average precision between 2004 baseline and combined (manual queries).

Concluding, the results showed that simple tokenization heuristics can significantly affect retrieval performance. Stemming and removing stop-words generally helped. Using overlapping character n-grams however, was not beneficial. Finding appropriate breakpoints and normalising the word parts separated by these points gave mixed results. A combination of indexing both the joint and separate word parts worked particularly well for abstract collections but slightly hurt performance on a full-text collection.

Combining stemming, stop-word removal and one of the breakpoint normalisation strategies (*js3*) improved the results on all collections (*combined* in Table 3.8). A per-topic analysis explained why, according to a sign test, the results are not significantly different: quite a few topics show a small performance drop. The topics which do improve show a relatively large change. Figure 3.5 illustrates the change in average precision per topic: almost a third of the topics showed a substantial improvement in average precision compared to the baseline. In section 4.3 the result of this combined tokenization approach is further analysed in comparison to a concept-based representation.

Table 3.8 also lists the results of combining the tokenization heuristics with relevance feedback (*base+fb* and *combined+fb*). For the 2004 and 2007 query sets, both the baseline and the combined tokenization heuristics benefitted from relevance feedback. However, the combined heuristic, showed larger relative improvements from applying relevance feedback. One explanation for this relatively large improvement was the fact that the pseudo-relevant documents used for reformulating the query were of higher quality: the initial search results were better, making a reformulated query based on these results better as well.

3.5 Discussion

In this discussion, we will compare our results to related work. After that, the investigated heuristics will be discussed one by one, followed by the limitations of this work.

Related work

The results presented in the previous section show some overlap with but also show some differences from work presented by Jiang and Zhai (2007), who carried out similar experiments on only the TREC Genomics 2004 document collection. Jiang and Zhai also investigated stemming, stop-word removal and breakpoint normalisation, but their conclusions only partially agree with our experimental work. They concluded about stop-word removal that "[it] either does not improve the performance, or only slightly improves the performance"; our results indicated that especially for the original queries, stop-word removal significantly improved retrieval effectiveness and in worst cases only slightly hurt performance. Jiang and Zhai (2007) concluded that breakpoint set 1 is most useful in the context of the normalisation methods they investigated. These normalisation methods included two variations of joining word parts (with or without hyphens) and joining the tokens in a single index term. Our results confirm that for *join* and *split*, breakpoint set 1 is most effective. However, when a normalisation method is used which outputs both word parts and the joined parts as separate tokens (*js* and *jse*), our results indicated that a more extensive breakpoint set (breakpoint set 3) is preferable. The latter approach generates both specific tokens (joined word parts) for precision and shorter, more general tokens (word parts) for recall. Jiang and Zhai argued for using different normalisation strategies for queries containing either verbose or gene symbol terms: this requires multiple indices and choosing the appropriate tokenization method (and corresponding index) at search time. It is unclear, however, how such an approach should cope with queries which contain both verbose and symbol terms. Our results indicate that, simply combining the tokenization heuristics shows to be an effective solution. Finally, our results support Jiang and Zhai (2007)'s conclusions that stemming can be effective for verbose queries. Conditional stemming, as proposed and used by Zhou and Yu (2006) and Urbain et al. (2006), might work even better, but the TREC Genomics topics probably contain too few applicable instances to confirm this.

Stop-word removal

The observed positive impact of stop-word removal might be interacting with the function of smoothing in language model IR. Zhai and Lafferty (2004) indicated that smoothing has a double function: "to make the estimated document language model more accurate and to "explain" the non-informative words in the query". To some extent, this is confirmed by the optimal smoothing values observed in our experiments (see appendix B.1). Verbose queries, such as the original 2004 and 2005 queries showed optimal retrieval effectiveness with a large proportion of collection smoothing ($\lambda > 0.65$). Removing stop-words from these queries resulted in a slight drop of the optimal smoothing values ($\lambda > 0.6$), but smoothing values are still reasonably high. This can be explained by the fact that the queries still contained query specific stop-words such as 'Find related articles'. The optimal smoothing value remained high to compensate for these fairly general words. In the manual queries, both general as well as general query stop-words were removed, resulting in a lower optimal smoothing value ($\lambda = 0.50$ for 2004, $\lambda = 0.05$ for 2005). In these cases the smoothing value compensated for data sparsity in the documents. Especially for the 2004 collection this role of smoothing is important. Documents are short, so more background smoothing is required to compensate for terms which are related to the document but simply do not

appear in it. By explicitly removing stop-words, the amount of smoothing can be focused on making the document language model more accurate.

Stemming

Both stemming and splitting words at breakpoints can be considered variance reduction techniques (Ponte, 2001; Kraaij, 2004). More accurate estimations can be made of stems or word parts, since more data is available, but at the cost of introducing a bias in favour of these terms. We observed stemming to perform well in this domain. We attribute this improvement to the many biomedical terms which occur both as nouns and verbs.

Breakpoint normalisation

The breakpoint normalisation and especially expansion heuristics (*js* and *jse*) appear to be primarily a recall enhancing device. Especially for the TREC 2004 collection, improved performance was observed in mean average precision (which favours high recall). The expansion with additional terms allowed for more flexible matching between the short citations and queries. For the longer documents full text documents in the TREC 2006 collection, such an expansion showed to be unnecessary. To some extent this could be attributed to the 2006 and 2007 topic sets, which did not contain many breakpoints. A second explanation is the fact that the full-text documents in this collection are more verbose and are more likely to contain more spelling variations. Rather than increasing the recall, the expanded queries suffer from query drift by overemphasising breakpointed words. This can, however, also be a reason for improved retrieval: in cases where the query word containing breakpoints describes an important aspect of the query, adding additional terms can cause a desired boosting effect.

Character n-grams

Using overlapping character n-grams performed below expectations. McNamee (2008) observed that on an English newswire document corpus overlapping 4- or 5-grams perform just as well as ordinary words for monolingual search. The difference can be explained by the fact that the Genomics queries are rather long in comparison to the newswire collection used by McNamee. Moreover, informative words in the query tend to be short, such as gene symbol names of 3 or 4 characters. As a result, the n-gramming approach might have overemphasised the phrases found in the original query.

Limitations

We identify three limitations of the current work.

Firstly, in the reported experiments no query operators were used or investigated, such as phrase and synonym operators. Breakpoint normalisation is strongly related to phrase-based search: rather than normalising a compound term to a single token, compound terms can be treated as a phrase. Phrase-based searching has been frequently used for biomedical IR, with varying degrees of success (Carpenter, 2004; Ide et al., 2007; Stokes et al., 2009). Searching with synonym operators may prevent the query drift experienced when using breakpoint normalisation with expansion. To keep the experiments transparent,

unstructured queries were investigated in our experiments. In future work, the relationship between these operators and breakpoint normalisation can be investigated.

Secondly, basic linear smoothing was used in our experiments. In future work more advanced smoothing methods could be taken into account. As the collections contain documents of varying length, a smoothing method which adapts to document length might be more effective in combination with different tokenization heuristics. In our experiments we used a fixed smoothing parameter for all query terms. Alternatively, this smoothing could be varied per query term.

Thirdly, throughout the experiments the same tokenizer was used for both queries and documents. The use of an aggressive tokenization strategy which generates many expansion terms for the documents and a more restrictive tokenizer for the query or vice versa could also be considered in future work.

3.6 Chapter summary

In this chapter, document preprocessing heuristics for word-based biomedical retrieval have been investigated. Preprocessing heuristics have been shown to influence retrieval effectiveness strongly. A baseline for word-based retrieval has been established, which includes expanding documents and queries with terms using breakpoints, stop-word removal, and stemming. In subsequent chapters this form of preprocessing will be used for obtaining word-based representations.

Chapter 4

Concept-based Biomedical IR

"There are very few things that are purely conceptual without any hard content."

Kevin Bacon

Parts of this chapter have been published in Trieschnigg, Schuemie, and Kraaij (2006); Trieschnigg, Kraaij, and de Jong (2007); Trieschnigg, Pezik, Lee, Kraaij, de Jong, and Rebholz-Schuhmann (2009); Trieschnigg, Meij, de Rijke, and Kraaij (2008) and Meij, Trieschnigg, de Rijke, and Kraaij (2010).

In this chapter, a *concept-based* representation of queries and documents will be explored. A *concept* represents information at a higher abstraction level than single words or phrases and should resolve difficulties caused by synonymy and lexical ambiguity. Theoretically, having a concept-based representation of both the queries and the documents should have its advantages over a word-based representation: concepts unambiguously represent information and if both the information need and the document content can be precisely and completely represented in terms of concepts (at the proper level of granularity), text-based retrieval should be outperformed both in terms of precision and recall. In practice, however, every representation, including a concept-based representation, has its limitations: the representation vocabulary might for example not be specific or exhaustive enough to represent all information needs and the document content. As a result, retrieval performance can be actually harmed by such an approach.

In this chapter, a concept-based representation is investigated in practice. Two concept-based representation languages based on a controlled vocabulary and a thesaurus are investigated for biomedical IR in comparison to word-based IR. We will answer RQ2 posed in chapter 1.

RQ2: *What is the added value of a concept-based representation based on terminological resources for biomedical IR?*

The overview of this chapter is as follows. In section 4.1, the two concept representation vocabularies used in this thesis will be described. In section 4.2, seven classification methods will be described for mapping text to a concept-based representation. These classifiers will

be used in subsequent experiments. We will describe a number of out-of-the-box classifiers and we will propose two classification methods based on statistical language models. In sections 4.3 to 4.7 the five research topics described in chapter 1 will be investigated.

RT2a: How documents are represented in a concept-based representation. In section 4.3, we will analyse the two concept-based document representations investigated in this chapter from a statistical perspective.

RT2b: To what extent such a document representation can be obtained automatically. In section 4.4, a selection of classification methods will be evaluated for mapping text-based document representations to concept-based document representations.

RT2c: To what extent a text-based query can be automatically mapped onto a concept-based representation and how this affects retrieval performance. In section 4.5, the classifiers will be used to classify queries and they will be evaluated in their effectiveness for retrieval.

RT2d: To what extent a concept-based representation is effective in representing information needs. In section 4.6, an analysis in an artificial setting will be carried out to determine the added value of the word-based and concept-based query representations.

RT2e: How the relationship between text and concepts can be used to determine the relatedness of concepts. In section 4.7 we will investigate different methods to predict the relatedness of concept-based representations.

The chapter will be summarised in section 4.8. Figure 4.1 shows the approaches which will be investigated in this chapter schematically.

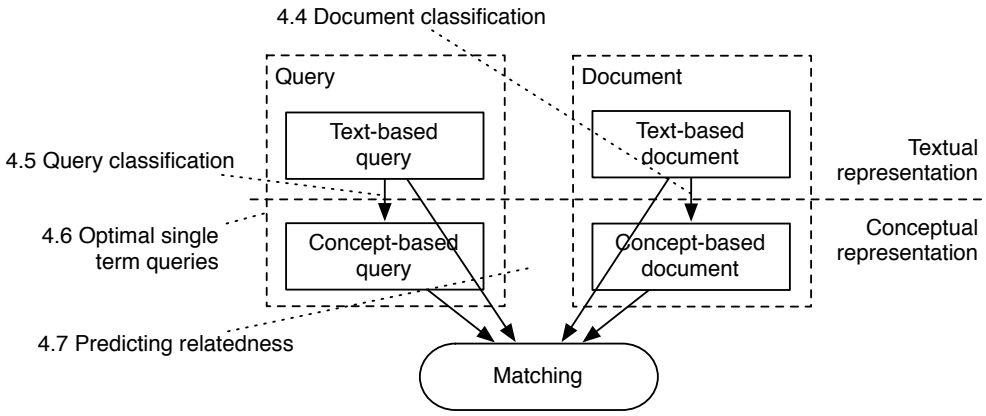


Figure 4.1: Using both a separate text and concept representation for retrieval.

4.1 Two concept languages for biomedical IR

In this thesis, we will investigate the two following concept representation vocabularies in their usefulness for biomedical IR.

MeSH The Medical Subject Headings thesaurus (see subsection 2.2.4), which contains around 24,000 MeSH headings (which we will refer to as concepts) and which is used to manually index MEDLINE citations. The fact that the complete MEDLINE database is manually indexed using this controlled vocabulary, indicates its diversity and broad coverage. Considering its small size, its specificity is limited, however.

UMLS$_{++}$ The Unified Medical Language System (UMLS)[1] extended with several gene and protein dictionaries for four species[2] (Schuemie et al., 2007c), referred to as UMLS$_{++}$ in this chapter. This concept language is a superset of MeSH, and allows for a more fine-grained representation of information. The combined thesaurus consists of 640,016 concepts from 59 vocabularies. In contrast to MeSH, no manually curated UMLS$_{++}$ document representations are available. For UMLS$_{++}$, the concept-based document representation is limited to the output from an automatic mapping process.

An important difference between the two concept vocabularies is that UMLS$_{++}$ does not provide a single consistent view of the world, since it is based on a combination of terminological resources. As a result, particular topics might be more densely covered. MeSH, in contrast, is maintained by a single authority and therefore is less likely to be inconsistent.

Our main claim in this chapter is as follows.

- *A concept-based representation based on a controlled terminological resource can improve the effectiveness of biomedical IR*

We are particularly interested in how this added value is influenced by the choice of representation language, its use and the mapping process to obtain concept-based query and document representations.

In the individual sections, additional research questions will be formulated.

4.2 Automatically mapping text to concepts

In this section, several methods for obtaining a concept-based representation will be introduced. In the next subsection, different types of approaches for mapping text-based to concept-based representations will be discussed. In subsections 4.2.2 to 4.2.8, the systems used for experiments later in this chapter will be discussed in more detail.

[1] UMLS version 2008AB

[2] Entrez-Gene, OMIM, Swiss-Prot (version 103), and Hugo. Where no version is noted, the most recent versions available on 30 March 2009 were used.

4.2.1 Classifying biomedical text

Translating text to concepts can be considered a large multi-class and multi-label text classification problem: one or more labels (concepts) have to be assigned to a piece of text (either a document or a query).

Some classification approaches explicitly link the concepts found to words or phrases in the original text. Such approaches have been referred to as "concept tagging", "concept mapping" and "name identification" (Aronson, 2001; Krauthammer and Nenadic, 2004). Others make an implicit link between concepts and text, by assigning the concepts to the text as a whole. During manual, controlled vocabulary indexing such an implicit link is made: controlled vocabulary index terms are assigned to a document as a whole.

Implicitly or explicitly linked concept classifications have their advantages and disadvantages. On the one hand, an implicit representation allows for abstraction and for including additional concepts, which can increase recall during search. On the other hand, it is more difficult for the user to relate the text to the concept representation, especially when the conceptual representation has been obtained automatically and may contain errors. Explicitly linked concepts are easier to relate by the user and allow for more specific searches. However, they lack the possibility to describe what is written at a more abstract level. A document about information retrieval in the biomedical domain can, for example, be classified with the MeSH term [Computational Biology], without actually explicitly referring it.

In either case, however, a large number of possible classes is available to assign to a piece of text. Most out-of-the-box text classifiers, such as decision trees, rule learning, neural networks, and Support Vector Machines (SVMs) are not directly suitable for a classification task involving thousands of classes, or in this case thousands of concepts. SVMs for example, have shown their superiority to Naive Bayes classifiers on binary classification tasks (Joachims, 1998), but without sophisticated adaptation it is not feasible to build and train a system using SVMs for thousands of concepts. Scalable concept classification is often limited to less sophisticated machine learning methods, such as Naive Bayes and K-Nearest Neighbour classifiers. Related work, especially on MeSH (see Sohn et al. (2008) for more related work), shows a separation between research on sophisticated techniques limited to a subset of the problem and more straightforward techniques which do offer a complete solution.

For example, several researchers have used the OHSUMED collection and investigated the performance of their classifiers on a subset of MeSH descriptors, such as the terms in the Heart Disease branch (Lam and Ho, 1998; Ruiz and Srinivasan, 2002), or by only considering generalised descriptors (Rak et al., 2007). Recently, Sohn et al. (2008) investigated optimal training sets for Naive Bayes' classifiers on a small set of 20 MeSH descriptors. Despite the reported improvements over the K-Nearest Neighbours approach, so far such a classifier has not been proven feasible for all 24,000 MeSH terms.

The focus of this work is on systems which can be used for assigning the full set of concepts found in the concept languages. We distinguish between the following four types of approaches.

Classifiers based on string matching or dictionary lookup simply scan for synonyms of a concept in the text. The synonyms used for scanning are found in the terminological resource. The resource often requires preprocessing and filtering before it can be used

for this kind of string matching. An additional step can be required to disambiguate the detected concepts: based on the surrounding context a choice can be made for the most appropriate concept. MetaMap (subsection 4.2.2), PubMed Automatic Term mapping (subsection 4.2.3) and Peregrine (subsection 4.2.6) strongly rely on this kind of string matching.

Concept-oriented classifiers build and depend on explicit models built for each concept in the thesaurus/controlled vocabulary. They are often inspired by information retrieval techniques and return a ranked list of the most appropriate concepts for the text to classify (Lam et al., 1999; Ruch, 2006). The actual classification, that is the binary assignment of a particular term to a piece of text, is achieved by cutting off the list at a particular rank or score. EAGL (subsection 4.2.4) and the method based on concept language models (subsection 4.2.7) are concept-oriented classifiers.

Nearest-neighbour classifiers or classifiers based on retrieval feedback classify objects based on the known classification of the most similar objects in a training set. Using pseudo-relevance feedback to obtain a concept-based representation can be viewed as a nearest-neighbour classifier: the concept-based representation of a query is based on the classifications of the documents textually most similar to the text-based query. Srinivasan (1996a) is one of the first to use pseudo-relevance feedback to obtain a concept-based representation for a textual query. In subsection 4.2.8, a KNN system based on language models will be described.

Hybrid classifiers combine two or more of the previously mentioned approaches. The Medical Text Indexer, described in section 4.2.5 combines sophisticated dictionary lookup with nearest-neighbour classification to map a text to a concept representation.

In the following subsections, the classification systems used in this chapter will be discussed. In subsections 4.2.2 to 4.2.6, five existing (out-of-the-box) classifiers will be described. Sections 4.2.2 to 4.2.5 will describe systems which use MeSH as a concept language; the Peregrine system described in subsection 4.2.6 uses the UMLS$_{++}$ representation.

In subsections 4.2.7 and 4.2.8, we will propose two approaches to text classification based on ranked statistical language models. The first creates a concept language model based on the documents to which the concept has been assigned. The classification of text is based on a ranking of these concept language models. The second approach is a variant of a K-Nearest Neighbour classifier: the classification is based on a parallel corpus of documents which is available in both a textual and conceptual representation. To perform a classification, these documents are ranked according to the similarity of their textual representation and the text to classify. The actual classification is based on the conceptual representations of these ranked documents.

4.2.2 MetaMap

MetaMap is a program developed by the National Library of Medicine to map text automatically to concepts in the UMLS Metathesaurus (Aronson, 2001). It is a principal component of the Medical Text Indexer described in subsection 4.2.5.

Thesaurus concepts are found in the following five-step process, described in detail in Aronson (2001).

1. **Parsing** The text is parsed (mainly) into noun phrases.

2. **Variant generation** Variants are generated for the noun phrases based on a lexicon containing both general and biomedical words. Variants include synonyms, acronyms and abbreviations, completed with inflected forms. For example, variants of the word 'ocular' include 'eye' (a synonym) and 'eyes' (the plural inflection of the synonym).

3. **Candidate retrieval** Entries which have at least one word in common with one of the generated variants are retrieved from the thesaurus as candidate concepts. Words retrieving too many entries are ignored.

4. **Candidate evaluation** The retrieved candidates are compared to the original text using a similarity function involving four features: centrality, variation, coverage, and cohesiveness.

5. **Mapping construction** Based on the calculated similarity score, a selection is made of the candidates to assign to each noun phrase and in effect to the original text.

MetaMap does not directly use the UMLS Knowledge Sources as distributed, but applies a number of manually crafted rules and filtering steps to remove entries causing frequent mapping errors and to adjust entries to improve mapping.

The program has a number of advantages and disadvantages. Advantages are its high recall and the flexibility to control its output. A disadvantage is that the many options make it difficult to set up. Moreover, because of its many and complex processing steps, it is slow to use. A third major disadvantage Aronson (2001) also mentions, is MetaMap's inability to cope with the ambiguity of the terminology. The high recall comes at the cost of precision: since the mapping between text and concepts is primarily based on single words these can easily lead to incorrect mappings. The filtering mentioned before can only partially prevent these errors. For example, the noun phrase 'ocular complications' is incorrectly mapped to the concept [Complications Specific to Antepartum or Postpartum] based on the word 'complications'.

In the experiments which will be described later, the output of MetaMap was filtered to concepts which occur in the MeSH thesaurus.

4.2.3 Automatic Term Mapping

PubMed applies "Automatic Term Mapping" (ATM) to improve queries to its search engine automatically.[3]. Query terms which have not been explicitly targeted at a particular MEDLINE field, so-called "Untagged" query terms, are automatically translated and expanded using a number of translation and lookup tables. When, for example, a journal name is encountered, the query is automatically extended with a term searching for the journal in the journal field of the MEDLINE citations. Analogously, untagged query terms are automatically mapped to MeSH terms and author names. Figure 4.2 shows the automatic expansion of the query 'mad cow disease' using ATM.

[3] `http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#AutomaticTermMapping`

```
"encephalopathy, bovine spongiform"[MeSH Terms]
OR ("encephalopathy"[All Fields] AND "bovine"[All Fields]
    AND "spongiform"[All Fields])
OR "bovine spongiform encephalopathy"[All Fields]
OR ("mad"[All Fields] AND "cow"[All Fields]
    AND "disease"[All Fields])
OR "mad cow disease"[All Fields]
```

Figure 4.2: The query "mad cow disease" after Automatic Term Mapping (ATM).

The ATM relieves the user from the overhead of explicitly searching particular fields using the appropriate terms. As a result, ATM can have a strong recall-enhancing effect. However, when an incorrect mapping is made it can lead to confusing results. For example, the single word query 'cell' is automatically expanded with the MeSH term [cells], but also with the MeSH term [cellular phone][4]. For most users, the latter will not be an intended expansion term. Also in search sessions the process can lead to unexpected behaviour. For instance, the subsequent queries 'mad cow' and 'mad cow disease' are expanded with different terms: the first query is expanded with the substance 'mycophenolic adenine dinucleotide', whereas the latter is not.

Neither the exact process of ATM nor the translation tables it uses are publicly documented or available. ATM can be used, however, as a black box system to map text to MeSH terms, which, to the best of our knowledge, only uses the information available in the MeSH thesaurus.

4.2.4 EAGL

Ruch (2006) introduced a retrieval-based system for MeSH classification solely using the information from the MeSH thesaurus called EAGL. EAGL indexes for each MeSH term, its synonyms and description have as a single document in a retrieval index. EAGL classifies a piece of text by issuing it as a query to a vector space retrieval system: the classification of the documents is based on the best ranked MeSH "documents" cut off at a particular retrieval score or rank. The system is available on line[5].

The advantages of this approach are its high speed and small index size. One drawback is that it can return MeSH terms which only have a single word in common with the text to classify. The phrase 'Breast cancer' could, for example, yield the MeSH term [Breast cancer], but also other MeSH terms containing the word 'cancer', such as [Testicular cancer] and [Stomach cancer].

4.2.5 MTI

Aronson et al. (2004) introduced the Medical Text Indexer (MTI). The tool is used to suggest MeSH terms to indexers annotating MEDLINE citations and is provided to registered users by the NLM.

[4]Reported to the NLM on 7 January 2010; the erroneous mapping has been removed from the current PubMed interface.

[5]`http://eagl.unige.ch/EAGL/`

MTI takes a hybrid approach by combining different classifiers, including MetaMap, the "PubMed Related Citations algorithm", and "Restrict to MeSH". Different processing steps including clustering and applying (manually defined) rule-based filtering are used in the system. Parts of the system have been evaluated based on user questionnaires and in a "machine learning setting" (Kim et al., 2001; Aronson et al., 2004). An evaluation against other classification systems or an assessment of its usefulness for information retrieval is missing however. Details of the system can be found on the Semantic Knowledge Representation website[6]. We used MTI as a black box system in our evaluation, using the default settings to obtain MeSH classifications which favours the MeSH term suggested by MetaMap (with weight 7) over the ones from the related citations component (weight 2).

4.2.6 Peregrine

Peregrine was developed by the Biosemantics Group of the ErasmusMC University Medical Center, originally intended as tool for gene name normalisation, that is recognising gene names and mapping them to controlled vocabulary identifiers. The mapping is based on dictionary lookup combined with a number of additional processing steps (Schuemie et al., 2007a). These steps include manually crafted rules to remove erroneous and highly ambiguous terms, rules to generate spelling variations and a method to perform basic disambiguation.

The disambiguation of ambiguous terms is based on rules and keywords found in the surrounding text. The rules define when a term is ambiguous (for example, when it can be mapped to many concepts or when a term is very short), and only assign a concept when a synonym is mentioned in the same document. The keyword-based method is less strict and uses single keywords, that is relatively infrequent words found in other terms used for the concept, to disambiguate terms. Participation in the gene normalisation task of the Biocreative 2 competition resulted in a precision of 75% and a recall of 76% when linking human gene mentions in text to specific genes (Schuemie et al., 2007a). The filtering and disambiguation steps turned out to be important to achieve higher precision at a small loss of recall. The Peregrine system was used to detect UMLS$_{++}$ concepts in text.

4.2.7 Concept language models

In this section, we will propose the first of two classification systems based on statistical language models.

The MeSH thesaurus has already been used extensively to classify MEDLINE citations, so an obvious approach is to use the available manual assignments of MeSH terms to citations as training data to build a classifier.

We propose to use a system based on *Concept Language Models* (CLM). Similar to EAGL, classification is based on a ranked retrieval system. For each MeSH concept, a concept language model is created offline. The CLM is a probability distribution over words which are associated to a MeSH term. The parameters of this language model are based on the titles and abstracts of citations to which the MeSH term has been assigned. Hence, a MeSH term is represented by the text words found in citations relevant to that MeSH term.

[6] `http://skr.nlm.nih.gov`

Formally, the probability of a word w in a CLM θ_M is estimated as follows.

$$P(w|\theta_M) = (1-\lambda)\frac{\sum_{D\in\mathcal{D}_M} f(w,D)}{\sum_{D\in\mathcal{D}_M}|D|} + \lambda P(w|\theta_C) \tag{4.1}$$

Where $\mathcal{D}_M$ is a set of documents assigned to the MeSH term M; $f(w,D)$ is the term frequency of the word w in the document D; and $|D|$ is the length of the document D. The estimation is linearly smoothed with a background language model θ_C. The amount of smoothing is controlled by the parameter λ. For the experiments reported in this chapter, we set λ to a single fixed value for all concepts.

A piece of text is classified by creating a query language model $P(w|\theta_Q)$ for this text and ranking the concept language models using the negated cross entropy, as follows (also see subsection 2.1.3).

$$RSV_{CE}(Q,M) = \sum_{w\in V} P(w|\theta_Q)\log P(w|\theta_M) \tag{4.2}$$

The ranked list of concepts is returned as the classification.

Advantages of this approach to classification are its simplicity and its ability to suggest MeSH classifications which are not explicitly mentioned in the text to classify. The latter can also be a drawback, since it can lead to classifications which are difficult to relate to. The method can be easily extended when new MeSH terms are introduced, by creating additional conceptual language models for these terms individually. Estimating the parameters of new concept language models can be an issue though, since newly introduced MeSH terms have been assigned to only a few citations. Finally, it should be noted that this approach shows close resemblances to a Naive Bayes classifier (Lewis, 1998); a distinct difference is that Naive Bayes classifier incorporates a prior probability of observing a particular class, in this case a particular MeSH term. How this impacts the classifications is further discussed in the evaluations in sections 4.4 and 4.5.

4.2.8 K-Nearest-Neighbours (KNN)

In this subsection, we will propose the second classification system based on statistical language models.

An alternative way to benefit from the available annotations of MeSH terms to MEDLINE citations is to use a K-Nearest-Neighbour (KNN) classifier. KNN classifiers are based on the K-Nearest-Neighbour rule: an unknown pattern is classified with the class of its k nearest neighbour(s) in the training data (Fix and Hodges, 1951; Duda et al., 2000). Note that a classical KNN classifier assigns only a single class to a pattern. The KNN classifier we propose here extends the rule to multi-label classification: an unknown pattern (the text to classify) is classified with the classes assigned to its k nearest neighbours (most similar documents). The classifier is similar to the PubMed Related Citations Algorithm used in MTI (Lin and Wilbur, 2007). KNN is considered for three reasons. Firstly, it can be easily scaled up to such a large classification task. Secondly, we expect it to be useful for document and query classification because its output is based on actual concept-based document representations. The classification is based on coherent groups of concepts related to the text to classify rather than a direct relationship between the text to classify and a model

of the individual concepts. For document classification, its output is therefore expected to look more similar to a typical manual classification. Implicitly, the classification takes into account the rules used for manual indexing. For query classification, such output can be useful, since it is more similar to typical document content. The last reason is practical: in many research environments a full text search system for MEDLINE is already available, making KNN straightforward to implement.

When the KNN approach is used for query classification and the nearest neighbours are found in the collection being searched, the KNN approach can be viewed as a form of pseudo-relevance feedback. The classifier we propose here is inspired by the relevance models and cross-lingual relevance models proposed by Lavrenko and Croft (Lavrenko and Croft, 2001; Lavrenko et al., 2002).

The classification is modelled as follows. We assume to have a document collection $\mathcal{D}$ available in both a conceptual and textual representation. For each document D, we can estimate a textual language model and a conceptual language model, $P(w|\theta_D)$ and $P(c|\phi_D)$ respectively.

For the text to classify (referred to as Q), we wish to estimate a conceptual language model $P(c|\phi_Q)$, which will be used for classification: a list of concepts is returned, ranked according to the (decreasing) probability in this language model. The approximation of the language model is based on the joint probability of observing the concept c with the query Q in the previously introduced document collection $\mathcal{D}$. In words, this approach determines which concepts are most likely to co-occur with the query. Formally, the language model is estimated as follows.

$$P(c|\phi_Q) \approx \frac{P(c,Q)}{\sum_{c'} P(c',Q)} \tag{4.3}$$

Where $P(c,Q)$ is the joint probability of observing a concept c with the query Q.

The joint probability of observing the concept with the query is approximated by independently sampling documents from the collection $\mathcal{D}$, followed by independently sampling the concept and the query from each document.

$$P(c,Q) = \sum_{D \in \mathcal{D}} P(D)\left(P(c|\phi_D)P(Q|\theta_D)\right) \tag{4.4}$$

Where $P(D)$ is a prior probability of sampling the document D from the collection (assumed to be uniform) and $P(Q|\theta_D)$ is the probability of sampling the query from the document, the query likelihood. In the chapter 2, we explained that the query likelihood is commonly approximated by independently sampling the query terms $q_1, \ldots, q_n$ from the document language model (see Equation 2.1 on page 15). The joint probability can therefore be rewritten as follows.

$$P(c,Q) = \sum_{D \in \mathcal{D}} P(D)\left(P(c|\phi_D)\prod_{i=1..n} P(q_i|\theta_D)\right) \tag{4.5}$$

Obviously, requiring the complete collection $\mathcal{D}$ to be processed for classifying a piece of text, makes the model infeasible in practice. The contribution of many documents to

$P(c, Q)$ is relatively small, however, since they are not likely to generate the query ($P(Q|\theta_D)$ is small). Therefore, following Lavrenko and Croft (2001), we can safely reduce this document collection to n documents with the highest probability of generating the query $P(Q|\theta_D)$. In practice, these are the top n documents ranked by query likelihood.

The classification obtained from this approach returns a list of concepts, ranked according to their descending concept language model probability.

4.3 Comparing concepts to text

To obtain a better feeling of the characteristics of a concept-based representation, a collection-level comparison will be made between the word-based representation and the two concept-based representations in MeSH and UMLS$_{++}$.

The statistics were collected from the 2004 and 2006 TREC Genomics collections. The MeSH-based document representation was obtained from the manually indexed MEDLINE citations; for the 2006 collection, which contains full-text articles rather than citations, the MeSH terms assigned to the corresponding MEDLINE citations were used. We only considered MeSH headings, and ignored additional subheadings. For both collections, the UMLS$_{++}$-based representation was automatically obtained using Peregrine. The word-based representation was obtained using the *combined* tokenizer explained in chapter 3. During this tokenization process, stop-words were removed and stemming was applied.

In subsections 4.3.1 to 4.3.3 the global term statistics will be analysed from a document, token and vocabulary perspective, respectively. In subsection 4.3.4, the consequences of using these representations for retrieval will be discussed.

4.3.1 Document perspective

Table 4.1 lists global document length statistics of the 2004 and 2006 collection, respectively. On average, a citation in the 2004 collection is represented by 118 text tokens, 63 UMLS$_{++}$ tokens, and 11 MeSH tokens. Figure 4.3 shows six histograms of the document lengths using different representations. The graphs for the 2004 collection illustrate that quite a few citations only have a title present in the database. Accordingly, the text and UMLS$_{++}$ representations show a peak at small document lengths. The MeSH representation shows a peak at length zero, indicating the almost 50,000 citations which do not have any MeSH terms assigned. Since the UMLS$_{++}$ representation was mapped from the text representation, the document lengths of the two representations are correlated, illustrated by a similar shaped curve in the graphs. The UMLS$_{++}$ document representation however, is considerably smaller.

For the 2006 collection containing full-text articles from Highwire Press the document length is higher for all three representations. On average, a full-text document is represented by 4,501 text tokens (38 times longer), 1,723 UMLS$_{++}$ (27 times longer) tokens, and 15 MeSH tokens. It is remarkable that the MeSH representation on average is longer (15 versus 11), since the collections have been through a similar indexing process and cover a similar date range: the 2006 collection contains publications from between 1995 and 2005, whereas the 2004 collection mostly contains citations from between 1994 and 2004. A likely explanation is that, since these documents are open-access, the indexers have

Table 4.1: Token statistics of the two TREC Genomics document collections used between 2004 and 2007.

(a) 2004 collection (4,591,008 MEDLINE citations)

	Tokens		Token types		Token/type
	Avg.	St. dev.	Avg.	St. dev.	ratio
Text	117.8	87.4	69.6	45.4	1.7
UMLS$_{++}$	62.9	46.2	34.9	22.4	1.8
MeSH	11.4	5.0	11.4	5.0	1.0

(b) 2006 collection (162,259 full-text journal articles)

	Tokens		Token types		Token/type
	Avg.	St. dev.	Avg.	St. dev.	Ratio
Text	4501.4	2052.6	1267.7	460.1	3.6
UMLS$_{++}$	1722.5	851.4	412.6	162.2	4.2
MeSH	15.2	6.2	15.2	6.2	1.0

had access to the full-text versions of the articles, and assigned more terms based on this information.

A second difference between the 2004 and 2006 token statistics is the change in ratio between text tokens and UMLS$_{++}$ tokens. For the 2004 collection the ratio between text tokens and UMLS$_{++}$ tokens is 1.87; for the 2006 collection, it is 2.61. Straightforwardly stated, relatively few UMLS$_{++}$ terms were found in the 2006 collection. A likely explanation is that the citations contain relatively many references to concepts. The full-text may, for example, contain more extensive discussions or references without explicitly mentioning biomedical concepts.

4.3.2 Token perspective

Table 4.2 lists for every representation, the ten most frequent tokens encountered in the 2004 TREC Genomics collection. The most frequent text tokens, such as 'studi', 'patient', and 'effect', clearly illustrate that it is a biomedical document collection. The top MeSH tokens indicate species such as [Human] or [Rats], or particular subject groups such as [Adult], [Aged], and [Middle Aged]. The top UMLS$_{++}$ terms are peculiar and clearly illustrate one of the shortcomings of automatic term mapping using a collection of terminological resources. The most frequently observed UMLS$_{++}$ term was [Donkeys], which clearly does not reflect the contents of a large part of the 2004 collection. The error was caused by the incorrect normalisation of the synonym 'Ass' to 'as', which was frequently encountered in the text. Similarly, the quite specific concepts [Clinical Trials], [Scientific Study], and [DICOM Study] were frequently encountered because they all have the synonym 'study'.

Figure 4.4 visualises the term frequencies of the different representations in the collections, sorted in descending frequency order. Zipf (1949) investigated the distribution of words in text corpora and showed that, for general English, the frequency of a term multiplied by its frequency rank approximates a constant. In practice this implies that a few words account for most of the term occurrences in a document collection, where a

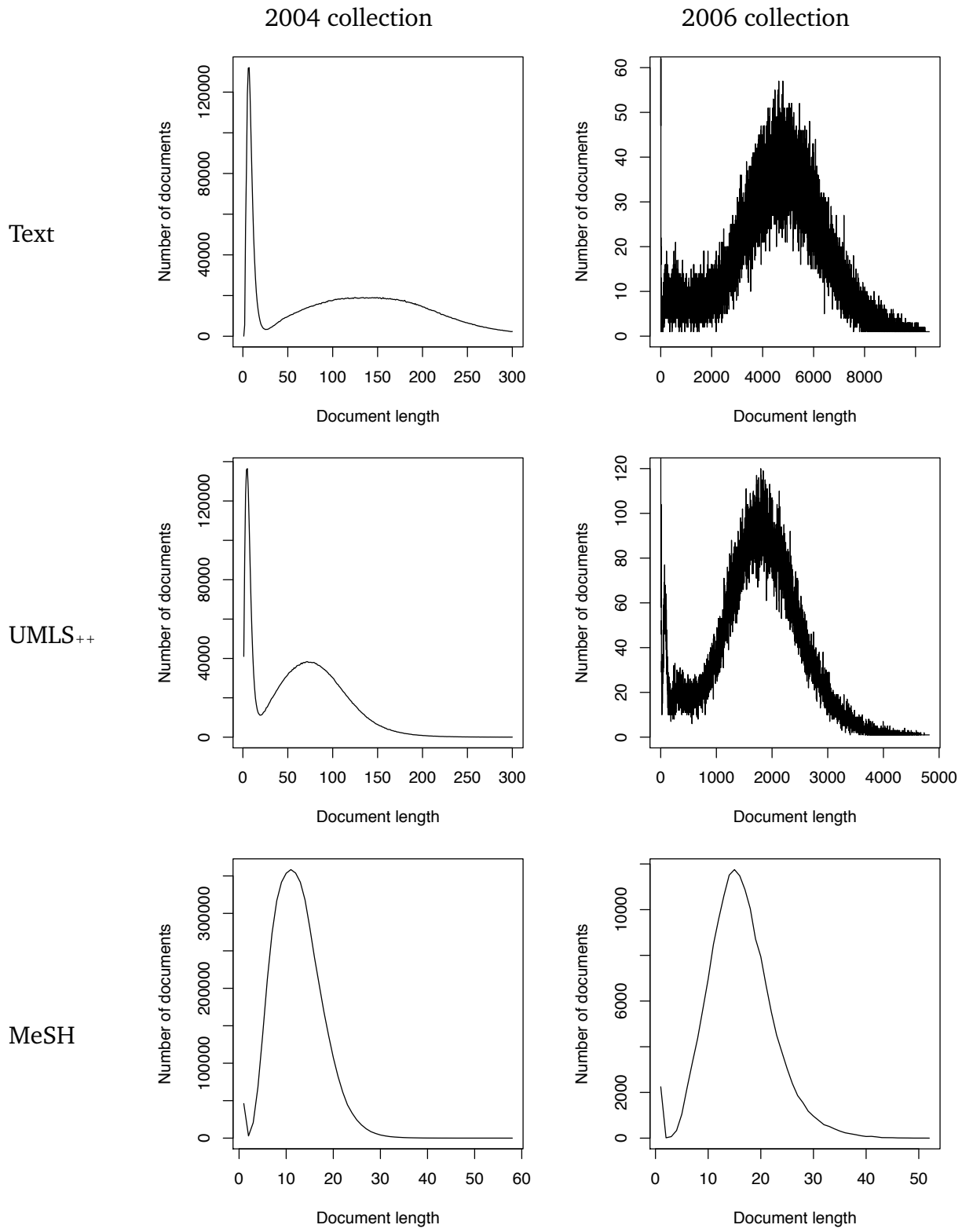


Figure 4.3: Histograms of the document lengths using different representations.

Table 4.2: Most frequently encountered tokens in text, MeSH and UMLS$_{++}$ representation in the 2004 collection.

Text		MeSH		UMLS$_{++}$	
1,617,898	studi	3,217,428	[Humans]	1,698,118	[Donkeys]
1,615,231	or	1,543,128	[Male]	1,658,998	[Clinical Trials]
1,569,398	result	1,529,872	[Female]	1,599,073	[Scientific Study]
1,461,485	1	1,249,478	[Animals]	1,599,073	[DICOM Study]
1,421,786	2	941,740	[Adult]	1,126,118	[Patients]
1,328,495	not	779,915	[Middle Aged]	923,188	[Cells]
1,137,354	patient	569,956	[Aged]	845,465	[Therapeutic procedure]
1,121,724	3	349,324	[Rats]	826,804	[Analysis]
1,066,718	effect	348,493	[Adolescent]	692,125	[Others]
1,047,865	s	288,739	[Child]	677,330	[Disease]

large number of word occurrences are made up by words with a low term frequency: these words are found in the tail of the curves in Figure 4.4.

As expected, large parts of the curve representing the text terms follow a Zipfian distribution (a straight line on log-log scale). Few terms appear with a high frequency and many terms appear with a low frequency. The curve for the 2006 collection is longer (illustrating the larger vocabulary) and has even more terms with a high frequency and many more with a low frequency. The tail of the curve is made up by hapaxes, that is words or rather non-words appearing only once, such as long number sequences, unique DNA sequences, and word number combinations. Many of these term are noise caused by errors in the decoding process from HTML to plain text.

The term frequency histogram of the UMLS$_{++}$ representation initially follows the text curve. At the end of the curve, the UMLS$_{++}$ shows a sharper drop, however, indicating fewer terms with a low frequency.

The MeSH curve shows a much denser frequency distribution: only a few terms have a rather high term frequency (illustrated by a fast drop at the beginning of the curve in comparison to the text and UMLS$_{++}$ representations) and a relatively small proportion of the terms occur with a small frequency (illustrated by a sharper drop at the end of the curve).

4.3.3 Vocabulary perspective

Below, the three representations were analysed from a vocabulary perspective. Heap's law describes the discovery of new terms after viewing increasing numbers of material (Heaps, 1978). According to Heap, the growth of the vocabulary can be described as a function over the number of encountered terms as follows.

$$V(n) = Kn^{\beta} \tag{4.6}$$

Where n is the number of encountered terms and K and β are two collection and vocabulary specific parameters.

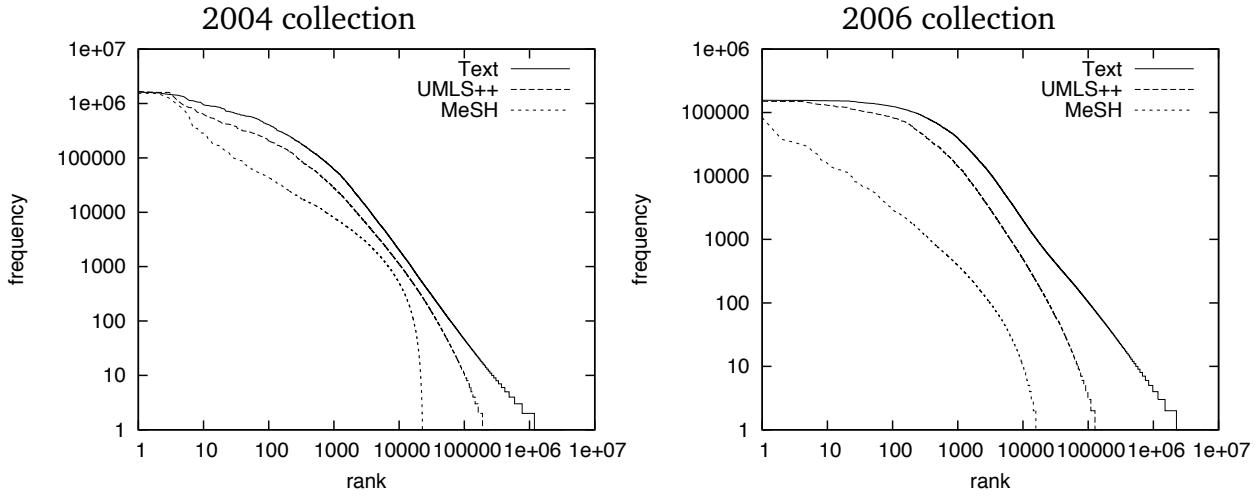


Figure 4.4: Zipf' curves for text, MeSH and UMLS$_{++}$ representations.

Figure 4.5 visualises the vocabulary growth of the three representations after observing more documents from the 2004 and 2006 collections. The MeSH representation has an especially fast growing coverage of the total vocabulary. After viewing 100,000 citations from the 2004 collection, 75% of all vocabulary terms have been encountered. This gives the impression that the MeSH representation of documents is quite diverse. Obviously this is related to the fact that fewer terms have a low document frequency. For the text representation, the vocabulary shows almost a linear growth on the 2006 collection, most likely caused by the noisy terms described earlier. The UMLS$_{++}$ representation shows a large early vocabulary growth on the document collection. This can be explained by the fact that the documents are considerably longer in the 2006 collection.

4.3.4 Consequences for retrieval

The above analysis indicates a number of consequences using these representations for retrieval.

It is clear that the word-based representation is the most exhaustive. In comparison to the concept-based representations, it has a longer tail of terms with a high specificity. Using these terms, one would expect to achieve high precision. To some extent, it is expected that the over-exhaustiveness can be compensated by frequency information: despite the long tail of terms with low frequencies, on average a word occurs several times (1.7 and 3.6 times for the 2004 and 2006 collection respectively). Especially in the case of the full-text collection this frequency information is expected to be beneficial to determine the relative importance of documents for a term.

The UMLS$_{++}$ representation is strongly related to the text representation, but is more compact. Synonymous terms are grouped at the document level, allowing for increased recall during search. The top terms do give the impression, however, that many incorrect mappings have been made. How this affects retrieval is unclear and also depends on how the queries are mapped to UMLS$_{++}$; if both queries and relevant documents are mapped to the same incorrect representation, even an erroneous representation can improve retrieval. This depends on the amount of ambiguity introduced by such incorrect mappings. Depending

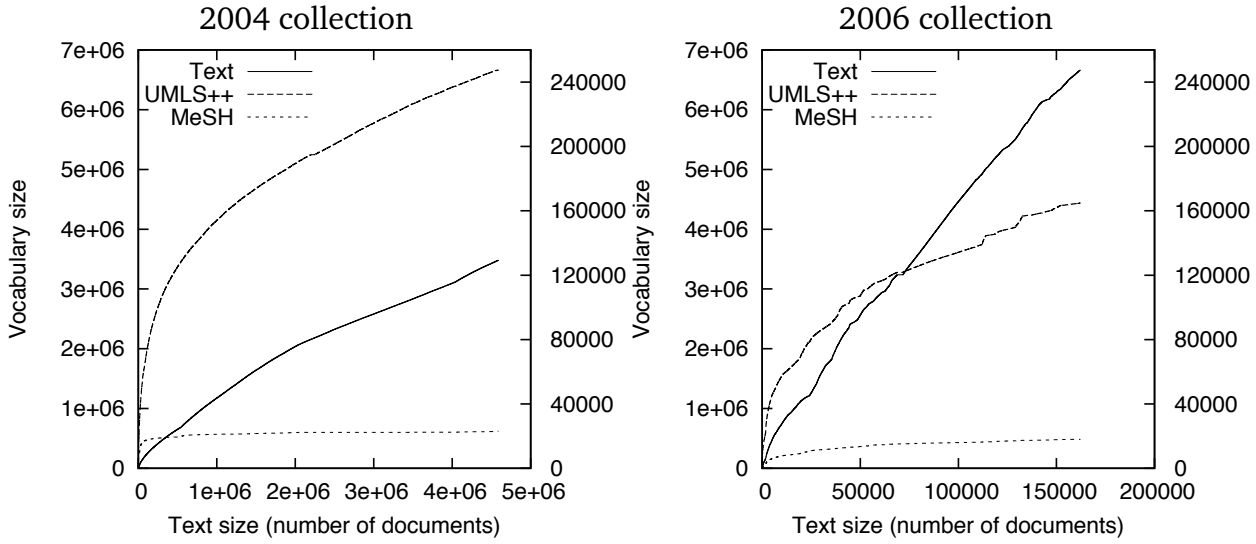


Figure 4.5: Illustration of Heap's law for text, MeSH and UMLS$_{++}$ representations. Note that the lines representing MeSH and UMLS$_{++}$ tokens use the right y-scale, whereas the line representing text tokens uses the left y-scale.

on the query, this might not be a problem after all (Sanderson, 1994; Stokes et al., 2009). Whether the documents are completely covered by the UMLS$_{++}$ vocabulary remains unclear: if important aspects are not represented by UMLS$_{++}$ concepts this is likely to hurt retrieval.

The MeSH representation is relatively short and offers a coarse representation in a small vocabulary. The manual assignment of MeSH terms is not exhaustive, but is quite precise: the assignment of a term indicates a high importance of the term. Advantages of the MeSH document representation are that it is unambiguous and that is has been manually assigned: assuming that manual index errors were not often made, the assignment of a term to a document is reliable information that the document is about that concept. It is difficult to anticipate the added value of the representation. On the one hand, it allows for high precision, limited however by the coarse granularity of the vocabulary. On the other hand, its lack of exhaustiveness will hurt recall: relevant documents which discuss a concept marginally are not represented by MeSH concepts. By searching with a more general concept from the MeSH hierarchy recall can be increased but this still will not find concepts marginally discussed in the documents.

4.4 Document classification

In section 4.2, we introduced a number of out-of-the-box text classification systems and proposed two classification systems based on statistical language models. In this section, we will compare a selection of these systems on their *automatic document classification*, or automatic indexing performance.

The goal of this evaluation is to determine to what extent the classification systems can translate a textual document representation to a concept-based representation useful for IR.

We will only consider the MeSH document representation vocabulary, simply because for MeSH a large set of manually curated documents is available. The MeSH classification

systems MetaMap, MTI, EAGL, Concept Language Models (CLM), and KNN will be compared. Automatic Term Mapping will not be evaluated, since it only supports classification of short queries.

We expect that the nearest neighbour approaches (KNN and MTI) perform well on this task: their output is expected to resemble manual document classification since their output is based on existing document classifications. Especially, MTI is expected to perform well, since it is actively used as a recommendation system for manual MeSH term assignment and has been geared towards this task. For this reason, MTI is used as a baseline in this evaluation. The methods based on string matching (MetaMap, CLM, EAGL) and dictionary lookup are expected to perform worse, since their output does not typically adhere to the style of manual document classification. In this category, EAGL is expected to perform worst, since its approach (word-based matching) and information used for classification (only the thesaurus) are limited. CLM uses a similar word-based matching approach but uses much more information (language models for each concept-based on classified documents) for its matching. We are ambivalent about the performance of MetaMap. Despite its sophisticated approach to classification, including noun phrase detection and extensive term variant generation, its mapping process is still limited to information in the thesaurus.

In this section, we will answer the following three research questions.

RQ2.1: *To what extent can manual document classification be reproduced by automatic classification?*

RQ2.2: *What kind of errors are made by the different types of classification systems?*

RQ2.3: *Is there added value of automatic document classification over manual classification?*

The overview of this section is as follows. The experimental setup is described in subsection 4.4.1. The classification results of the different systems will be analysed in subsection 4.4.2, followed by a discussion in subsection 4.4.3.

4.4.1 Experimental setup

A commonly used method to evaluate MeSH text classification is to see how well a classifier reproduces the manual annotations of MEDLINE citations (Lewis et al., 1996; Ruiz and Srinivasan, 2002; Ruch, 2006). Selected citations of the OHSUMED collection (Hersh et al., 1994a) have been used as training and test data, but as Ruiz and Srinivasan (2002) noted, different test collections and variable numbers of categories have been used, making comparisons difficult. Moreover, the OHSUMED collection is not up-to-date anymore. At the time of its creation, the MeSH thesaurus consisted of around 14,000 MeSH terms. The 2008 edition of the thesaurus contains around 24,000 terms, making an evaluation using OHSUMED not representative for the current state of the MeSH thesaurus. Similar to the approach taken by Ruch (2006), we sampled a random set of one thousand citations from the 2008 MEDLINE baseline distribution (consisting of more than 16 million citations) and used these citations as a test set. The selected citations were required to have at least one MeSH term assigned to each of them. The list of citations can be downloaded for followup research[7]. The test set covers 3951 distinct MeSH terms (9596 assignments).

[7] `http://www.ebi.ac.uk/~triesch/meshup/testset_v1.xml`

To keep training and testing data separated, the test collection was excluded from the collection used for building the concept language models (used for CLM) and the set of neighbouring citations (used for KNN). The remaining citations in the 2008 baseline distribution were used for training: to build an index for the KNN approach and for sampling citations (at most 1000 citations per MeSH term) to build the concept language model for each MeSH term.

The metrics used to evaluate the suggested indexing terms will be explained in the following block.

Evaluation metrics

Lam et al. (1999) described three types of metrics to compare the ground truth to the output of the classification systems: document, category and decision perspective metrics.

The document perspective metrics evaluate the assignment of MeSH terms at the document level. Since most of our classification systems rank the suggested MeSH terms, evaluation metrics can be borrowed from IR evaluation: rather than retrieving relevant documents for a query, relevant MeSH terms have to be retrieved for a citation. Accordingly, the Mean Average Precision (MAP) and Precision at 10 (P@10) can be used as document perspective metrics. It should be noted however, that MAP favours systems with high recall. Ranked classification systems, such as EAGL, CLM and KNN retrieve many more MeSH terms in comparison to MTI and are therefore more likely to achieve a higher recall. This should be taken into account when observing MAP scores. However, rank precision (precision at 10) does give a clear indication of the performance when only a few top terms are considered.

The category perspective metrics suggested by Lam et al. (1999) are the (macro) F-measure, Precision, and Recall for each MeSH term. They are defined as follows.

Given:

$a = \#$ documents assigned to the category both manually and automatically

$b = \#$ documents assigned to the category automatically but not manually

$c = \#$ documents assigned to the category manually but not automatically

Precision (P), recall (R), and F_β-measure are defined as follows.

$$\mathrm{P} = \frac{a}{a+b}, \qquad \mathrm{R} = \frac{a}{a+c}, \qquad F_\beta = \frac{(1+\beta^2)\mathrm{PR}}{\beta^2\mathrm{P}+\mathrm{R}} \tag{4.7}$$

Finally, the decision perspective metrics are the micro-averaged F-measure, Precision, and Recall. They are based on the number of correct and incorrect decisions a classification system makes, where each possible document and category pair form a decision. Lam et al. (1999) defined them as follows.

Given:

$p = \#$ assignments made automatically and manually

$q = \#$ assignments made automatically

$r = \#$ assignments made manually

Micro precision (P*), micro recall (R*), and micro F_β-measure are defined as follows.

$$\mathrm{P}^* = \frac{p}{q}, \qquad \mathrm{R}^* = \frac{p}{r}, \qquad F^*_\beta = \frac{(1+\beta^2)\mathrm{P}^*\mathrm{R}^*}{\beta^2\mathrm{P}^* + \mathrm{R}^*} \tag{4.8}$$

Both the F-measure and micro F-measure require a discrete number of classifications per instance and our classifiers return a ranked list of classes. Similar to Lam et al. (1999), the measures will be reported using optimal cutoff values, by assuming the number of top classes which yields the highest average F-measure. These optimal cutoff values were determined on a per-system basis. Hence, the results of the category and decision perspective metrics should be regarded as upper bounds for the systems.

4.4.2 Results and analysis

To illustrate the output from the evaluated classification systems, in appendix C.1 an example MEDLINE citation with the output of the evaluated systems is listed.

RQ2.1: *To what extent can manual document classification be reproduced by automatic classification?*

Table 4.3 lists the classification results of the different systems when presented with either the title alone, or with both title and abstract of the 1000 MEDLINE citations. Figure 4.6 shows the PR-curve of the recall and precision from document perspective.

MTI served as the baseline to compare the other systems to and turned out to perform reasonably on the classification task. On average, two of its top 10 suggestions corresponded to the manual term assignments. The results indicated that MTI is sensitive to the amount of input provided. All four evaluation measures show large increases (between 54% and 77%) when presented with both the title and abstract of a citation, rather than the title alone. This difference is also clearly visible in the PR-curves.

MetaMap performed worse than MTI on all metrics. It shows that, since MetaMap is a component of MTI, MTI strongly benefitted from its other sources of MeSH terms for classification. We expected that MetaMap would benefit more from longer input, since more text would provide more noun phrases to detect MeSH terms in. In contrast, the improvements were relatively small (up to 37%).

EAGL and CLM performed similarly to or slightly worse than MetaMap when presented with only the title of the citation to classify. They both performed considerably worse than MTI when the abstract was also available for classification. The performance of CLM remained almost the same when the abstract was also available. EAGL, which uses a similar approach showed larger improvements when this information was provided.

KNN performed surprisingly well in comparison to the other systems: it showed 99% improvement in terms of MAP, 41% improved precision at 10 (P@10) and 12% improvement in micro F_1 over MTI when presented with the title and abstract of the citations. On average, more than four of KNN's top 10 terms corresponded to manual classification, whereas MTI returned slightly over three matching MeSH terms. In terms of Category F_1, KNN performed 10% worse than MTI: when considering one MeSH term, MTI was better in choosing whether to assign it to a citation or not. Considering the performance from a document perspective, KNN outperformed MTI: given the title and abstract of a citation,

Table 4.3: MeSH classification performance on 1000 random MEDLINE citations.

(a) Title used as input

	Document				Category		Decision	
	MAP		P10		F_1		micro F_1	
MTI	0.1625		0.1809		0.2663		0.2859	
MetaMap	0.1426	-12%	0.1735	-4%	0.2330	-13%	0.2660	-7%
EAGL	0.1722	+6%	0.1800	-0%	0.2413	-9%	0.2588	-9%
CLM	0.1763	+8%	0.1690	-7%	0.3326	+25%	0.2877	+1%
KNN	**0.4795**	+195%	**0.4326**	+139%	**0.3693**	+39%	**0.4758**	+66%

(b) Title and abstract used as input

	Document				Category		Decision	
	MAP		P10		F_1		micro F_1	
MTI	0.2536		0.3200		**0.4503**		0.4415	
MetaMap	0.1623	-36%	0.1910	-40%	0.3187	-29%	0.2968	-33%
EAGL	0.1976	-22%	0.2119	-34%	0.2987	-34%	0.2977	-33%
CLM	0.1783	-30%	0.1748	-45%	0.3429	-24%	0.2982	-32%
KNN	**0.5052**	+99%	**0.4515**	+41%	0.4074	-10%	**0.4963**	+12%

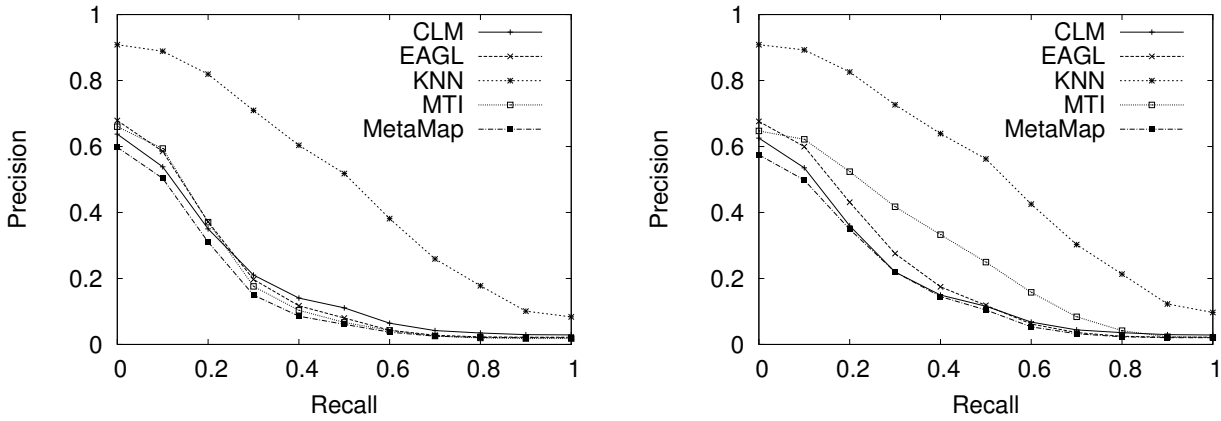
KNN found more correct/manual MeSH terms and ranked them higher. KNN showed to be less sensitive to the amount of available information than the other systems: the metrics increased between 4 and 10% when the abstract was available for determining the classification. The improvements (up to 15% MAP) were only small, however, compared to the increased length of the input text (16 times as many words on average). Therefore, one might argue for the use of only the title for automatic annotation using KNN.

Performance on specific and general MeSH terms

RQ2.2: *What kind of errors are made by the different types of classification systems?*

Table 4.4 shows the category F_1 performance measure for MeSH terms organised according to specificity. For example, for very specific MeSH terms, that is terms with a relatively low document frequency in MEDLINE (between 0 and 1000), the average F_1 measure for KNN is the highest with 0.5578. The table clearly shows that MetaMap, CLM and EAGL performed relatively well on specific MeSH terms. KNN and MTI however, were capable of reproducing both general and specific MeSH terms. MTI especially was shown to perform more consistently across the range of more or less specific terms.

Figure 4.6: Document perspective PR-curves of MeSH classification.



(a) Title used as input

(b) Title and abstract used as input

Table 4.4: Classification performance in terms of category F_1 grouped according to more or less specific MeSH terms (classifications based on title and abstract as input).

DF in MEDLINE	# concepts in test set	MetaMap	CLM	EAGL	KNN	MTI
0-1000	196	0.4099	0.4955	0.3889	0.5578	0.5408
1,000-5,000	975	0.3546	0.4513	0.3295	0.4630	0.4628
5,000-10,000	766	0.3196	0.3955	0.3011	0.4391	0.4753
10,000-50,000	1653	0.2953	0.2734	0.2755	0.3585	0.4327
>50,000	361	0.2774	0.1736	0.2679	0.3319	0.3954
>1,000	3755	0.3140	0.3349	0.2940	0.3995	0.4456
>5,000	2780	0.2997	0.2941	0.2816	0.3772	0.4396
>10,000	2014	0.2921	0.2555	0.2742	0.3537	0.4260

False positives analysis

RQ2.3: *Is there added value of automatic document classification over manual classification?*

Despite the fact that manual annotations of MEDLINE are carefully created and that on average the most important terms are assigned, it should be noted that using these manual annotations for evaluation is an idealisation. Manual annotators do accidentally assign irrelevant MeSH terms or miss relevant terms. To investigate this issue, we asked an experienced annotator to judge some of the false positives, that is automatic annotations which are not in the set of manually assigned terms. For 50 of the 1000 citations in the test, the annotator judged the three highest ranking false positives from MetaMap, CLM, and KNN[8] on a 5-point scale. To test the reliability of the annotator three manual annotations were added to each citation as well. For each of the 50 citations, the title and abstract were presented with 12 (9 false positives and 3 true positives) randomly ordered MeSH terms. Each MeSH term was then judged on a 5-point scale ranging from "Strongly irrelevant/Incorrect" to "Strongly relevant". This scale can be found in Appendix C.3. The analysis provides additional insights into the performance of the different classification systems. Some of the automatically identified terms may have been judged as irrelevant (false positives), because they were not included in the original MeSH annotations. By taking a closer look, however, we may actually find them to be highly relevant, that is appropriate to represent the text to classify.

Table 4.5 shows the results of this annotation process. The first column of Table 4.5 shows that in 88% of the cases our annotator judged the original MeSH annotations as (very) relevant. Using more common inter-annotator agreement measures, such as Cohen's Kappa is not applicable in this case, since we do not know which MeSH terms were explicitly labeled as non-relevant by the NLM's indexers. The high percentage does, however, indicate a strong agreement between the NLM annotators and our annotator.

In general we noticed that a fair share of the false positives was judged "Relevant" or better (58% for MetaMap, 37% for CLM and 34% for KNN), indicating that automatic annotations do contribute relevant terms in addition to manual annotations.

Despite MetaMap's relatively poor performance on reproducing the exact manual annotations, the results show that in many cases its terms were useful for representing the text (58% of its false positives are judged as "Relevant" or better). Only few false positives (3%) were indicated as totally incorrect. Compared to CLM and KNN, only few terms (14,7%) were labeled "Undecided". This is because MetaMap requires an almost direct link between words in the text to classify and the MeSH terms it suggests. As expected, quite a few terms were suggested of which only some can be related to the text to classify.

The largest proportion of the false positives from the CLM system were judged as "Undecided" (35.5%). The system returned too many specific terms and some of the suggestions could not be directly linked to the text to classify. For KNN, most of the false positives (31%) were indicated as "Irrelevant". This value can be explained because KNN frequently returned general terms which were found in similar documents but were not appropriate to the text to classify.

[8]Restricted to these systems because of resource limitations.

Table 4.5: Results of the analysis of false positives.

Judgment	True positives		False positives MetaMap		CLM		KNN	
Very relevant	94	75%	40	29%	44	24%	37	20%
Relevant	17	13%	39	29%	26	14%	27	14%
Undecided	12	10%	20	15%	66	35%	49	26%
Irrelevant	1	1%	33	24%	35	19%	58	31%
Incorrect	2	2%	4	3%	16	9%	17	9%

4.4.3 Discussion

Our MeSH classification experiments clearly indicated the limitations and advantages of the different methods.

EAGL and MetaMap were limited in their capability to produce general MeSH terms or indirectly related terms. The false-positive analysis underlines that it is easy for the user to link the suggested concepts to the text through the words that they share. Advantages of EAGL are its classification speed and moderate index size. Many general terms were missed and incorrect terms were suggested only on the basis of a partial match with the text to classify, however.

The system based on Concept Language Models required a large amount of training data but was straightforward to train. The system performed on a par with the EAGL system and returned specific classifications. The false positive analysis confirmed that the method returned relevant classifications which could only be related indirectly to the text to classify. Again these methods failed to produce general MeSH terms. We expect that a better trade-off between general and specific MeSH terms can be accomplished by adding a prior to the CLM system, similar to Kraaij et al. (2002).

The classification system based on similar documents (KNN) showed the best trade-off between general and specific MeSH terms. It strongly outperformed the other classifiers in reproducing manual annotations. Documents related to the text to classify, yielded not only relevant specific MeSH terms, but also potentially very relevant general MeSH terms. In addition, relevant terms were returned which were not explicitly mentioned in the text. Some of the drawbacks include its classification speed (around a second per abstract on a desktop system) and the required index size. Finally, quite a few of the false positives were either irrelevant or incorrect, due to general MeSH terms which were appropriate for related documents, but not for a document in particular.

We note that the false positive analysis might be biased in favour of the thesaurus-oriented classifiers. For both KNN and CLM, it was more difficult to judge a false positive if part of the suggested MeSH term did not occur in the text. This would favour the thesaurus-oriented approaches, since they rely on more explicit overlap. Moreover, we note that our annotator did not have access to the same information as the annotators responsible for the MEDLINE annotations; the latter are (sometimes) provided with the full-text of the citation under annotation as well.

The false positive analysis indicated the value of automatic classification: quite a few of the suggested MeSH terms which did not correspond to the manual annotation

turned out to be relevant for the citations. Despite the fact that these terms might not be correct according to the official NLM's indexing practice, they might be useful for an extended conceptual document representation useful for IR. However, the inevitable errors of automatic classification might even hurt retrieval.

4.5 Query classification

In the previous section, we investigated the performance of the two proposed classification systems together with a selection of out-of-the-box classifiers on document classification. Now, we will turn to analysing their ability to classify queries. In this section, we will investigate both MeSH and UMLS$_{++}$ query representations and compare them using all of the classification systems described earlier. We will use the different classifiers to obtain a conceptual query representation from a text query. This conceptual query will subsequently be used to retrieve documents. The classification systems will be evaluated based on the resulting retrieval performance. The first part of this evaluation will be split into two experiments. In the first experiment, the retrieval performance will be determined when using the concept-based query representation on its own for retrieval. We expect that such an approach will perform poorly in comparison to word-based retrieval, since the concept-based representations are limited in their ability to express information needs completely. The experiment does provide valuable information, however, to compare the relative performances of the classification systems. In the second experiment, we will investigate a word-based query representation combined with an automatically obtained concept-based representation. We expect that word-based retrieval benefits from such an expanded representation. In particular, we expect MeSH, because of its limited specificity, to be a recall enhancing representation. For the UMLS$_{++}$ representation, we also expect a precision enhancing effect: the representation is more fine-grained than a word-based representation.

We will answer the following research questions in this section.

RQ2.4: *What is the effectiveness of concept-only retrieval?*

RQ2.5: *Can an automatically obtained concept-based query representation improve word-based retrieval?*

The overview of this section is as follows. In subsection 4.5.1 we will describe the experimental setup. In subsections 4.5.2 and 4.5.3 the results of the two experiments will be analysed and discussed. In subsection 4.5.4 we will report results from an additional experiment motivated by the results from the first two experiments. Subsection 4.5.5 contains a section conclusion.

4.5.1 Experimental setup

Again, the TREC Genomics benchmark collections used between 2004 and 2007 were used for retrieval experiments (see subsection 2.2.5). Similar to the previous chapter, Mean Average Precision and Precision at 10 were used as evaluation measures. In the following blocks, the retrieval system and representations for queries and documents will be described.

Retrieval system

A basic language model retrieval system was used for concept-based retrieval. In this model, documents are ranked according to the negated cross entropy between query and document concept language models. For retrieval using both text-based and concept-based representations, the retrieval scores were linearly combined as follows.

$$RSV(D,Q) = -\alpha H(\phi_Q|\phi_D) - (1-\alpha)H(\theta_Q|\theta_D) \tag{4.9}$$

Where ϕ_Q and ϕ_D are the concept language models of query and document respectively; θ_Q and θ_D are the text language models described in subsection 2.1.3; α, with a value between 0 and 1, controls the importance of either representation: when set to 1, only the concept representation is used, when set to 0, only the text-based representation is used. Other fusion methods were investigated as well (see appendix C.4); interpolation turned out to be an effective method to combine the results.

Analogue to text-based language models, the parameters of the document concept language models were based on a smoothed maximum likelihood estimate.

$$P(c|\phi_D) = (1-\lambda_c)\frac{f(c,D_C)}{|D_C|} + \lambda_c P(c|\hat{\phi}_{\mathcal{C}}) \tag{4.10}$$

Where $f(c, D_C)$ is the concept term frequency of the concept c in the conceptual representation of the document $D_{\mathcal{C}}$; $|D_{\mathcal{C}}|$ is the total number of concepts in the conceptual representation of the document (the concept document length); $P(c|\hat{\phi}_{\mathcal{C}})$ is the probability of the concept in the collection (estimated similar to the text-based language model).

It should be noted that for MeSH, the concept term frequency in a document is never higher than 1, since a MeSH term is either assigned to a document or not. Consequently, the unsmoothed concept document language model is a uniform distribution over the MeSH terms assigned to that document. For UMLS$_{++}$, this is not the case: concepts might have been assigned multiple times to (different) words and phrases found in the document, also allowing concept term frequencies higher than one.

Query representation

Both MeSH and UMLS$_{++}$ classification were investigated. For MeSH, the queries were automatically classified using MetaMap, Automatic Term Mapping (ATM), EAGL, CLM, MTI, and KNN. For UMLS$_{++}$, KNN and Peregrine were used.

The parameters of the conceptual query language model $P(c|\phi_Q)$ were based on the relative score the classification method assigned to the concepts.

$$P(c|\phi_Q) = \frac{s(c,Q)}{\sum_{c' \in C} s(c',Q)} \tag{4.11}$$

Where $s(c, Q)$ is the classification score assigned to concept c and $\sum_{c' \in C} s(c', Q)$ is the sum of scores assigned to all classified concepts. In the case that no scores were assigned to the returned concepts (for example for ATM), all scores were assumed to be 1. In the case of KNN, the parameters were estimated as explained in subsection 4.2.8.

Document representation

The parameters of the MeSH-based document language models, $P(c|\phi_D)$, were based on the manual MeSH index terms assigned by NLM indexers. The conceptual document representation in terms of UMLS$_{++}$ terms, was totally based on automatic mapping using Peregrine. As a result, this representation contained errors. These representations have been analysed in section 4.3.

4.5.2 Concept-only retrieval

RQ2.4: *What is the effectiveness of concept-only retrieval?*

Table 4.6 lists the retrieval effectiveness the systems achieved when using only an automatically obtained conceptual representation for searching. This corresponds to setting α in Equation 4.9 to 1. As a baseline, we used a word-based retrieval system using the combined tokenization method described in chapter 3.

The message is clear: searching and matching solely with an automatically obtained conceptual query representation is by far outperformed by using a word-based representation alone. Especially in the case of MeSH, this could have been expected: a vocabulary of around 24,000 concepts is probably not extensive enough to express precise aspects of every information need.

EAGL performed poorly using its classifications for retrieval. Its approach simply yielded too many unrelated terms. On only a single topic, the concept representation performed better than the textual representation. In this case the (3) relevant documents indeed ranked higher than using the textual query. For the particular topic (topic 178, see Appendix A), many of the suggested concepts were associated to insulin which corresponds to a major aspect of the query. The resulting average precision of 0.09 was still quite low however.

ATM performed much better than EAGL, but was still far behind the text-based baseline. ATM's queries were considerably shorter but often contain incorrect mappings. For only 2 (out of 164) topics, the method performed on par or better than the text-based baseline. In these cases, all aspects of the query were represented by MeSH terms, with no incorrect mappings.

The system based on CLM performed slightly better than ATM, but also stayed far behind the baseline. A per topic analysis showed that CLM performed better than the baseline on only a single topic (topic 160, see Appendix A). The topic contains two aspects, the gene/protein PrnP and mad cow disease, but since these two are inherently related the query can be represented by concepts referring to either aspects.

MTI and KNN exhibited the best MeSH-only retrieval but again for only few (both 6) topics did they show improvements over the baseline in terms of MAP. Interestingly, however, KNN did not show a significant decrease at P@10 compared to the baseline.

The two systems using the UMLS$_{++}$ thesaurus, Peregrine and KNN, also stayed behind the text-only baseline, but performed better than most runs based on the MeSH-based representation. For 37 (out of 164) topics, the queries based on a UMLS$_{++}$ representation from Peregrine outperformed the text-based representation. For the representation based on KNN, this was the case for 48 topics. It appears that especially capturing specific phrases was beneficial: topics containing phrases such as 'antibody activity' (topic 17), 'nerve growth factor pathway' (topic 44), 'Nerve Growth Factor' (topic 44), 'coronary artery disease' (topic

205), and 'signal recognition particle' (topic 226) which were treated as a single concept showed improvements.

Based solely on MAP and P@10 performance, one may conclude that these representations cannot contribute that much to word-based retrieval. Table 4.7 shows that the concept-based representations can indeed add something to retrieval: it shows the number of relevant documents retrieved by the concept-only representation but which were not retrieved by the text-based representation. The KNN approaches and Peregrine performed especially well in retrieving documents not found with the word-based representation. The table shows that the representations are indeed useful for enhancing recall.

4.5.3 Combining concepts with text

RQ2.5: *Can an automatically obtained concept-based query representation improve word-based retrieval?*

The results in the previous section showed that, on its own, a concept-based representation is too limited to represent information needs completely and as a result, cannot outperform word-based retrieval. The concept-based representation does however, retrieve relevant documents which were missed by a word-based representation. A text-based representation might therefore be improved by combining it with a concept-based representation. Table 4.8 lists the retrieval performance of such a combined approach. The table lists the retrieval performance for the values of α which result in the highest MAP for that system (the value of alpha was varied between 0.05 and 0.95 with a step size of 0.05).

In general, it could be observed that a combined representation can result in improved retrieval effectiveness up to 9.9% in MAP and 5.7% in P@10. The results on the 2004 collection appeared to benefit more from a conceptual representation than the results on the 2006 collection. For the 2006 and 2007 topics, no significant improvements could be observed using a combination with the MeSH-based representation. For the 2004 and 2005 topics, the improvement of MeSH depended on the classification method: the two methods based on a KNN approach (KNN (MeSH) and MTI) showed significant improvements. Except for the method based on CLM, none of the string matching methods using MeSH (MetaMap, ATM, and EAGL) showed significant improvements over the baseline. Using the representation based on UMLS$_{++}$ resulted in significant improvements for the 2004, 2005, and 2006 collections, but no classification method consistently yielded significant improvements on all topic sets.

4.5.4 Combining blind feedback

The KNN approach or blind feedback approach discussed in subsections 4.5.2 and 4.5.3 performed well in comparison to a text-only baseline *without* text-based feedback. The improvements might therefore be explained from the expansion effect using pseudo-relevant documents, rather than the use of a concept-based representation. To assure that the improvement was indeed caused by the individual concept-based representations, the feedback methods using different representations were combined.

Table 4.9 shows the result of combining different feedback methods. Note that as a baseline, text-only *with* text-based feedback was used. In all these experiments, the

Table 4.6: Retrieval effectiveness when only using the conceptual query representation for retrieval. [1], [2] and [3] indicate significant differences to the baseline at confidence levels 0.05, 0.01 and 0.001 respectively, determined with a paired sign test. The highest value of each column is printed in boldface.

(a) 2004 and 2005 queries

	2004		2005	
	MAP	P@10	MAP	P@10
baseline	**0.3576**	**0.5800**	**0.2219**	**0.3551**
MetaMap	0.0169 [3] -95.3%	0.0620 [3] -89.3%	0.0228 [3] -89.7%	0.0551 [3] -84.5%
ATM	0.0173 [3] -95.2%	0.0460 [3] -92.1%	0.0265 [3] -88.1%	0.0653 [2] -81.6%
EAGL	0.0031 [3] -99.1%	0.0100 [3] -98.3%	0.0034 [3] -98.5%	0.0102 [3] -97.1%
CLM	0.0277 [3] -92.2%	0.1020 [3] -82.4%	0.0250 [3] -88.8%	0.0571 [3] -83.9%
KNN (MeSH)	0.1889 [3] -47.2%	0.4380 -24.5%	0.1268 [3] -42.8%	0.2878 -19.0%
MTI	0.0357 [3] -90.0%	0.1460 [3] -74.8%	0.0596 [3] -73.1%	0.1531 [1] -56.9%
Peregrine	0.1630 [3] -54.4%	0.3600 [1] -37.9%	0.0857 [3] -61.4%	0.1878 -47.1%
KNN (UMLS$_{++}$)	0.2799 [2] -21.7%	0.5120 -11.7%	0.1670 [3] -24.7%	0.3286 -7.5%

(b) 2006 and 2007 queries

	2006		2007	
	MAP	P@10	MAP	P@10
baseline	**0.3889**	**0.4769**	**0.2796**	**0.4500**
MetaMap	0.0646 [3] -83.4%	0.1154 [2] -75.8%	0.0278 [3] -90.1%	0.0500 [3] -88.9%
ATM	0.0888 [3] -77.2%	0.1077 [2] -77.4%	0.0256 [3] -90.8%	0.0528 [3] -88.3%
EAGL	0.0208 [3] -94.7%	0.0731 [3] -84.7%	0.0161 [3] -94.2%	0.0583 [3] -87.0%
CLM	0.1071 [3] -72.5%	0.1538 [3] -67.7%	0.0390 [3] -86.1%	0.1028 [3] -77.2%
KNN (MeSH)	0.2518 [3] -35.3%	0.4077 -14.5%	0.1901 [3] -32.0%	0.3750 -16.7%
MTI	0.1059 [3] -72.8%	0.2231 [1] -53.2%	0.0607 [3] -78.3%	0.1722 [2] -61.7%
Peregrine	0.3085 [1] -20.7%	0.4000 -16.1%	0.1619 [2] -42.1%	0.3250 -27.8%
KNN (UMLS$_{++}$)	0.3535 -9.1%	0.4692 -1.6%	0.2355 [2] -15.8%	0.4222 -6.2%

Table 4.7: Unique relevant documents retrieved by concept-only approaches that were not retrieved by the text-only baseline.

	2004	2005	2006	2007
Relevant documents	8,268	4,584	1,449	2,490
MetaMap	290	198	23	46
ATM	255	179	16	27
EAGL	192	108	17	37
CLM	334	237	115	146
KNN (MeSH)	**725**	**468**	94	198
MTI	371	302	31	117
Peregrine	527	253	119	196
KNN (UMLS$_{++}$)	710	382	**133**	**224**

weight of the original text-based query was set to 0.5, the remaining 0.5 weight was evenly distributed over the query representations obtained from feedback. These values were based on earlier experiments with pseudo-feedback, where for both text and concept-based feedback these values performed well. We expect that careful tuning of these weights can further improve the combined results, but that the relative contribution of each representation will remain the same.

Except from the 2006 topic set, it turned out to be quite effective to combine pseudo-feedback from the three different representations: the combination of all three representations performed best for the 2004 and 2005 topic sets and close to best for the 2007 topic set. Feedback with only MeSH and UMLS$_{++}$ representations did significantly outperform text-only feedback on the 2004 and 2005 topic sets.

From these results we may conclude that concept-based feedback does give an additional contribution to conventional text-based feedback.

4.5.5 Section conclusion

After all these experiments, the question is what conclusions may be drawn about the usefulness of a concept-based representation for retrieval.

The experiments with concept-only retrieval and single term queries showed the limitations of the used concept-based representation. On average, concept-only retrieval stayed far behind word-based retrieval. In many cases no suitable concepts were available to represent the information need and even when such concepts were available a single word-based representation could outperform it. This does not make a conceptual representation completely useless, since it did help retrieve documents which were not retrieved by the text-based representation on its own. Occasionally, concept-based retrieval did perform better when all query aspects were represented in terms of concepts and no or few incorrect concepts were added to the query.

As expected, the quality of the process of obtaining a concept-based representation turned out to be important for its use to improve IR. For the MeSH representation, string matching turned out to perform poorly both on concept-only retrieval and combined

Table 4.8: Retrieval effectiveness when combining the word-based and concept-based representations for retrieval. See Table 4.6 for legend.

(a) 2004 and 2005 queries

	2004				2005			
	MAP		P@10		MAP		P@10	
baseline	0.3576		0.5800		0.2219		0.3551	
MetaMap	0.3504	-2.0%	0.5880	+1.4%	0.2250	+1.4%	0.3551	-0.0%
ATM	0.3536	-1.1%	0.5700	-1.7%	0.2253	+1.5%	0.3592	+1.1%
EAGL	0.3636	+1.7%	0.5940	+2.4%	0.2303	+3.8%	0.3612	+1.7%
CLM	0.3655 [1]	+2.2%	0.5900	+1.7%	0.2256	+1.7%	0.3531	-0.6%
KNN (MeSH)	0.3868 [2]	+8.2%	0.6000	+3.4%	**0.2429** [1]	+9.5%	**0.3755**	+5.7%
MTI	0.3723 [2]	+4.1%	0.5900	+1.7%	0.2306	+3.9%	0.3653	+2.9%
Peregrine	0.3554	-0.6%	**0.6080**	+4.8%	0.2300 [1]	+3.6%	0.3694	+4.0%
KNN (UMLS$_{++}$)	**0.3929** [2]	+9.9%	0.5840	+0.7%	0.2285	+3.0%	0.3592	+1.1%

(b) 2006 and 2007 queries

	2006				2007			
	MAP		P@10		MAP		P@10	
baseline	0.3889		0.4769		0.2796		0.4500	
MetaMap	0.3934	+1.2%	0.4615	-3.2%	0.2725	-2.5%	0.4639	+3.1%
ATM	0.3944	+1.4%	0.4769	-0.0%	0.2456	-12.1%	0.4167	-7.4%
EAGL	0.3986	+2.5%	0.4615	-3.2%	0.2803	+0.3%	0.4556	+1.2%
CLM	0.3982	+2.4%	0.4654	-2.4%	0.2820	+0.9%	0.4694	+4.3%
KNN (MeSH)	0.3736	-3.9%	0.4615	-3.2%	0.2916	+4.3%	0.4750	+5.6%
MTI	0.3960	+1.8%	0.4692	-1.6%	0.2845	+1.8%	0.4750	+5.6%
Peregrine	**0.4111** [1]	+5.7%	**0.4885**	+2.4%	0.2920	+4.5%	**0.4833**	+7.4%
KNN (UMLS$_{++}$)	0.4048	+4.1%	0.4692	-1.6%	**0.2981**	+6.6%	0.4750	+5.6%

Table 4.9: Retrieval effectiveness when combining feedback from different representations. Text feedback is used as a baseline. See Table 4.6 for legend.

(a) 2004 collection

Feedback			2004				2005			
Text	MeSH	UMLS$_{++}$	MAP		P@10		MAP		P@10	
No feedback			0.3576 [2]	-7.1%	0.5800	+2.5%	0.2219 [1]	-7.2%	0.3551	-7.9%
X			0.3851		0.5660		0.2392		0.3857	
	X		0.3868	+0.4%	0.6000	+6.0%	0.2429	+1.5%	0.3755	-2.6%
		X	0.3929	+2.0%	0.5840	+3.2%	0.2285	-4.5%	0.3592	-6.9%
X	X		0.4098	+6.4%	**0.6260**	+10.6%	0.2516 [2]	+5.2%	0.4061	+5.3%
X		X	0.4079	+5.9%	0.6080	+7.4%	0.2529 [2]	+5.8%	0.4020	+4.2%
	X	X	0.4122 [1]	+7.0%	0.6120	+8.1%	0.2495 [1]	+4.3%	0.3959	+2.6%
X	X	X	**0.4144** [1]	+7.6%	0.6200	+9.5%	**0.2559** [3]	+7.0%	**0.4082**	+5.8%

(b) 2006 collection

Feedback			2006				2007			
Text	MeSH	UMLS$_{++}$	MAP		P@10		MAP		P@10	
No feedback			0.3889 [1]	-11.0%	0.4769	-3.9%	0.2796	-5.6%	0.4500	-8.5%
X			**0.4371**		**0.4962**		0.2961		0.4917	
	X		0.3736 [1]	-14.5%	0.4615	-7.0%	0.2916	-1.5%	0.4750	-3.4%
		X	0.4048	-7.4%	0.4692	-5.4%	0.2981	+0.7%	0.4750	-3.4%
X	X		0.4222	-3.4%	0.4692	-5.4%	0.3131	+5.7%	0.5056	+2.8%
X		X	0.4278	-2.1%	0.4962		**0.3187** [1]	+7.6%	**0.5111**	+4.0%
	X	X	0.4111	-6.0%	0.4654	-6.2%	0.3085	+4.2%	0.4833	-1.7%
X	X	X	0.4266	-2.4%	0.4885	-1.6%	0.3167 [1]	+7.0%	0.4944	+0.6%

retrieval. To some extent this can be attributed to the fact that incorrect concepts were detected which deteriorated the results. A second explanation is that the query and document representations did not match up. In the case of MeSH, the documents were manually indexed with MeSH terms. Such a manual process yields different terms than an automatic method. As a result, the query representation might use too few and too specific MeSH concepts whereas relevant documents are indexed with concepts at a different granularity. The KNN approach which obtains a concept-based representation based on retrieval feedback did not suffer from this discrepancy: the representation of the query was similar to the representation of the documents. Moreover, many concepts were used at the same time for searching, further increasing recall.

Obtaining a concept-based query representation through KNN classification (or pseudo-feedback) demonstrated to perform well. This can be explained by the fact that aspects of the query which are not explicitly available in the conceptual representation can be represented by a combination of concepts. The resulting concept-based query might therefore not exactly describe the original information need, but indicate groups of documents in the same area. In these cases, the concept-based representation acts as a recall enhancing device. A second explanation for its performance is the fact that the query remains balanced: as long as the initial text-based query returns a few relevant documents, the derived concept-based representation is likely to keep this balance. In contrast, if only a single concept is detected in the query, the combined query can be skewed to a particular aspect and therefore perform poorly.

The final experiments described in subsection 4.5.4 showed that the concept-based representations can complement a word-based representation. Based on the same feedback documents, retrieval performance became more robust.

4.6 Optimal single term queries

In this section we investigate the added value of the text and concept-based representations by determining how well a single term can be used to retrieve information for a particular information need. We expect that a useful single concept representation for IR is better capable of grouping documents for retrieval than a single word-based representation. Depending on the search goal, recall can be preferred over precision and vice versa: a user interested in finding all relevant information is more willing to accept irrelevant information than a user who wants to find a single piece of relevant information quickly.

We will answer the following research question.

RQ2.6: *How well can single term queries in different representations answer an information need?*

The overview of this section is as follows. First the approach of determining optimal single term queries for a test collection will be described. After that, two examples will be provided to illustrate the approach. The analysis will be described in subsections 4.6.3 and 4.6.4, finalised with a discussion.

4.6.1 Approach

To determine the usefulness of single word or concept representations for IR, we analysed the retrieval performance of optimal single term queries for a set of information needs. The TREC Genomics topic sets were designed to model such needs, and were used in this analysis. For each topic, the precision and recall of a Boolean search using all terms as single term queries was determined. After that, the optimal single query term in each representation was determined for different search goals. To model different search goals, we determined these optimal terms based on the highest F-measure (see Equation 2.9 on page 18), using different values for the parameter of β. A value of β below 1 indicates a preference of precision over recall; values higher than 1 indicate a preference of recall over precision. β was varied between 0.25, indicating a strong preference for precision and 4, indicating a strong preference for recall.

4.6.2 Two examples

To illustrate our approach, Table 4.10 and Table 4.11 show the optimal terms for two TREC Genomics topics. For the topic "Find articles about the function of FancD2" the optimal word terms are 'fancd2', 'fancd', 'fanca', and 'fanconi'. The term 'fancd2' should be used for a single term search which prefers precision over recall ($F_{0.25}$); 'Fanconi' returns a document set with high recall at a lower precision (F_4). Similarly, the MeSH and UMLS$_{++}$ based representations have different optimal terms for the trade-offs between precision and recall. The search terms are what you expect: precise concepts are used for precision-oriented searches [FANCD2 Protein]), more general but related concepts are used for recall-oriented searches [FANC Proteins] and [Fanconi Anemia]). The results for this particular topic are surprising: the MeSH and UMLS$_{++}$ concepts [FANCD2 Protein] and [FANCD2] almost precisely cover the topic. One would therefore expect the highest precision and recall from the concept-based representations. The contrary is true: the text-based representation outperforms the concepts in all except for the recall-oriented searches. For a topic from the full-text collection "What is the role of PrnP in mad cow disease?", similar observations can be made. A single text term achieves higher F-measures for precision-oriented searches. Again only when recall is strongly preferred over precision, the concept-based representation outperforms the text-based representation. It is notable that the same concept [Scrapie] is optimal for both concept-based representations for different search goals. For MeSH, it has relatively high precision and low recall, whereas for UMLS$_{++}$ it gives a higher recall at a lower precision. The example does show a limitation of the experimental method. Without domain knowledge or in fact collection knowledge, the terms 'scheinker'[9], 'prpsc'[10], and 'spongiform'[11] are not obvious search terms to actually use. The same can be said about the optimal query concepts [Scrapie] and [Creutzfeldt-Jakob Disease] which are (Prion) diseases related to mad cow disease found in sheep and humans respectively. The analysis does however show the limitations of the representations of precisely and completely representing information needs.

[9] From Gerstmann-Straussler-Scheinker disease, a related prion disease
[10] The disease-producing protein encoded by PrnP
[11] From bovine spongiform encephalopathy, a synonym of mad cow disease

Table 4.10: Optimal single query terms for the query "Find articles about the function of FancD2" (a gene involved in Fanconi Anemia, a genetic disease); topic 6 from the 2004 topic set.

Measure		Term	P	R	F_β
$F_{0.25}$	Text	fancd2	**0.841**	0.394	**0.788**
	MeSH	[FANCD2 Protein][1]	0.771	0.287	0.702
	UMLS$_{++}$	[FANCD2]	0.597	**0.457**	0.587
$F_{0.5}$	Text	fancd	**0.808**	0.447	**0.695**
	MeSH	[FANCD2 Protein][1]	0.771	0.287	0.577
	UMLS$_{++}$	[FANCD2]	0.597	**0.457**	0.563
F_1	Text	fancd	**0.808**	0.447	**0.575**
	MeSH	[FANCD2 Protein][1]	0.771	0.287	0.419
	UMLS$_{++}$	[FANCD2]	0.597	**0.457**	0.518
F_2	Text	fanca	0.562	**0.532**	**0.538**
	MeSH	[FANC Proteins][2]	0.241	0.436	0.375
	UMLS$_{++}$	[FANCD2]	**0.597**	0.457	0.480
F_4	Text	fanconi	0.093	**0.979**	0.629
	MeSH	[Fanconi Anemia]	0.110	0.904	0.635
	UMLS$_{++}$	[Fanconi's Anemia]	**0.131**	0.968	**0.704**

[1] Officially: Fanconi Anemia Complementation Group D2 Protein

[2] Officially: Fanconi Anemia Complementation Group Proteins

Table 4.11: Optimal single query terms for the query "What is the role of PrnP in mad cow disease?"; topic 160 from the 2006 topic set.

Measure		Term	P	R	F_β
$F_{0.25}$	Text	cheinker	**0.836**	0.232	**0.725**
	MeSH	[Scrapie]	0.711	0.273	0.649
	UMLS$_{++}$	[creutzfeldt-jakob disease]	0.619	**0.616**	0.619
$F_{0.5}$	Text	prpsc	**0.638**	**0.828**	**0.669**
	MeSH	[Prions]	0.614	0.682	0.626
	UMLS$_{++}$	[creutzfeldt-jakob disease]	0.619	0.616	0.619
F_1	Text	prpsc	**0.638**	0.828	**0.721**
	MeSH	[Prions]	0.614	0.682	0.646
	UMLS$_{++}$	[Scrapie]	0.521	**0.929**	0.668
F_2	Text	spongiform	0.550	**0.944**	**0.826**
	MeSH	[Prions]	**0.614**	0.682	0.667
	UMLS$_{++}$	[Scrapie]	0.521	0.929	0.803
F_4	Text	spongiform	0.550	0.944	0.906
	MeSH	[Prions]	**0.614**	0.682	0.677
	UMLS$_{++}$	[Prion Diseases]	0.436	**0.985**	**0.917**

Table 4.12: Average F-measure for optimal single term queries. [1], [2] and [3] in the MeSH and UMLS$_{++}$ columns indicate significant differences to the text-based representation at confidence levels 0.05, 0.01 and 0.001 respectively.

		Word	Word*	MeSH	UMLS$_{++}$
2004	$F_{0.25}$	**0.426**	0.222[3]	0.118[3]	0.312[3]
	$F_{0.5}$	**0.324**	0.221[3]	0.106[3]	0.252
	F_1	**0.304**	0.232[1]	0.113[3]	0.248
	F_2	**0.355**	0.286	0.156[3]	0.293
	F_4	**0.456**	0.393	0.256[3]	0.394
2005	$F_{0.25}$	**0.460**	0.090[3]	0.067[3]	0.187[3]
	$F_{0.5}$	**0.293**	0.095[3]	0.068[3]	0.150[3]
	F_1	**0.229**	0.112[3]	0.086[3]	0.152[3]
	F_2	**0.259**	0.166[2]	0.137[3]	0.208[1]
	F_4	**0.360**	0.267	0.239[3]	0.315
2006	$F_{0.25}$	**0.775**	0.171[3]	0.305[3]	0.532[3]
	$F_{0.5}$	**0.627**	0.180[3]	0.276[3]	0.397[3]
	F_1	**0.555**	0.211[3]	0.264[2]	0.372[3]
	F_2	**0.609**	0.292[2]	0.304[3]	0.442[3]
	F_4	**0.708**	0.409	0.389[3]	0.572[2]
2007	$F_{0.25}$	**0.660**	0.096[3]	0.356[3]	0.495[3]
	$F_{0.5}$	**0.467**	0.100[3]	0.253[3]	0.344[3]
	F_1	**0.386**	0.126[3]	0.222[3]	0.290[3]
	F_2	**0.424**	0.197[3]	0.272[3]	0.341[1]
	F_4	**0.524**	0.320[3]	0.364[3]	0.476

4.6.3 Results

RQ2.6: *How well can single term queries in different representations answer an information need?*

Table 4.12 shows the average optimal F-measures over all topics in the 2004 to 2007 query sets. It shows that on average a single text term can easily outperform concepts in precision and recall. In only rare occasions does a concept outperform a text term[12]. Obviously this can be attributed to some extent to the much larger text vocabulary: there are simply more terms to choose an optimal term from. The results show, however, the potential of using a single word-based representation which cannot be equaled by a single concept-based representation.

The second column in Table 4.12, labelled *Word** puts the results in a different perspective. For this column, the selection of optimal word terms was restricted to words actually occurring in the original topic description. Compared to these values, the optimal concept terms perform reasonably well. A notable exception is the MeSH representation, which on average performed worse than the optimal query words on the 2004 and 2005

[12] in five cases there is not a significant difference between the text-based and concept-based representation

collection. For the 2006 and 2007 collections, the MeSH terms did perform better than the more precision oriented searches. We think that the latter is caused by the difference in collection size in relation to the vocabulary size. Since the 2006 collection consists of only around 160,000 documents, a term from a vocabulary of around 24,000 terms can more precisely select a group of documents than when it is used for selecting documents from a collection of 4.5 million documents. Similarly it is possible to explain why the optimal UMLS$_{++}$ terms performed better at more precise searches in the 2006 and 2007 topic sets.

4.6.4 Analysis of the optimal concept terms

Similar to the optimal word-based representation, one can argue that the optimal concept-based representations are overfitted to the relevant documents rather than being appropriate for the information needs. To investigate what kind of optimal terms were selected, a domain expert was asked to categorise the optimal terms on the following four-point scale.

Exact The term appears exactly in the topic description of the information need. A concept-representation was categorised as exact when one of its synonyms exactly appeared in the topic description. For example, the concept [SLC40A1] was categorised as an exact representation of 'Ferroportin-1' in the query, since they are synonyms.

Derived The term is derived from the topic description. For example, when the topic description mentions 'RSK2' and the optimal word term is 'rsk'. In the case of concepts, this classification was used, for example, when the concept was more general than what was described in the information need. For example, the concept [Iron-Binding Proteins] was categorised as *derived* from 'Ferroportin-1'.

Related The term is not found in the topic description, but is related to the described information need. For example, the concept [GAL80] was categorised as *related* to a topic about the [GAL1] gene.

Unrelated No relationship can be established between the optimal term and the information need. For example, when author names were optimal query terms to find the information.

To reduce the amount of categorisation work, only the best term (for each representation) for a topic was categorised. Table 4.13 shows the categorisation frequencies on all 2004 to 2007 query collections combined, based on the optimal terms for F_2.

Quite a large proportion (27%) of the optimal word terms was classified as unrelated. Further inspection showed that especially on the 2006 and 2007 collection, many of the optimal word terms were in fact rare terms uniquely identifying particular publications, such as author names.

An interesting comparison is when both the optimal word terms and concept terms were categorised as "Exact": both the optimal word term and optimal concept term could be directly related to the original information need. This was the case for 27 topics with word and MeSH terms; for 36 topics this was the case for word and UMLS$_{++}$ terms. We expected that in these cases, the concept-based representation would outperform the text-based representation. The contrary was true however: for the 27 topics, the average F_2 for the MeSH terms was 0.28 and for the word terms 0.45[13]. For the 36 topics, the average F_2

[13] significantly different at the $p < 0.001$ level, based on a paired T-Test

Table 4.13: Classification of the optimal single term queries.

	Exact		Derived		Related		Unrelated	
Word	55	34.2%	14	8.7%	49	30.4%	43	26.7%
MeSH	58	36.0%	32	19.9%	42	26.1%	29	18.0%
UMLS$_{++}$	65	40.4%	22	13.7%	45	28.0%	29	18.0%

measures for word and UMLS$_{++}$ terms were 0.47 and 0.42 respectively[14].

This leads to an important observation: when only considering the optimal word and concept terms which can be exactly related to the information need, on average a single word representation gave a better retrieval performance than a single MeSH concept-based representation. For UMLS$_{++}$ such a difference exists, but the difference is not significant.

This observation further supports the results from the MeSH experiments in the previous sections, where systems such as ATM and MetaMap, which try to precisely map the query text to concepts could not or could only marginally improve a word-based retrieval system. Systems, which map the textual query in a more lenient way to (more) MeSH concepts, such as KNN, CLM, and MTI do show improvements when combined with a text-based system. Similarly, this observation supports why Peregrine, which also tries to precisely map text to UMLS$_{++}$ concepts, does show significant improvements over text-only retrieval when combined with a text-based baseline.

4.6.5 Discussion

In this section we analysed the usefulness of single word or concept representations for IR. The analysis showed that in practice a concept-based representation is quite limited in representing information needs accurately. For a user, the concept-based representation is probably more intuitive to relate to than (stemmed) single words, but on average it is not the most effective representation to search with.

4.7 Predicting concept relatedness

Until now, concept-based representations have been used as a hidden variable to improve information retrieval: the user of the information retrieval is not aware of the fact that a conceptual representation is used to improve retrieval effectiveness. A conceptual representation can also be useful, however, for communicating with the user of an IR system. It could for example be used to explain how the system interpreted the users (textual) query. In this context it can be useful to have a measure which indicates the semantic relatedness of pairs of concepts. This measure can, for instance, be used for expanding a query with related concepts.

In this section, different *concept relatedness* measures will be compared which are either based on ontology structure, or on a document collection in which to each document one or more concepts have been assigned.

[14] not significantly different at the $p < 0.05$ level based on a paired T-Test

Pedersen et al. (2007) defined (concept) relatedness as follows. “Semantic relatedness refers to human judgements of the degree to which a given pair of concepts is related”. Often a distinction is made between semantic relatedness and semantic similarity (Resnik, 1995; Pedersen et al., 2007). Semantic similarity requires concepts to be of the same type, for example [cars] and [bicycles], whereas semantic relatedness defines a more general relationship between concepts based on common characteristics or context, such as [cars] and [gas].

We will answer the following research question in this section.

RQ2.7: *How well can different relatedness measures predict human judgements of relatedness?*

First, an overview will be provided of several relatedness measures. After that, we will explain how the conceptual language models introduced in section 4.2.7, can be used for determining concept relatedness. In subsection 4.7.3 the experimental setup of comparing the measures will be described. In subsection 4.7.4 the results will be described, followed by a discussion and conclusion.

4.7.1 Relatedness measures

In the literature, four categories of concept relatedness measures can be distinguished: based on structure, information content, association, or context. These will now be described.

Measures based on structure

Firstly, concept relatedness can be based on taxonomy structure, or edge counting (Li et al., 2002). The measures assume the concepts to be linked in a graph structure. Nodes in the graph indicate concepts, edges are used to indicate relationships between concepts. Concepts close to each other in the structure are assumed to be strongly related.

The most primitive indicator for relatedness is the shortest path length: the relatedness of two concepts is determined by the length of the shortest path when traversing from one concept to another through the concept taxonomy. Despite its simplicity the shortest path has been frequently used as a relatedness measure (Rada et al., 1989; Hirst and St Onge, 1998; Caviedes and Cimino, 2004).

More sophisticated measures also take into account the depth of the concepts in the graph structure, the lowest common subsuming concept (lcs) or the direction of the relationships in the graph (Wu and Palmer, 1994; Yang and Powers, 2005; Nguyen and Al-Mubaid, 2006). Additionally, these structure based features have been combined in machine learning methods, to train an effective relatedness measure (Li et al., 2002; Liu et al., 2007)

Nguyen and Al-Mubaid (2006) proposed a similarity measure which takes into account the depth of the lowest common subsuming concept and the path length between the two concepts. It is defined as follows.

$$D_{\text{Nguyen}}(c_1, c_2) = \log_2 \big((D_{\text{path}}(c_1, c_2) - 1)\,(m - depth(lcs(c_1, c_2))) + 2 \big) \tag{4.12}$$

Where $D_{\text{path}}(c_1, c_2)$ is the shortest path length between two concepts, m is the maximum depth of the taxonomy, $lcs(c_1, c_2)$ is the lowest common subsuming concept of the two concepts, and $depth(c)$ is the shortest path length from the root of the taxonomy to the concept c.

The method proposed by Wu and Palmer (1994) takes into account almost the same features. It is defined as follows.

$$D_{\text{Wu}}(c_1, c_2) = \frac{2\ depth(lcs(c_1, c_2))}{D_{\text{path}}(c_1, c_2) + 2\ depth(lcs(c_1, c_2))} \tag{4.13}$$

Measures based on information content

Secondly, measures based on information theory have been proposed, often extending measures based on taxonomy structure.

Resnik (1995) proposed a measure taking into account the Information Content (IC) of the concepts. The information content of a concept c is defined as follows.

$$ic(c) = -\log p(c) \tag{4.14}$$

Where $p(c)$ is the probability of encountering a concept c and can in the case of MeSH be based on the ratio of documents assigned to a concept. For instance, when a concept is assigned to a third of the documents in a collection, $p(c) = 1/3$.

Resnik (1995) defined the semantic similarity of a pair of concepts as follows.

$$\begin{aligned} D_{\text{Resnik}}(c_1, c_2) &= \max_{c \in S(c_1, c_2)} ic(c) \\ &= ic(lcs(c_1, c_2)) \end{aligned} \tag{4.15}$$

Where $S(c_1, c_2)$ defines the set of concepts that subsume both c_1 and c_2. So for two concepts which are only connected by traversing the root concept, that is the root concept is the only concept subsuming both, the relatedness is $1 \log(p(c_{root})) = -\log(1) = 0$. When the subsuming concept is more specific, the measure returns larger values. Assuming concepts lower in the taxonomy have a higher information content, D_{resnik} is equal to the information content of the lowest common subsumer, lcs of the concept pair.

Lin (1998b) extended this measure by also taking into account the information content of the individual concepts, as follows.

$$D_{\text{Lin}}(c_1, c_2) = \frac{2 \log p(lcs(c_1, c_2))}{\log p(c_1) + \log p(c_2)} \tag{4.16}$$

There are variations based on IC and edge-counting, using different (machine-learning) strategies to combine features such as depth, path length, and information content, but these are not covered by this work (Jiang and Conrath, 1997; Li et al., 2003).

Measures based on (document) association

Thirdly, different association-based or corpus-based methods can be used to determine the relatedness of concepts. In this case the co-occurrence of concepts (or words) in sentences, paragraphs or documents serves as a relatedness indicator.

Van Rijsbergen (1979) discussed a number of measures for calculating the similarity or diversity of sample sets, such as the Dice coefficient, the Jaccard index, and overlap coefficient. For brevity, only the Dice coefficient is mentioned and used here.

$$D_{\text{Dice}}(c_1, c_2) = \frac{2|D_{c1} \cap D_{c2}|}{|D_{c1}| + |D_{c_2}|} \tag{4.17}$$

Where D_{c1} and D_{c2} are document sets assigned with concept c_1 and concept c_2, respectively.

Alternatively, collocation-based measures such as Pointwise Mutual Information (PMI) and Log Likelihood Ratio (LLR) can be used as well (Manning and Schütze, 1999). PMI is defined as follows.

$$D_{\text{PMI}}(c_1, c_2) = \log \frac{p(c_1, c_2)}{p(c_1)p(c_2)} \tag{4.18}$$

Where $p(c_1)$, $p(c_2)$ are the probabilities of encountering concepts c_1 or c_2 in a large collection, and $p(c_1, c_2)$ is the probability of encountering the assignment of two concepts to a document at the same time.

The Log of the Likelihood Ratio is defined as follows (Manning and Schütze, 1999, p. 173).

$$\begin{aligned} D_{\text{LLR}}(c_1, c_2) = {} & \log L(\mathrm{f}_{12}, \mathrm{f}_1, p) + \log L(\mathrm{f}_2 - \mathrm{f}_{12}, N - \mathrm{f}_1, p) \\ & - \log L(\mathrm{f}_{12}, \mathrm{f}_1, p_1) - \log L(\mathrm{f}_2 - \mathrm{f}_{12}, N - \mathrm{f}_1, p_2) \end{aligned} \tag{4.19}$$

$$\text{where } L(k, n, x) = x^k(1-x)^{n-k}$$

$$\text{and } p = \frac{\mathrm{f}_2}{N} \qquad p_1 = \frac{\mathrm{f}_{12}}{\mathrm{f}_1} \qquad p_2 = \frac{\mathrm{f}_2 - \mathrm{f}_{12}}{N - \mathrm{f}_1}$$

Here, f_1, f_2 are the number of documents assigned with concepts c_1 and c_2 respectively (out of a collection of N documents), and f_{12} is the number of documents to which both c_1 and c_2 have been assigned.

Measures based on context

Finally, the relatedness of concepts has been estimated by considering the context of concepts, where the context of a concept consists of text surrounding or discussing it. Pedersen et al. (2007) presented an approach in which the relatedness of biomedical concepts is defined as the cosine of the angle between two *context vectors*. The context vectors of a concept consist of an aggregation of *word vectors* for the descriptive terms of a concept. The word vectors hold counts of words found in a window surrounding the descriptive terms.

The distance measure is then defined as the cosine between the two context vectors.

$$D_{\text{Pedersen}}(c_1, c_2) = \frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{|\overrightarrow{v_1}||\overrightarrow{v_2}|} \tag{4.20}$$

v_1 and v_2 are the context vectors corresponding to c_1 and c_2 respectively. In earlier work, the cosine was used in a similar fashion for document routing tasks (Buckley et al., 1994).

4.7.2 Relatedness based on conceptual language models

As a concept relatedness measure we propose to use a symmetrical version of the Cross Entropy Reduction (CER) between two concept language models. A concept language model θ_c is defined as a distribution over words based on a concatenation of a subset of documents annotated with a concept c, as described in subsection 4.2.7.

Similar to the measures based on context, the rationale behind our CER-based notion of concept relatedness is that related concepts are surrounded by similar language. The CER quantifies how much better a certain language model is in modelling a certain observed text in comparison with modelling by a collection model. CER has already been successfully applied to ad hoc retrieval and topic detection and tracking (Kraaij, 2004). The CER is defined as follows.

$$\begin{aligned} \text{CER}(\theta_c; M, \theta_{c'}) &= H(\theta_{c'}, M) - H(\theta_{c'}, \theta_c) \\ &= \sum_t P(t|\theta_{c'}) \log \frac{P(t|\theta_c)}{P(t|M)} \end{aligned} \tag{4.21}$$

θ_c is the concept language model of a concept c, M is a background language model and $H(\theta_1, \theta_2)$ is the cross entropy between two language models. Kraaij (2004) argued that the incorporation of $H(\theta_{c'}, M)$ is essential for making the resulting scores comparable, in our case across different concept pairs. A symmetrical version of CER is used as a concept distance. This symmetrical measure is defined as follows.

$$D_{\text{CER}}(c, c') = \frac{\text{CER}(\theta_c; M, \theta_{c'}) + \text{CER}(\theta_{c'}; M, \theta_c)}{2} \tag{4.22}$$

4.7.3 Experimental setup

For comparing relatedness measures for general English words, test sets of 65 word pairs by Rubenstein and Goodenough (1965), and a subset of 30 pairs composed by Miller and Charles (1991), have frequently been used (Resnik, 1995; Lin, 1998b). A common way to assess the quality of a relatedness measure is to compare the scores to the semantic relatedness as indicated by human assessors. The performance is measured by looking at the level of agreement between two gold standard sets and each method, using a correlation coefficient.

For this experiment, two biomedical test sets assembled by Caviedes and Cimino (2004) and Pedersen et al. (2007) were used. They both contain biomedical concept pairs, judged by human assessors on their relatedness. Caviedes and Cimino's test set consists of 55 concept pairs (11 unique concepts), originally intended to measure relatedness measures for the UMLS, judged by three physicians on a 1 to 10 scale. Secondly, a test set from Pedersen was used, consisting of 30 concept pairs, judged by 12 experts on a 1 to 4 scale (Pedersen et al., 2007). This test set was developed for evaluating relatedness measures on the SNOMED-CT ontology. Both test sets were manually mapped to their MeSH term equivalents. Pairs which could not be mapped to MeSH, were removed from the test set. This resulted in a test set of 55 concept pairs (11 unique concepts) and 24 concept pairs (47 unique concepts).

Three variations using Concept Language Models (*CER*, *KL* and *QL*) were compared to two structure-based methods (*path*, *Nguyen* and *Wu*), one information content approach (*Lin*), three association-based approaches (*Dice*, *LLR*, and *PMI*), and one (other) context-based approach (*Pedersen*). For the creation of concept language models, the calculation of association scores, and the calculation of information content, the 2007 MEDLINE baseline distribution was used[15] (consisting of 16,120,074 citations). The tree structure of the 2008 MeSH structure was used for the structure-based methods. For the Pedersen method, we created context vectors based on the same information as for the concept language models. In earlier experiments, it turned out that the suggested windowing method using MeSH terms and MEDLINE gave poor results.

The correlation between the gold standard judgements and relatedness scores was measured using both Kendall's tau rank correlation coefficient and Pearson's correlation coefficient.

4.7.4 Results

RQ2.7: *How well can different relatedness measures predict human judgements of relatedness?*

Table 4.14 shows the correlation between the scores of the different relatedness measures and the ground truth assessments.

The first observation which can be made is that, except for PMI, the performance of all metrics is worse on the second test set. Possibly, the second test set is more difficult than the first. The measures based on information content and structure especially show smaller correlation on the second test set. The methods based on association (in particular PMI) and methods based on conceptual language models performed well on both test sets.

Scatter plots of the system scores against scores obtained from the human evaluators judgements can be found in appendix C.5. We will first discuss the results for test set 1, followed by a discussion of the results on test set 2 (as listed in Table 4.14).

Path length (*Path*) showed a strong linear correlation, but is limited because of its discrete values in separating more and less related concepts. *Nguyen* showed a similar correlation and shows a finer granularity in its scores. *Wu* assigned similar scores to pairs which are less related (judgements > 6, at a scale from 1 to 10). The association based measures (*Dice*, *PMI* and *LLR*) performed well according to their correlation coefficients in Table 4.14, but only PMI shows a clearly recognisable correlation between scores and judgements. This is caused by some of the outlying scores returned by Dice and LLR for strongly related pairs of concepts. Similar to *Path* length, the scores returned by *Resnik* are limited to a few unique values, caused by a concept which is frequently used as a lowest common subsuming concept. This is caused by the limited number of unique concepts in the test collection. *Lin* showed more variation in scores, but has a slightly lower correlation than Resnik. Both Resnik and Lin quickly reached a minimum after a particular judgement level (around 6). Depending on the application of the relatedness measure one could argue that such behaviour is not undesirable. For applications which are mostly interested in strongly related concepts, the Lin metric might be sufficient. *Pedersen* showed a linear correlation, but with quite some noise at all judgement levels. The *CER*, *KL* and *QL* measures all perform

[15] http://www.nlm.nih.gov/archive/20090811/bsd/licensee/2007_stats/baseline_doc.html

Table 4.14: Absolute correlation between metrics and ground truth (τ = Kendall tau rank correlation coefficient, ρ = Pearson's correlation coefficient). ^{1} to ^{3} indicate whether the correlations are significant at the p-levels, 0.05, 0.01 and 0.001 levels respectively.

		Test set 1		Test set 2	
		τ	ρ	τ	ρ
Structure	path	0.598^{3}	0.719^{3}	0.271^{1}	0.397^{1}
	Nguyen	0.609^{3}	0.741^{3}	0.249	0.357^{1}
	Wu	0.681^{3}	0.804^{3}	0.241	0.319
Association	Dice	0.678^{3}	0.833^{3}	0.493^{2}	0.639^{3}
	PMI	0.510^{3}	0.687^{3}	**0.654**^{3}	**0.782**^{3}
	LLR	0.672^{3}	0.833^{3}	0.440^{2}	0.561^{2}
Information content	Resnik	0.719^{3}	0.833^{3}	0.282^{1}	0.366^{1}
	Lin	0.666^{3}	0.822^{3}	0.246	0.337
Context	Pedersen	0.634^{3}	0.797^{3}	0.387^{2}	0.535^{2}
CLM	CER	**0.767**^{3}	0.914^{3}	0.522^{3}	0.662^{3}
	KL	0.717^{3}	0.866^{3}	0.544^{3}	0.693^{3}
	QL	0.760^{3}	**0.918**^{3}	0.484^{2}	0.646^{3}

quite well. The *KL* method however, does show a smaller linear correlation. Topic pairs which are less related are not separated as well as by *CER* and *QL* scores. Comparing the correlation plots of *CER* and *KL* to each other, one would argue that the normalisation factor present in *CER* indeed is valuable. However, when comparing the correlation coefficients of *CER* to *QL* the normalisation is not really justified.

The scatter plots for the second test set illustrate that the test set is quite small and since there are quite a few ties in the judgements scores (a group of five, a group of four, and three groups of 2 pairs have the same score), evaluation based on the test set did not clearly indicate which relatedness measure is better.

4.7.5 Discussion and conclusion

In the previous experiments we have investigated the effectiveness of various relatedness measures in the biomedical domain. The results indicated that the comparison of concept language models (the *CER*, *KL* and *QL* methods) is quite effective for predicting the relatedness of concepts as indicated by human annotators. Structure-based measures did not perform as well as the other measures based on document association (co-occurrence of concepts in documents) and information content. We point out that the association and information content measures are strongly related to each other: the information content is based on the same document collection as the association measures. They are both based on the document frequencies of MeSH terms in documents. Such a relationship can also be observed between concept language models and association based measures. The concept language models of two concepts are more similar when they are based on the same documents. The association-based measures takes the number of these overlapping documents into account. The concept language models take more evidence into account by

using the term frequencies of these (partially overlapping) documents. A drawback of the approach based on concept language models is its complexity: to compare a single pair of concepts, many term probabilities have to be compared. This can become an issue when many concept pairs have to be efficiently compared. A brute force solution to this problem would be to simply calculate all the concept relationships offline and store them. For a small vocabulary such as MeSH, this is quite feasible, but such a solution does not scale up to larger concept vocabularies. Alternatively, a possible solution would be to combine for example an association-based measure with the CLM measure. The association-based measure can be used to quickly select a smaller group of concepts, followed by a comparison of concept language models to determine the relatedness more precisely.

A limitation of this work is the size of the used test sets. It is difficult to say whether results on an experiment with 55 and 24 concepts generalises to pairs from the complete set of MeSH terms. This is however common practice: the early test sets by Rubenstein and Goodenough (1965) (65 pairs), and the subset of 30 pairs composed by Miller and Charles (1991), are still frequently used as test sets in the general English domain.

A more important issue is the actual usefulness of a relatedness measure for IR. "Relatedness" describes a quite general relationship between terms. In our own experience, we found it difficult to compare different specific types of relationships on the same relatedness scale. For example, how does the relatedness of a hypernym and hyponym (for example, [Heart diseases] and [Heart neoplasms]) compare to the relatedness of a concept which is a meronym of the other (for example, [Heart] and [Cardiovascular system])? The MeSH hierarchy contains both these relationships as parent and child nodes. And how does this compare to siblings in a type-of relationship (for example, [Heart] and [Blood vessels])? For IR, expanding with these related terms can give different results. This problem is actually supported by statements in the paper describing the second test set (Pedersen et al., 2007). At the beginning of their paper, Pedersen et al. note that "Studies have shown that, surprisingly, most humans agree on the relative semantic relatedness of most pairs of concepts". In describing their experimental setup, however, they note that "The correlation on the medical test set of 120 concept pairs was 0.51. To derive a more reliable test set we extracted only those pairs whose agreement was high". One could argue that this extraction resulted in an "easy" test set of 30 terms, but this might also be caused by the aforementioned problem of putting different types of relationships on the same scale. This does not make the discussed and evaluated relatedness measures less useful: they can still be useful to compare pairs of concepts with the same type of relationship. Hence, it can be a motivation for future work on developing relatedness measures which also automatically determine the types of concept relationships in the context of IR.

4.8 Chapter summary

In this chapter, a concept-based representation for biomedical IR was introduced and investigated. Theoretically, a concept-based representation has the added value of being capable of representing information in a normalised, unambiguous fashion. For information retrieval such a representation would overcome vocabulary mismatch between query and documents.

Two concept-based representation vocabularies were investigated in this chapter: MeSH, a controlled vocabulary for indexing biomedical documents, and UMLS$_{++}$, the UMLS

Metathesaurus extended with a number gene and protein dictionaries.

We compared different classification systems to automatically obtain concept-based document and query representations. We proposed two classification methods based on statistical language models, one based on K Nearest Neighbours (*KNN*) and one based and concept language models. KNN classifies text based on similar, pre-classified documents. The method based on concept language models classifies text by ranking language models which have been built for each concept. The systems were compared to a number of out-of-the-box classification systems.

In a document classification experiment, we investigated to what extent a number of classification systems could reproduce manually created concept-based document representations. The proposed KNN system performed surprisingly well in comparison to the out-of-the-box systems. Further analysis indicated that the automatic classification systems returned additional concepts which were useful for representing the documents.

In a query classification experiment, we investigated the usefulness of having a concept-based representation for retrieval. The investigated classification systems showed strongly varying performance in effectively mapping a text-based query to a concept-based representation for retrieval. Retrieval based on only concepts was demonstrated to be less effective than word-based retrieval. However, depending on the classification method used, significant improvements in retrieval effectiveness could be observed when the concept-based representation was combined with a word-based representation. Again, the proposed KNN classifier, performed well in comparison to the other investigated systems.

In an artificial setting, we compared the optimal retrieval performance which could be obtained with word-based and concept-based representations. In contrast to our intuition, a single word-based query showed to perform better on average than a single concept-based representation, even when the best concept term precisely represented part of the information need.

In general, we conclude that in practice a concept-based representation is very limited in comparison to a word-based representation. On its own, it cannot completely and precisely represent information needs. However, when combined with a text-based representation it can bring significant improvements to retrieval. Obtaining a concept-based representation through pseudo-relevance feedback (KNN) turned out to be especially effective.

In a final experiment, we investigated to what extent the relatedness between pairs of concepts as indicated by human judgements could be automatically reproduced. Results on a small test set indicated that a method based on comparing concept language models performed particularly well in comparison to systems based on taxonomy structure, information content and (document) association. It was noted, however, that additional work is necessary to make the relatedness measures useful for biomedical IR.

Chapter 5

A Cross-Lingual Framework for Biomedical IR

"The original is unfaithful to the translation."

Jorge Luis Borges

The original idea for this chapter has been published in Trieschnigg (2008); parts of this work have been published in Schuemie, Trieschnigg, and Kraaij (2007b) and Trieschnigg, Hiemstra, de Jong, and Kraaij (2010).

In the previous chapter we investigated the possible gains of using a concept-based representation for biomedical IR. We demonstrated that a concept-based representation on its own cannot outperform a text-based retrieval system, but that a careful combination of the two representations can improve retrieval effectiveness.

We will investigate a tighter integration of the two representation types from a "cross-lingual" perspective. In traditional cross-lingual information retrieval (CLIR), queries and documents are expressed in different languages. The retrieval system has to perform some kind of (automatic) translation before the two can be matched. Similarly, coping with the mismatch of terminology in IR can be viewed as a form of cross-lingual IR: translation is required to allow for matching of different terms for the same concept. We propose to view the integration of a concept-based representation in biomedical IR as a cross-lingual retrieval problem. The elements that need to be translated into each other are the concept-based representations and textual representations. The integration of a concept-based representation in biomedical IR is then reduced to translating the query and/or documents, and matching them in the same representation type. Such a cross-lingual perspective gives the opportunity to adopt a large set of established CLIR methods and techniques for this domain. In this chapter, we will answer RQ3 posed in chapter 1.

RQ3: *Is it possible to cast the integration of knowledge from terminological resources in biomedical IR into a retrieval framework?*

The structure of this chapter is as follows. First, a short background will be provided about traditional CLIR. After that, in section 5.2 we will introduce a cross-lingual framework

for biomedical IR and outline the differences with traditional CLIR. In section 5.3 we will describe a number of translation models for biomedical CLIR. The first two retrieval models that will be described in section 5.4 use these translation models directly for retrieval. The second set of three retrieval models will combine translation models to improve translation. An experimental evaluation of these translation and retrieval models will be described in sections 5.5 and 5.6, followed by a discussion and summary of this chapter.

5.1 Established cross-language IR

Traditional cross-language IR is concerned with retrieving documents in a language different from the user's query language. For example, a user can formulate his information need in Dutch and the retrieval systems retrieves English documents. A cross-lingual retrieval system relieves the user of the burden of formulating the query in the document language. Such an approach has two important advantages. Firstly, despite the fact that the user might be able to read and understand the documents in the foreign language, the user is likely to be more comfortable with formulating a query in his own (or most fluent) language. Secondly, in a case where documents are available in multiple languages, the user only has to formulate the query once.

5.1.1 Approaches to CLIR

In general, three approaches to CLIR translation can be distinguished: query translation, document translation or a combination of both (Kraaij, 2004; Wang and Oard, 2006). In a system based on query translation, only the query is translated into the document language. Matching takes place in the language of the documents. One drawback of this approach is the limited context provided by a short query: only limited information is available to disambiguate query terms and to select the proper translation. Systems based on document translation, translate the documents to the query language. In this case, matching takes place in the query language. Documents provide considerably more context to perform a more accurate translation, possibly resulting in better cross-lingual matching. However, translating and storing the documents in multiple languages can be prohibitively expensive. Finally, both the documents and the queries can be translated before matching. By performing the matching in both document and query language, and combining the results, retrieval can become more robust (McCarley, 1999; Wang and Oard, 2006).

The translation may not be limited to the query (source) and document (target) languages. During transitive translation, an intermediate or pivot language is used for translating the source language expression to the target language. For instance, Dutch queries can be first translated into English before translating them into Arabic. This can for example be useful when no translation resources are available for particular language pairs, or when the resources to translate to an intermediate language are of a higher quality than the resources for direct translation.

5.1.2 Translation resources

Translation for CLIR can be based on different language resources. We distinguish between four types of language resources (Moreau, 2009).

Firstly, a machine translation (MT) system, such as Babel Fish or Google Translate can be used to translate a representation. MT systems translate a text to a single, most likely, translation. A drawback of such systems is that they offer limited control over the output: only a single translation can be obtained. Uncommon (senses of) words are therefore likely to be incorrectly translated. Advantages are that MT systems are readily accessible and provide good accuracy at general translation tasks.

Secondly, machine-readable dictionaries (bilingual lexicons) and multilingual thesauri (also referred to as ontologies) have frequently been used for translating representations (Hull and Grefenstette, 1996; Pirkola et al., 2001). The main difference between dictionaries and thesauri is their structure: dictionaries contain definitions of language expressions, whereas thesauri group language expressions according to similarity of meaning and can organise entries hierarchically by themes and topics. In either case, translations can be obtained by looking up terms and translating them one by one. A drawback of these resources is that they do not provide any information about the actual use of the stored translations: they do not provide translation probabilities between terms. Moreover, they provide little support for translating phrases and particular expressions.

Thirdly, various corpus-based approaches have been proposed for obtaining translation models. A *parallel corpus*, a collection of translated documents in multiple languages, can be used to learn translations between terms. The translation probabilities are then based on the alignment of documents at the sentence and the word level (Brown et al., 1993; Oard, 1997). Alternatively, a *comparable corpus* can be used to train translation models. In this case no (explicit) alignment is available between the documents in different languages. For example, news articles in different languages discussing the same event might not be exact translations, but still allow translations to be learnt to some extent. Given the large amounts of textual information available in multiple languages on the Web, these approaches are promising since they can provide good coverage of frequently used translations. Moreover, they provide translation probabilities which can be used for IR, which will be discussed in the next section.

Fourthly, conceptual interlingua have been used as a language resource for CLIR. An interlingua is a "knowledge base of language-independent concept representations" (Ruiz et al., 1999). Each concept is linked to its respective translations in various languages. Ruiz et al. (1999) for instance, built an interlingua by extending WordNet (a combination of dictionary and thesaurus in English) with multiple languages. Despite their theoretical attractiveness, interlingua are not commonly used for traditional CLIR (Kraaij, 2004).

5.1.3 CLIR models

As said, the integration of translation in CLIR can be carried out in different ways. A straightforward way is to use a machine translation system to translate the query into the document language. This translated query can then be used in a monolingual retrieval system. A drawback of such an approach is that errors in the translation process are severely penalised: if the incorrect sense and thus incorrect translation is chosen for a term, retrieval performance is likely to be hurt.

Alternatively, multiple translations can be used in the IR process. It is important to keep in mind that for cross-lingual matching, translation does not need to be completely accurate or understandable to the user. Inaccurate translation can have a beneficial query

expansion effect: a less precise translation with a related term (possibly caused by errors in the translation model) can still retrieve relevant documents. Structured queries can be used to group the translations of a single term. The query collocation effect can push up documents with the intended translation: for instance, when all the translations of 'bank' are combined with all the translations of 'money', it is likely that documents with the financial sense of 'bank' are retrieved first, because they also contain a translation of 'money'.

Pirkola (1998) and Kwok (2000) investigated the use of a structured query language including operators for synonyms, proximity and a combination of both. In their approach each translation is assumed to be equally likely. The translation alternatives can also be weighted to represent the largest confidence in that particular translation, or, weighted according to their actual use (Darwish and Oard, 2003; Gao et al., 2006). Alternatively, weighted word-by-word translation can be effectively directly embedded into the retrieval model (Kraaij, 2004).

It is also possible to integrate translations in IR without an explicit dictionary or translation model. By searching a parallel corpus in the source language, a translated query can be obtained from the top-ranked documents in the target language (Ballesteros and Croft, 1997; Lavrenko et al., 2002).

Finally, it can be decided *not* to translate the query. For languages which are quite similar this can in fact be quite effective. Also when proper names are used, and in many cases the query should remain unaltered, such an approach can be beneficial.

5.1.4 CLIR challenges

Kraaij (2004); Gao et al. (2006) and Oard (1997) identified three important challenges for traditional CLIR.

Firstly, coverage of the translation resources. Dictionary-based approaches to traditional CLIR especially suffer from incomplete coverage of the language pairs. Secondly, there is the problem of translation selection or lexical disambiguation. The problem is more severe than for monolingual IR, since two languages are involved. Finally, translating multi-word terms and phrases has been shown to be important for effective CLIR. Detecting and translating phrases rather than single terms can considerably improve translation. However, a more extensive dictionary, or a larger parallel corpus for training statistical translation models is required to obtain high quality translations.

In the SIGIR forum of June 2003, a group of senior researchers discussed long-term challenges for IR. In their report (Allan et al., 2003), they note about the challenges of CLIR: "the lessons learned from CLIR suggest new ways to approach monolingual retrieval. Given the successes to date in CLIR, any improvements in monolingual retrieval should generate comparable improvements across languages and vice versa". This is a clear motivation for the biomedical CLIR framework we will now describe.

5.2 A Biomedical CLIR framework

Figure 5.1 gives an overview of the cross-lingual framework for biomedical IR which will be discussed in this section.

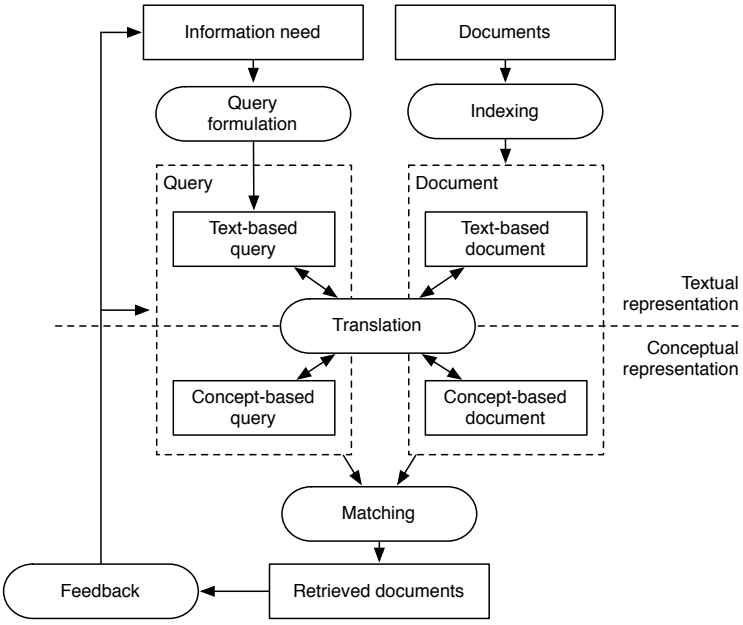


Figure 5.1: A cross-lingual framework for biomedical IR.

As discussed in the background chapter, biomedical IR suffers from terminology or vocabulary mismatch: the terminology used in queries does not, or does not fully agree with the terminology used in relevant documents. To put it simply, the query and document languages can be considered to be different languages and some kind of translation step is required to match them. In monolingual retrieval this kind of translation has been suggested by Berger and Lafferty (1999), who proposed viewing the query formulation process as a noisy translation from the language used in relevant documents. From such a viewpoint, expanding or updating a representation within one language can also be viewed as a translation process.

In our framework, a second concept-based language is introduced. The concept language is expected to overcome (some of) the limitations of word-based matching discussed in the previous chapters. Moreover, the second language offers the possibility of using additional data for enrichment of the original query and document representations. For example, by leveraging the relationships between concepts based on a taxonomy structure. Or, by using an additional collection of documents in a concept-based representation to determine concept relatedness, as explored in section 4.7.

Incorporation of concept-based representations can be modelled as a cross-lingual matching problem: matching queries and documents can be viewed as translating and matching word and concept representations, either within a language or across the two languages.

The remainder of this section is structured as follows. First, the languages and resources to translate between languages will be described. Then, the possible translations between representations will be discussed. Finally, a comparison will be made between traditional CLIR and the biomedical CLIR investigated in this chapter.

5.2.1 Languages and translation resources

Two languages are distinguished in biomedical CLIR. Firstly, the textual language in which queries are formulated by a user in free text and in which documents have been written. Secondly, a conceptual language which is defined by the concepts or entries in a terminological resource. For example, the manual indexing of MEDLINE documents with MeSH forms the document representation in the conceptual language.

A number of *resources* are available to link these languages to each other. Firstly, a corpus of documents in the two languages is available. MEDLINE citations both provide a textual representation and a MeSH-based representation. MEDLINE can be considered as a corpus of comparable documents: both languages represent the same information in their own way. Secondly, terminological resources, such as thesauri, domain-specific databases, and controlled vocabularies, which group synonyms into concepts can be used to link the representations. In this case each concept is linked to a number of synonymous terms.

Also within a single language resources are available to link representations. Firstly, a collection of documents can be used to infer useful relationships between representations. For instance, relationships can be inferred from the co-occurrence of text or concepts in documents to build similarity thesauri and statistical thesauri (Qiu and Frei, 1993). Additionally, the relationships between concepts explicitly defined by the taxonomy provide information to link within the concept event space.

5.2.2 Translating and expanding representations

Strictly speaking, a translation of a representation in a source language should result in a semantically equivalent representation in the language being translated to. In some cases such an equivalent representation is not available: for some expressions there may simply be no precise translations in the target language. One can simply decide not to translate these expressions, or to translate to a strongly related representation. The latter approach can be even beneficial to retrieval effectiveness: the translation may result in a beneficial query expansion effect.

Assuming that expanding or updating a representation in one language is also a form of translation, four types of translations can be distinguished, either applied to the query or to the document representation.

- Text to text translation;
- Text to concept translation;
- Concept to text translation;
- Concept to concept translation.

As explained in the previous sections, ambiguity is an important hurdle for correctly translating between representations. The amount of *context* used in the translation can therefore strongly affect translation quality. For instance, translating a query in a word-by-word fashion does not take into account the context of the query words. By translating phrases or multiple words from the query at the same time, more context is taken into account, allowing for more accurate translations.

Table 5.1: Translations with different contexts and resources.

From text to text	From text to concepts
• Similarity thesaurus • Thesaurus/dictionary • Local query/document context • Text corpus (relevance model)	• Word-to-concept • Thesaurus lookup • Parallel/comparable corpus (relevance model)
From concepts to text	**From concepts to concepts**
• Concept-to-word • Concept-to-phrase • Parallel/comparable corpus (relevance model)	• Using a relatedness measure • Using taxonomy structure • Concept corpus (relevance model)

A second important factor which influences the quality of the translation is obviously the quality of the translation models. Incorrect translations likely result in poor retrieval performance. A third factor influencing the translation quality is the extent the translation models offer the possibility of using information from the context during translation. An automatically obtained translation model may contain more errors than a hand-crafted dictionary. The automatically obtained model can, however, be built specifically to leverage context provided for the translation.

In the following sections, different approaches to translation between representations will be discussed with a focus on the available contexts and the resource(s) used.

Figure 5.1 schematically shows the translation types and lists a number of approaches to carrying out the translation using different amounts of contextual information.

"Translating" from text to text

There are quite a few ways available to translate a textual representation to another textual representation. Strictly speaking a translation should only find synonyms or near-synonyms of a textual representation. However, allowing for translation into more general, related terms can lead to a beneficial query expansion effect, which is beneficial for recall.

A naive way of carrying out a query translation is to carry out a word-by-word translation using a dictionary or a statistical thesaurus based on a global analysis of the corpus (see section 2.3.2). Such an approach totally ignores the query context in which the words occur. As a result, such a translation may suffer from lexical ambiguity. An advantage of using a statistical thesaurus over a dictionary is that the first approach can be better tuned to the collection being searched. Moreover, a statistical thesaurus provides weights to indicate the relative importance of terms.

More sophisticated query expansion or translation methods take into account a larger context. Bai et al. (2007) and Bai and Nie (2008), for example, derived the expansion terms from pairs of query terms rather than from individual terms. These context-dependent translations performed considerably better than a word-by-word approach.

Re-estimating a textual query representation based on pseudo-relevance feedback (or relevance models) can also be viewed as a form of translation which takes into account the context of a query as a whole. If the original query is of high quality, the pseudo-relevant documents relate to the query as a whole and as a result expansion terms take into account

the query context well. However, if the retrieved documents are skewed towards a particular aspect of the query, the results of relevance feedback are likely to result in query drift.

Translating from text to concepts

Analogous to the translation within the textual event space, translation from a textual to a conceptual representation shows varying degrees of taking context into account during translation.

Translation can, for example, be based on word-by-word translation: each word can be individually translated to the most appropriate concept(s) found in a thesaurus. Or, a larger context can be taken into account by matching text phrases to synonyms of concepts found in a dictionary. The KNN approach investigated in the previous chapter, takes the complete query context into account. The translation of a query is based on concepts co-occurring with all query words.

Similar to traditional CLIR, document translation can benefit from the context provided by the document. The disambiguation of terms can for instance be based on the presence of unambiguous synonyms found in the same document.

Translating from concepts to text

Context is less important for the translation of a concept-based representation to text, since the representation is (supposed to be) unambiguous. Translating concepts to text independently from the context they appear in can work quite well. The translation can, for example, be based on a translation model trained on a comparable corpus of text and concept-based documents. Or concepts can be translated to the synonymous phrases found in a thesaurus.

Translating from concepts to concepts

The translation between concept-based representations can be based on the hierarchal structure in which the concepts have been organised. Translation to these concepts can lead to a beneficial query expansion effect. For example, concepts can be translated to parent, child or sibling concepts. Alternatively, the relatedness measures discussed in section 4.7 can be used for finding concepts to translate to.

5.2.3 Comparison to established CLIR and research questions

In the previous sections we explained how biomedical IR can be viewed as a cross-lingual retrieval problem. In this chapter we investigate how we can adopt CLIR methods and techniques for more effective monolingual biomedical information retrieval. The challenges of integrating these techniques lie in the differences between traditional and biomedical CLIR.

The main difference between biomedical and traditional CLIR is obviously that for biomedical CLIR, translation is not strictly required: reasonable results can be achieved without translation. In traditional CLIR, translation is essential to allow for matching; in biomedical CLIR, translation is expected to enhance the matching process. To be more precise: the translation is expected to result in a recall enhancing effect because of the

integration of synonyms. Additionally, it can enhance precision by translating multi-word terms to a single concept.

The first challenge of using CLIR methods and techniques for biomedical CLIR is how to build effective translation models. Compared to conventional CLIR, the languages are considerably different. Firstly, concepts organise information at a different granularity than a word-based representation: a concept groups a number of synonymous phrases and textual expressions. These concepts vary from fine-grained, such as [Immunoglobulin Enhancer-Binding Protein] to general, such as [Genes]. The words in a word-based vocabulary also group information at different granularity levels, but the differences are not expected to be as large in a concept-based vocabulary. Secondly, concepts in a concept-based vocabulary have a different relationship to each other than words in a word-based vocabulary. Concepts in a concept-based vocabulary are intended to discriminate between concepts we encounter in the world. The words in an (automatically obtained) word-based vocabulary are not organised in such a way: there is a considerable overlap in what they refer to. It is not known how the approaches available to build translation models in CLIR handle these differences in vocabularies. Moreover, it is unknown how these translation models handle different concept vocabularies. Formulated as two research questions.

RQ3.1: *How can we build translation models for biomedical CLIR?*

RQ3.2: *How effective are these translation models for improving word-based retrieval?*

Both traditional and biomedical CLIR suffer from lexical ambiguity. In the case of traditional CLIR, this ambiguity is present in both the source and target language. For biomedical CLIR, the concept-based representation is (supposed to be) unambiguous, or at least less ambiguous than the text-based representation. In both cases, the extent to which the translation can deal with this ambiguity depends on the amount of (query or document) context taken into account. By translating queries as a whole, for instance, more query context is taken into account than when the query words are translated individually. A second way to improve translation quality is to combine translation resources. When two translation models agree on a translation, the translation is more likely to be correct.

Formulated as two research questions.

RQ3.3: *How does context affect translation quality for biomedical CLIR?*

RQ3.4: *Can translation for biomedical CLIR be improved by combining translation models?*

In subsection 5.1.4, we noted that one of the challenges for traditional CLIR is the coverage of the language pairs by the translation resource. For biomedical CLIR, coverage of a concept-based representation poses a similar challenge: simply not all "concepts" expressed in text have to be present or can be represented in a concept-based vocabulary. As a result, particular aspects of the original query cannot be precisely represented in terms of concepts. In the worst case, important aspects of the original textual query are neglected or even ignored in the translation, resulting in query drift towards aspects which can be accurately translated to concepts. We hypothesise that by combining translation models this drift can be prevented.

Formulated as a research question.

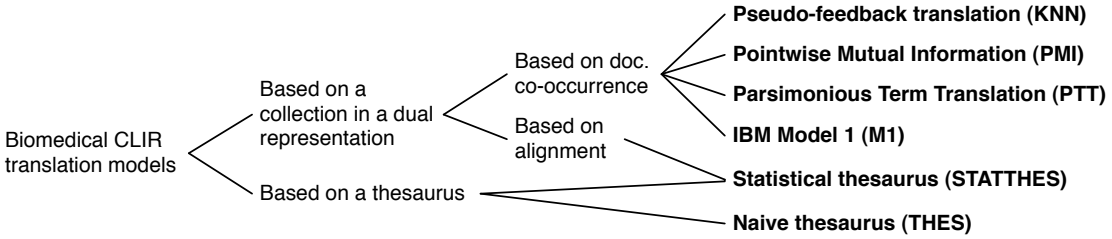


Figure 5.2: Taxonomy of translation models for biomedical CLIR investigated in this chapter.

RQ3.5: *Can translation models be used to prevent query drift?*

In the next section, we will propose a number of translation models for biomedical CLIR. In section 5.4, we will propose a number of retrieval models using these translation models which will be used to investigate these research questions.

5.3 Translation models for biomedical CLIR

Two types of translation models for biomedical CLIR will be investigated in this chapter (illustrated in Figure 5.2). The first group of translation models is based on a corpus of documents available in both text and concept-based representations. This group can be further divided into a group which is based on the co-occurrence of words and concepts in documents (a comparable corpus) and a group based on the alignment between concepts and words (a parallel corpus). In appendix D.4, an example of a comparable document can be found.

The second group of translation models is based on the thesaurus in which the concepts have been described. The translation model based on a statistical thesaurus (STATTHES) combines information from a comparable corpus and a thesaurus.

The first translation model we will investigate, based on pseudo-feedback translation (KNN), translates a text-based representation as a whole to a concept-based representation. Its translation is based on concepts co-occurring frequently in a comparable corpus with text-based representation. The model was described in subsection 4.2.8.

The other five translation models we will investigate (M1, PMI, PTT, STATTHES, and THES) translate words to concept representations in a term-by-term fashion. They employ different methods to estimate probabilities for $P(w|c)$ (the probability of translating a concept c to the word w) and $P(c|w)$ (the probability of translating the word w to a concept c). It is expected that on their own these term-by-term translation models will not be beneficial for biomedical IR. A single word is expected to be too ambiguous and to provide too little information for accurate translation to a concept-based representation. Consequently, we expect matching to deteriorate in comparison to simple word-based retrieval. We hypothesise, however, that these simple translation models can be useful in combination with pseudo-feedback translation. Firstly, they can be used to clean up noisy translations. A term-by-term translation model can be used to determine which concepts do not have a relationship with the original query and can therefore be removed.

Secondly, we expect that the term-by-term translation models are useful for creating a structured query which combines the word-based query with the concepts found through

pseudo-feedback translation. By grouping words with tightly associated concepts, query drift is expected to decrease.

These translation models will now be described in more detail.

5.3.1 Pseudo-feedback translation (KNN)

The first translation model that we will investigate was described in subsection 4.2.8. In this translation model, the concept-based translation is based on the joint probability of observing a concept c together with the text to classify (Q). This joint probability is approximated is follows.

$$P(c, Q) = \sum_{D \in \mathcal{D}} P(D)P(c|\phi_D)P(Q|\theta_D) \tag{5.1}$$

Where $\mathcal{D}$ is a document collection, $P(D)$ is the probability of sampling a document D from this collection, $P(c|\phi_D)$ is the concept-based language model of a document D and $P(Q|\theta_D)$ is the probability that the query Q is sampled from the word-based language model of document D.

When this joint probability is estimated on the searched collection itself, the approach can be viewed as a form of pseudo-feedback: the concept-based representation is based on documents deemed most relevant to the text to classify. It can also be viewed as an instance of multi-label K-Nearest-Neighbour (KNN) classification: the classification (translation) is based on classes assigned to neighbouring documents. In the remainder of this chapter, we will use 'pseudo-feedback translation' and 'KNN' interchangeably to refer to this translation model.

Major advantages of this approach are its simplicity and the fact that no intermediate representations have to be trained. In the previous chapter it was shown that pseudo-feedback translation can be effectively combined with word-based retrieval. One disadvantage is that, since it is based on pseudo-feedback, the model is dependent on the original query: if initial retrieval performance is poor, the obtained feedback representation will reflect this. A second disadvantage is that the representation can be noisy. Since the concept representation is based on feedback documents, it is likely to contain concepts which are not directly related to the original query. These concepts can result in a beneficial query expansion effect, but also skew the obtained representation into the wrong direction.

Figure 5.3 shows an example of translating a query to a weighted concept-based representation.

5.3.2 IBM Model 1 (M1)

The second translation model we will investigate is based on IBM Model 1, a statistical model of the translation process commonly used for traditional CLIR. Brown et al. (1993) proposed five models for determining statistical translation models based on a bilingual collection of sentences. Central to these models is the estimation of an *alignment* of the sentences in two languages. This alignment connects terms in the sentences in one language to terms in the translated sentence in the other language. An EM-algorithm is employed to iteratively improve the alignment and the parameters of the translation model, respectively. After a uniform initialisation of the translation probabilities, the probability of each possible

"Ferroportin-1 in humans" 0.095 [Humans], 0.091 [Cation Transport Proteins], 0.079 [Iron], 0.078 [Animals], 0.050 [Membrane Proteins], 0.038 [Enterocytes], 0.038 [Hemochromatosis], 0.036 [Carrier Proteins], 0.034 [Male], 0.021 [Iron-Binding Proteins], 0.021 [Mice], 0.018 [Biological Transport, Active], 0.018 [Cloning, Molecular], 0.018 [Zebrafish], 0.018 [Ferric Compounds], 0.017 [Duodenum], 0.017 [Models, Biological], 0.016 [Middle Aged], 0.016 [Forecasting], 0.016 [Intestines], 0.015 [Brain], 0.015 [Blotting, Western], 0.015 [Rats, Sprague-Dawley], 0.015 [Aging], 0.015 [Animals, Newborn], 0.015 [Rats], 0.012 [Iron Overload], ...

Figure 5.3: Translation of the textual query "Ferroportin-1 in humans" using feed-back translation (KNN). In section 5.5 is explained how the translation model was trained.

alignment between two sentences is determined. This probability is used for updating the translation model: the translation probabilities between terms which occur in more likely alignments is increased. This process can be repeated until the translation probabilities do not change anymore.

IBM Model 1 is the simplest of the five models proposed by Brown et al. (1993), which does not take word order into account. Models 2 to 5 are increasingly sophisticated, incorporating absolute and relative word reordering and a fertility model. For biomedical CLIR, the concept-based representation does not have a term order. Since we limited our experiments to term-by-term translation models, we will only use Model 1 for our translation models from text to concepts and vice versa. We do note, however, that it is possible to train more sophisticated phrase-based translation models using this approach which we will suggest for future work (see section 5.7).

An advantage of using Model 1 for training biomedical translation models is its theoretical soundness. The subsequent models proposed by Brown et al. (1993) illustrate that Model 1 is highly suitable to be extended to more sophisticated models. Disadvantages are that training the translation model is resource intensive and that with new concepts the whole training process has to be repeated.

Figure 5.4(a) and Figure 5.5(a) show two examples of translation probabilities for translating from a MeSH concept to words, and from a word to concepts, respectively. The training set used to build these translation models is described in section 5.5.

5.3.3 Pointwise Mutual Information (PMI)

The third translation model we will investigate is derived from the pointwise mutual information (PMI) between the concept-based and word-based event space (Church and Hanks, 1990). PMI indicates the association of two events based on their joint distribution in comparison to their individual probabilities. PMI and mutual information have been frequently used as an association measure for IR (van Rijsbergen, 1979; Lin, 1998a; Chen and Thiel, 2004; Liu et al., 2005) and in particular for filtering ambiguous translations in a CLIR setting (Bian and Chen, 1998; Fung et al., 1999; Jang et al., 1999; Gao et al., 2006). Berger and Lafferty (1999) used the mutual information statistic for constructing a distribution function of words over documents to sample queries for documents. We will

use such a distribution directly as a translation model. We argue that strongly associated concepts and words can be used as translations of each other.

In the literature, definitions of mutual information and pointwise mutual information are frequently confused. In this work, the following definition will be used for PMI.

$$PMI(w,c) = \log_2 \frac{p(w,c)}{p(w)p(c)} \tag{5.2}$$

$$= \log_2 \frac{N\mathrm{f}(w,c)}{\mathrm{f}(w)\mathrm{f}(c)} \tag{5.3}$$

$p(w,c)$ is the probability of encountering the word and concept together in a document collection, and $p(w)$ and $p(c)$ are the probabilities of encountering them separately in the collection. In the subsequent estimation of these probabilities $f(w,c)$ denotes the number of documents in which the words w and c appear together; $f(w)$ and $f(c)$ indicate the number of documents in which the word and concept appear respectively, and N is the size of the collection.

Manning and Schütze (1999) noted that PMI is not an ideal measure for measuring the association between terms, since it is biased towards low-frequency words. Similar to Ballesteros and Croft (1998) and Berger and Lafferty (1999), we circumvent this bias towards low-frequency words by introducing an additional factor based on occurrence frequency of the pair.

$$PMI'(w,c) = \mathrm{f}(w,c) \log_2 \frac{p(w,c)}{p(w)p(c)} \tag{5.4}$$

Based on these scores, we create the translation model for a term in an ad hoc fashion: the n translation terms with the highest PMI' scores are selected and normalised by dividing the sum of the top n scores.

Figure 5.4(b) and Figure 5.5(b) show two examples of translation probabilities for translating from a MeSH concept to words, and from a word to concepts, respectively.

5.3.4 Parsimonious term translation models (PTT)

The fourth translation model we will investigate is based on the conditional probabilities of encountering the target (translation) term after observing the source term in a large set of documents. The translation probabilities are estimated as follows.

$$P(w|c) = \frac{\mathrm{f}(w,c)}{\sum_{w' \in V} \mathrm{f}(w',c)} \tag{5.5}$$

$\mathrm{f}(w,c)$ is the number of times a word and a concept term occur together in a document, and the denominator indicates the sum of co-occurrences of the concept with any word in the word vocabulary.

An important simplification of this approach is that the co-occurrence between a word and a concept is approximated independently from other co-occurrence information between terms and documents. In contrast, the IBM model described in subsection 5.3.2 attempts to align the words and concepts in the comparable corpus.

Using this formula, the translation probabilities for the concept [Mad cow disease] look as follows[1]:

[Mad cow disease] 0.014 bse, 0.011 diseas, 0.011 spongiform, 0.011 encephalopathi, 0.010 bovin, 0.006 transmiss, 0.006 j, 0.006 prion, 0.006 akob, 0.005 creutzfeldt, 0.005 creutzfeldt-jakob, 0.005 infect, 0.005 anim, 0.005 or, 0.005 human, 0.005 cattl, 0.005 case, 0.004 s, 0.004 protein, ...

A number of observations can be made about this translation model. Firstly, the probabilities are quite low. Theoretically, all terms which co-occur with the concept are possible translations. Since there is a large group of terms with a low frequency, this accounts for a large proportion of the probability mass. The estimation requires pruning to remove these low-frequency translations. Secondly, words such as, 'diseas', 'anim', and 'case' receive a high translation probability, simply because they frequently occur in the collection and therefore also frequently co-occur with the concept to translate.

An Expectation Maximisation (EM) algorithm proposed by Hiemstra et al. (2004) is employed to prune these low probability translations and remove these common terms. Na et al. (2007) further explored this approach for monolingual query expansion. Bai and Nie (2008) used a similar method to determine domain models for information retrieval.

We propose to use the EM algorithm as follows. After initialising the translation probabilities with the maximum likelihood estimate defined in Equation 5.5, the EM algorithm will be applied: during the expectation step, the probability mass will be redistributed depending on the global probability of a term. During the maximisation step, the probability distribution will be normalised, that is, normalising the sum of the translations to one.

$$\text{E-step: } e_w = \mathrm{f}(w,c)\frac{(1-\lambda)P(w|c)}{\lambda P(w) + (1-\lambda)P(w|c)} \tag{5.6}$$

$$\text{M-step: } P(w|c) = \frac{e_w}{\sum_{w'} e_{w'}} \tag{5.7}$$

Where $P(w)$ is the probability of encountering the term w in a large collection and λ determines how parsimonious the translation model will be: a value of 0 results in the maximum likelihood estimate; a value close to 1 results in a translation model in which probability mass has been redistributed to fewer translations.

After applying the EM-algorithm[2] the translation model for [Mad cow disease] looks as follows:

[Mad cow disease] 0.251 bse, 0.181 spongiform, 0.073 akob, 0.071 creutzfeldt, 0.069 creutzfeldtjakob, 0.059 cjd, 0.046 scrapi, 0.043 encephalopathi, 0.043 prion, 0.031 vcjd, 0.014 bseinfect, 0.013 tse, 0.011 prpsc, 0.010 nvcjd, 0.006 bseaffect, 0.005 offal, 0.005 ruminantderiv, 0.004 kuru, 0.004 scrapielik, ...

Note that the translation probabilities are considerably higher and note that words such as 'diseas', and 'case' are not listed anymore.

We will refer to these translation models as *parsimonious term translation* models (PTT).

[1] using the TREC Genomics 2004 collection as training data, see section 5.5 for additional details.

[2] with λ set to 0.99, repeating the EM process for 10 iterations and pruning probabilities smaller than 0.001

Figure 5.4(c) and Figure 5.5(c) show two examples of translation probabilities for translating from a MeSH concept to words, and from a word to concepts, respectively.

5.3.5 Translation models based on a thesaurus (THES and STATTHES)

The last two translation models we will investigate use the thesaurus for determining translation probabilities between concepts and terms. In traditional CLIR, similar approaches have been used to use machine readable dictionaries to estimate translation models (Kraaij, 2004).

In the *naive* translation model based on a thesaurus (THES), the translation from words to concepts and vice versa, is estimated by their relative co-occurrence frequencies in entries in the thesaurus. This estimation is made as follows.

$$P(w|c) = \frac{\mathrm{f}(w,c)}{\sum_{w'} \mathrm{f}(w',c)}, \tag{5.8}$$

$\mathrm{f}(w,c)$ is the number of times the word w is used to describe c in the thesaurus. For instance, when a concept [Mice] has synonyms 'mice', 'house mouse', and 'mouse', the probability of $P(\text{mouse}|[\text{Mice}])$ is equal to $\frac{2}{1+1+2} = 0.5$.

Similarly, the probability of translating a word to a concept can be approximated as follows.

$$P(c|w) = \frac{\mathrm{f}(w,c)}{\sum_{c'} \mathrm{f}(w,c')} \tag{5.9}$$

Figure 5.4(d) and Figure 5.5(d) show two examples of translation probabilities for translating from a MeSH concept to words, and from a word to concepts, respectively.

The model based on a *statistical thesaurus* (STATTHES), also takes into account how frequently a particular word is used to refer to a concept in a corpus of documents. This requires the text to be tagged with concepts found in a thesaurus. $\mathrm{f}(w,c)$ is then defined as the frequency that the word w was tagged with the concept c. For instance, when a concept [Mice] has been encountered in a corpus of documents 100 times as 'mice', 50 times as 'house mouse' and 10 times 'mouse', the probability of $P(\text{mouse}|[Mice])$ is equal to $\frac{50+10}{100+50+50+10} = 0.29$.

5.4 Retrieval models for biomedical CLIR

We will investigate a number of retrieval models which incorporate the translation models described in the previous section with a focus on the following points.

Term by term translation We will use the term-by-term translation models (M1, PMI, PTT, THES and STATTHES) introduced in the previous section to carry out document translation and query translation for ad hoc document retrieval. We expect that these translations will not be of high quality, since they are extremely limited in the amount of context they use during translation. However, using these models can provide insight into the quality and usefulness of these term translation models.

[Mad cow disease] 0.228 bse, 0.096 spongiform, 0.096 encephalopathi, 0.038 diseas, 0.030 transmiss, 0.028 cattl, 0.027 infect, 0.025 case, 0.020 agent, 0.019 bovin, 0.019 anim, 0.014 mad, 0.012 epidem, 0.011 variant, 0.011 clinic, 0.010 human, 0.009 scrapi, 0.009 prion, 0.009 tse, 0.009 ban, 0.009 new, ...

(a) IBM model 1 (M1)

[Mad cow disease] 0.141 bse, 0.104 spongiform, 0.082 encephalopathi, 0.053 bovin, 0.048 akob, 0.046 creutzfeldt, 0.045 creutzfeldtjakob, 0.044 prion, 0.036 cjd, 0.030 scrapi, 0.029 cattl, 0.025 transmiss, 0.019 cow, 0.019 diseas, 0.018 vcjd, 0.015 variant, 0.015 mad, 0.013 pr, 0.012 sheep, 0.012 prp, 0.012 tse, ...

(b) Pointwise Mutual Information (PMI)

[Mad cow disease] 0.251 bse, 0.181 spongiform, 0.073 akob, 0.071 creutzfeldt, 0.069 creutzfeldtjakob, 0.059 cjd, 0.046 scrapi, 0.043 encephalopathi, 0.043 prion, 0.031 vcjd, 0.014 bseinfect, 0.013 tse, 0.011 prpsc, 0.010 nvcjd, 0.006 bseaffect, 0.005 offal, 0.005 ruminantderiv, 0.004 kuru, 0.004 scrapielik, 0.004 prpre, 0.003 meatandbon, ...

(c) Parsimonious term translation (PTT)

[Mad cow disease] 0.250 spongiform, 0.250 bovin, 0.125 enceph, 0.125 encephalopathi, 0.062 bse, 0.062 cow, 0.062 mad, 0.062 diseas.

(d) Thesaurus translation model (THES)

Figure 5.4: Translation probabilities of the different translation models for the MeSH concept [Mad cow disease]. M1, PMI and PTT were based on a comparable corpus of documents in a text and concept-based representation. THES was only based on the MeSH thesaurus. The training of these translation models is described in section 5.5.

ferroportin 0.184 [Cation Transport Proteins], 0.085 [Carrier Proteins], 0.076 [Homeostasis], 0.074 [Genetic Heterogeneity], 0.073 [Mutation, Missense], 0.054 [Amino Acid Substitution], 0.052 [Mononuclear Phagocyte System], 0.047 [Membrane Proteins], 0.047 [Receptors, Transferrin], 0.044 [Iron Overload], 0.043 [Italy], 0.042 [Chromosomes, Human, Pair 2], 0.040 [Codon], 0.036 [Lod Score], 0.028 [Iron], 0.024 [Exons], 0.019 [Mice], 0.009 [HLA Antigens], 0.008 [Histocompatibility Antigens Class I], 0.007 [Zebrafish], 0.003 [Hemochromatosis], ...

(a) IBM model 1 (M1)

ferroportin 0.171 [Cation Transport Proteins], 0.113 [Hemochromatosis], 0.096 [Iron], 0.067 [Iron-Binding Proteins], 0.065 [Iron Overload], 0.058 [Receptors, Transferrin], 0.044 [Membrane Proteins], 0.041 [Ferritins], 0.036 [Histocompatibility Antigens Class I], 0.032 [Genes, Dominant], 0.028 [Duodenum], 0.025 [Transferrin], 0.025 [Mutation, Missense], 0.023 [Iron-Regulatory Proteins], 0.022 [Carrier Proteins], 0.022 [Enterocytes], 0.018 [Kupffer Cells], 0.016 [RNA, Messenger], 0.013 [Family Health], 0.013 [Mutation], 0.012 [Homeostasis], ...

(b) Pointwise Mutual Information (PMI)

ferroportin 0.313 [Cation Transport Proteins], 0.205 [Hemochromatosis], 0.122 [Iron-Binding Proteins], 0.119 [Iron Overload], 0.097 [Receptors, Transferrin], 0.043 [Iron-Regulatory Proteins], 0.039 [Enterocytes], 0.031 [Ferritins], 0.015 [Kupffer Cells], 0.010 [Iron], 0.004 [Transferrin], 0.002 [Mutation, Missense].

(c) Parsimonious term translation (PTT)

ferroportin *no translations available*

(d) Thesaurus translation model (THES)

Figure 5.5: Translation probabilities of the different translation models for the text word 'ferroportin'. M1, PMI and PTT were based on a comparable corpus of documents in a text and concept-based representation. THES was only based on the MeSH thesaurus. The training of these translation models is described in section 5.5.

Enhancing translation through pruning The pseudo-feedback translation model (KNN) was shown to be able to improve word-based retrieval in chapter 4. A drawback of the approach is that the obtained translation is quite large (up to 50 concepts are used to represent a query) and that many of these concepts do not show a clear relationship to the information need described in text. We hypothesise that the term-by-term translation models (M1, PMI, PTT, THES and STATTHES) can be used to prune these noisy and redundant concepts from the pseudo-feedback translation, without a loss of retrieval effectiveness.

Further enhancing word-based retrieval We further explore how a combination of word-based retrieval and concept-based retrieval based on pseudo-feedback can be improved. In particular, we propose to use the term-by-term translation models to reweigh and structure the word-based query representations.

These focus points will now be described in more detail.

5.4.1 Term-by-term translation

Inspired by models used for integrating term-by-term translation in traditional CLIR (Kraaij, 2004), we will investigate two approaches to incorporate these translations in biomedical CLIR.

Query translation

The first term-by-term translation model we will investigate is based on *query translation* to estimate a query in a concept-based language. In this translation, each query word is independently translated to a concept-based representation. The translated concept language model is estimated as follows.

$$P(c|\phi_Q) = \sum_{w \in Q} P(c, w|Q) = \sum_{w \in Q} P(c|w, Q)P(w|\theta_Q) \tag{5.10}$$

$$\approx \sum_{w \in Q} P(c|w)P(w|\theta_Q) \tag{5.11}$$

$w \in Q$ are the words in the textual query, and $P(c|w)$ is the translation probability of translating the word w to the concept c. It is important to note the simplification of $P(c|w, Q)$ to $P(c|w)$ – the translation of a word is assumed to be independent from the query it appeared in. Clearly, this simplification may intensify problems related to ambiguity. For instance, the translation of the word 'labor' to both [Labor Force] and [Labor, Induced]. Even a single additional word found in the query might aid in disambiguating the word. However, the simplification allows fast translation by simply performing a lookup of the term in a translation table.

Document translation (DT)

The second term-by-term translation model we will investigate is based on *document translation*. In this case the conceptual document representation is translated concept-by-concept to the equivalent textual document representation which is matched to the textual

query representation. Hence, matching occurs in a text-based rather than concept-based representation. Formally, the (translated) textual document language model is estimated as follows.

$$P(w|\theta_D) = \sum_{c \in D} P(w, c|\phi_D) = \sum_{c \in D} P(w|c, \phi_D) P(c|\phi_D) \tag{5.12}$$

$$\approx \sum_{c \in D} P(w|c) P(c|\phi_D) \tag{5.13}$$

Where $c \in D$ are the concepts assigned to the document D; $P(w|c)$ is the translation probability of translating the concept c to the word w. Such an approach to translation is quite attractive: the concept-based representation is unambiguous and therefore translation to text is likely to be correct. Ambiguity is introduced, however, at the point where single words in the query are matched against single (translated) words in the document.

5.4.2 Enhancing translation by pruning

In traditional CLIR, combining different translation resources has shown to be an effective way to improve translation quality (Ballesteros and Croft, 1998; Gollins and Sanderson, 2001; Boughanem et al., 2002). In the previous chapter we showed that pseudo-feedback could be effectively used to obtain a concept-based representation of a text-based query. But since this representation is based on documents, it is expected to contain noisy concepts which are only indirectly related to the original query.

Guided by the experiences in traditional CLIR, we hypothesise that this translation can be further improved when combined with the term-by-term translation models introduced in this chapter. More specifically, we propose to use the concept to word translation models to prune concepts from the translated concept-based query obtained through feedback.

The concept-based representation obtained by feedback translation is filtered as follows.

$$P(c|\phi_Q) = \begin{cases} \kappa \; P_{\text{KNN}}(c|\phi_Q) & \text{if } \sum_{w \in Q} P(w|c) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.14}$$

Where $P_{\text{KNN}}(c|\phi_Q)$ is the conceptual query language model estimated through feedback; $P(w|c)$ is a concept to word translation model; and κ is a query dependent normalisation constant, which normalises $\sum_c P(c|\phi_Q)$ to 1.

Note that this type of pruning based on term-by-term translation models is not restrictive: concepts are only pruned from the translation when this concept cannot be translated to any of the query words; the translation probability in the concept-to-word translation model is not taken into account.

In appendix D.1 two examples of pruning a concept-based query representation using a term-by-term translation model are listed.

5.4.3 Enhancing word-based retrieval: reweighting

A well-known drawback of using pseudo-relevance feedback is possible query drift: an expanded query can overemphasise or neglect particular aspects from the original query, or

skew towards aspects not mentioned in the original query (Manning et al., 2008). In the case of a pseudo-feedback translation to a conceptual representation, the neglect of a particular query aspect can be substantiated by the fact that aspects cannot be represented accurately by the concept vocabulary. As a result, combining a word and concept-based representation based on feedback may understate aspects present in the word-based representation. The goal of the reweighting procedure we will now describe is to prevent that a word-based query combined with a concept-based query (obtained through feedback) neglects aspects found in the word-based query. To achieve this, the word-based query model is reweighted: depending on how well the concept-based representation *covers* the words in the query, the word weights are updated: well-covered words receive a lower weight, whereas poorly covered words receive an increased weight.

The reweighting process is used as follows.

- The parameters of the original word-based query model $P(w|\theta_Q)$ are based on words provided by the user. Using the feedback translation model the parameters of a concept-based query model $P(c|\phi_Q)$ are estimated.
- The coverage of the words in the original word-based query model $P_{cov}(w|\phi_Q)$ is determined by translating the concept-based query model using the term-by-term translation models described earlier.
- An updated word-based query model $P(w|\theta'_Q)$ is determined based on the coverage of the words by the concept-based query model. The updated word-based query model is combined with the concept-based query model for retrieving documents.

How the coverage and updated word-based query model are determined will now be described.

Determining the coverage of the word-based query

The coverage of a word-based query by a concept-based representation is defined as a probability distribution over the words in the original query. If the word-based query is evenly covered by a concept-based representation this probability distribution is uniform: all query words are covered by concepts in the concept-based representation. Well-covered words by the concept-based representation receive a high coverage probability; poorly covered words receive a lower probability.

We use a term-by-term translation model to determine this coverage as follows.

$$P_{cov}(w|\phi_Q) = \frac{\sum_c P(w|c,\phi_Q)P(c|\phi_Q)}{\sum_{w' \in Q}\sum_c P(w'|c,\phi_Q)P(c|\phi_Q)} \tag{5.15}$$

$$\approx \frac{\sum_c P(w|c)P(c|\phi_Q)}{\sum_{w' \in Q}\sum_c P(w'|c)P(c|\phi_Q)} \tag{5.16}$$

$P(c|\phi_Q)$ is the concept language model obtained through pseudo-feedback translation of the original word-based query and $P(w|c)$ is the term-by-term translation probability of translating a concept c to a word w. In the (unlikely) case that none of the concepts can be translated to a query word, $P_{cov}(w|\phi_Q)$ is equal to 0 for all w[3].

[3]This can be viewed as a coverage of a *null*-query word with probability 1.

Updating the word-based query language model

The coverage of the original word-based query language model is used to determine an updated word-based query language model.

We assume that all the aspects mentioned in the original text-based query are equally important: when searching with a combined word and concept-based query representation this balance should be maintained. When the concept-based representation does not cover all query aspects this balance is disturbed: some aspects are overemphasised leading to query drift. This query drift of a combined word and concept-based query representation can be prevented by decreasing the weight of words which are well covered by the concept-based representation.

We assume that the aspects of a query can be represented by the original word-based query language model (based on a maximum likelihood estimate). To retain the original query balance, the updated word-based query language model combined with the coverage by the concept-based query language model should approximate the original query word distribution. Formally, this balance is modelled as follows.

$$P(w|\theta_Q) = \beta_Q \; P_{cov}(w|\phi_Q) + (1-\beta_Q)P(w|\theta'_Q) \tag{5.17}$$

$P(w|\theta_Q)$ is the original query word language model, which should be covered by the translation of a conceptual query language model $P_{cov}(w|\phi_Q)$ and by an updated query language model $P(w|\theta'_Q)$. The query dependent parameter β_Q indicates the relative importance of the updated word-based query language model in comparison to the translated concept-based query language model.

To approximate Equation 5.17, initial estimates of the updated word-based query language model are calculated as follows.

$$e_w = \begin{cases} P(w|\theta_Q) & \text{if } P_{cov}(w|\phi_Q) = 0 \\ \dfrac{P(w|\theta_Q) - \beta_Q P_{cov}(w|\phi_Q)}{1-\beta_Q} & \text{otherwise} \end{cases} \tag{5.18}$$

The updated query language model is determined by normalising these initial estimates.

$$P(w|\theta'_Q) = \frac{e_w}{\sum_{w' \in Q} e_{w'}} \tag{5.19}$$

Note that the second line of the equation is obtained by rewriting Equation 5.17. The value β_Q has to be restricted to prevent $P(w|\theta'_Q)$ becoming less than zero.

$$0 \leq \beta_Q \leq \min_{w \in \phi_Q} \frac{P(w|\theta_Q)}{P_{cov}(w|\phi_Q)} \tag{5.20}$$

A β-value of 0 indicates that the updated word-based query language model is exactly the same as the original word-based query model; the largest possible value of β modifies $P(w|\theta_Q)$ as much as possible to retain the original query term balance.

Table 5.2 illustrates this reweighting in practice for a query consisting of three words (w_1 to w_3). Their original importance weights, based on the original query formulation is found

Table 5.2: Example of query term reweighting. The original weight is updated according to the coverage. β_Q determines the amount of reweighting.

	Original $P(w\|\theta_Q)$	Coverage $P_{cov}(w\|\phi_Q)$	Updated weight $P(w\|\theta'_Q)$ $\beta_Q = 0$	$\beta_Q = 0.1$	$\beta_Q = 0.25$
w_1	0.5	0.1	0.5	0.54	0.63
w_2	0.4	0.5	0.4	0.39	0.37
w_3	0.1	0.4	0.1	0.07	0

in the second column. The third column indicates to what extent the words are covered by concepts found in the query. w_1 for example, has an original probability of 0.5, but is only covered by the translation with a probability of 0.1. The updated probability should therefore be higher than 0.5. The last three columns of the table show the re-estimated weights for three different values of β_Q. The highest possible value of β_Q for this query is 0.25, resulting in a reweighted probability for the word w_3 of 0.

To control the value of β_Q at a global level (that is across different queries), we introduce the parameter α (between 0 and 1) which linearly scales β_Q between its minimum and maximum value.

$$\beta_Q = \alpha \min_{w \in Q} \frac{P(w|\theta_Q)}{P_{cov}(w|\phi_Q)} \tag{5.21}$$

In appendix D.2 a number of examples of reweighting word-based query language models are listed.

5.4.4 Enhancing word-based retrieval: structuring

The last approach we investigate to combine translation models combines the original textual query with a conceptual query based on pseudo feedback into a structure. The approach is motivated by the idea that the translated concepts should be linked to the query words they represent. We hypothesise that such an approach balances the original textual query with its translation, and prevents query drift.

To allow for such an integration we need to model words and concepts in the same event space. We achieve this by simply merging the two representations, that is mixing the identifiers of the concepts with the tokens extracted from the text. From a principled modelling perspective, mixing the representations is not attractive: words and concepts are different units of information and should therefore be kept separated. However, the mixed representation is easy to understand and straightforward to implement.

The parameters of the mixed document language model $P(t|\psi_D)$ are again based on a maximum likelihood estimation, smoothed with a background language model based on the mixed corpus as a whole (analogous to eq. 2.1.3 on p. 15). Note that in this case, the length of the document is the sum of the number of tokens and concepts in the document.

The initial parameters of the mixed query language model $P(t|\psi_D)$ are based on a linear interpolation of the word-based query model and the concept-based query model. Formally, this interpolation is defined as follows.

$$P(t|\psi_Q) = \alpha\ P(t|\theta_Q) + (1-\alpha)P(t|\phi_Q) \tag{5.22}$$

α indicates the relative importance of the text-based representation with respect to the concept-based representation.

We will now use a translation model $P(w|c)$ to create an *alignment* between the concepts and the words in this mixed query language model. Based on the translation model, each concept is assigned to (at most) one word. Assuming that the l terms in the word-based query are w_1 to w_l, and that the m concepts in the concept-based query are c_1 to c_m, we can define an alignment function between c_i and w_j as follows.

$$\delta(c_i, w_j) = \begin{cases} 1 & \text{if } j = \arg\max_{j'} P(w'_j|c_i) \\ 0 & \text{otherwise} \end{cases} \tag{5.23}$$

In words: the concept c_i is aligned to the word w_j with the highest translation probability. We now define $\sigma(w_j)$ of a word w_j as the set containing the word itself and the concepts which have been assigned to it.

$$\sigma(w_j) = \{w_j\} \cup \{\ c_i\ ; \delta(c_i, w_j) = 1\} \tag{5.24}$$

Similar to Kraaij (2004, p. 133), we use this set to define an equivalence class of the word and the concepts mapped to it.

$$P(class(w_j)|\psi_D) = \sum_{t \in \sigma(w_j)} \frac{P(t|\psi_Q)}{\sum_{t' \in \sigma(w_j)} P(t'|\psi_Q)} P(t|\psi_D) \tag{5.25}$$

The query language model of the equivalence class is defined as follows.

$$P(class(w_j)|\psi_Q) = P(w_j|\psi_Q) \tag{5.26}$$

In appendix D.3 a number of examples of structuring a word and concept-based query are listed.

5.5 Experimental setup

This section will describe the experimental setup for comparing the different translation and retrieval models.

As in the previous chapters, the TREC Genomics benchmark collections and topic sets were used for determining retrieval performance. The translation models which required training data (all except for the naive thesaurus translation model), were trained with documents from the TREC Genomics 2004 document collection.

Word-based representations of these documents were obtained using the combined tokenizer described in chapter 3. The MeSH-based representations of the documents were based on the major MeSH headings assigned by NLM indexers; subheadings were discarded. The UMLS$_{++}$-based document representation was obtained using the Peregrine

tool described in subsection 4.2.6. The document collection in word-based and UMLS$_{++}$-based representations was used both as a parallel and a comparable corpus. For training the STATTHES translation models, the explicit alignment between words and concepts (obtained from Peregrine) were used. For the other translation models, the alignment was discarded and the representation was treated as a comparable corpus[4]. The document collection in word-based and MeSH-based representations was only used as a comparable corpus. Translation models were built for translation between MeSH and words, UMLS$_{++}$ and words and vice versa.

The translation models for PMI and PTT were based on co-occurrence counts of words and concepts in the complete 2004 document collection. Because of scalability issues, the IBM model 1 translation models were built on a subset of the collection. 1,200,000 randomly selected documents from the collection were used to build the translation models, which is around one quarter of the collection. A slightly modified version of the GIZA++[5] machine translation toolkit was used to train the models based on IBM model 1. The default setting of 5 iterations of the EM algorithm was used.

For the translation models based directly on the thesauri (THES), the conditional probabilities were determined as follows: all synonymous terms for a concept found in the thesaurus were tokenised with the combined tokenizer described in chapter 3. This resulted in tokens labeled with concept identifiers. The number of times a token was encountered with a concept identifier was used as f(w, c) in Equation 5.8 and Equation 5.9. For MeSH, the 2008 thesaurus was used, treating all 'ENTRY' lines as synonymous terms for a MeSH concept. For UMLS$_{++}$, all 'TM' entries in the ontology file provided by the Biosemantics Group of the Erasmus University were used as synonyms[6].

The conditional probabilities used as translation probabilities in the statistical thesaurus translation model for UMLS$_{++}$ (STATTHES) were determined in a similar fashion.

- The TREC 2004 collection was tagged with UMLS$_{++}$ concepts using Peregrine. This resulted in a large number of text phrases labeled with UMLS$_{++}$ concept identifiers.
- The phrases were tokenised with the combined tokenizer described in chapter 3. This resulted in a large number of token-concept assignments. The number of times a token was encountered with a UMLS$_{++}$ concept identifier was used as f(w, c) in Equation 5.8 and Equation 5.9.

All translation models went through the following post-processing to remove noise.

- Translations with a probability smaller than 0.001 were removed;
- Words or concepts which occurred in fewer than 3 documents in the collection were pruned;
- Single character words and numbers were removed.

The remaining translations were normalised for each term, that is for both words and concepts (assuring $\sum_{t'} P(t'|t) = 1$).

[4]Note that IBM Model 1 does attempt to align the representations

[5]`http://www.fjoch.com/GIZA++.html`

[6]`https://wiki.nbic.nl/index.php/ErasmusMC_ontology_file_format`

5.6 Results

This results section is structured as follows. First, we will investigate the retrieval performance of the two retrieval models based on term-by-term translation models. In sections 5.6.2 to 5.6.4 we will look into the effectiveness of combining the pseudo-feedback translation model with the term-by-term translation models for pruning, reweighting and structuring respectively.

5.6.1 Term-by-term translation

In subsection 5.4.1, we described two retrieval models based on term-by-term translation: query translation and document translation. In section 5.3, we described five translation models that can be used for term-by-term translation: translation models based on IBM Model 1 (*M1*), Parsimonious Term Translation (*PTT*), Pointwise Mutual Information (*PMI*), the thesaurus (*THES*), and a statistical thesaurus (*STATTHES*). To compare the individual quality of these translation models we will first discuss their retrieval performance when only these translations are used for retrieval. As a baseline, the translation based on pseudo-feedback will be used. After that, we will determine whether these fairly basic approaches to translation can be used to improve word-based retrieval by combining a translated and word-based representation.

Query translation

Table 5.3 lists the mean average precision obtained on the four TREC Genomics topic sets when using the translation models for query translation. Results using a statistical thesaurus (STATTHES) are only reported for UMLS$_{++}$ since only for this vocabulary was such a translation model available.

As expected, the results of translating the textual representation in a word-by-word fashion to a MeSH concept-based representation (Table 5.3(a)) were significantly worse (between 33 and 84%) than translating the textual query as a whole through pseudo-feedback (KNN). The results in early precision (P@10, not reported) show similar significant drops in performance. Only considering single, isolated, query words for translation clearly is not beneficial for this concept representation vocabulary. The translation models trained on the comparable corpus (M1, PTT and PMI), performed significantly better than the translation model based solely on thesaurus information (THES). No significant differences were observed, however, between M1, PTT, and PMI.

Most of the term-by-term query translations to the UMLS$_{++}$ language also performed worse (up to a loss of 58%) than the baseline based on feedback. In many cases, the differences were not significant however. In particular, IBM model 1 (M1) and the statistical thesaurus (STATTHES) did not perform statistically different from the feedback baseline. Surprisingly, term-by-term translation based on the statistical thesaurus (STATTHES) even outperformed (+4.9%) the feedback baseline on the 2006 topic set. This small improvement can be attributed to the type of topics in the 2006 set: the topics request information about the relationship between two concepts, often described in single words (for example, “how does p53 affect apoptosis?”). In these cases, term-by-term translation results in a balanced query containing both required concepts, rather than a much more noisy translation

obtained from pseudo-feedback translation (KNN). STATTHES performed poorly because it translated rather general words ('cancer') to specific concepts (such as [prostate] and [breast neoplasms]). From the group of models based on comparable corpora (M1, PTT and PMI), M1 performed slightly better than PTT or PMI.

Document translation

Table 5.4 lists the results of *document translation* using term-by-term translation models. Note that these experiments were based on a translation from concepts to the text, rather than from text to concept (which was the case for the query translation). As a consequence, different translation models were used: for query translation we used $P(c|t)$, for these experiments $P(t|c)$ was used.

The results show that document translation performs poorly, both in contrast to pseudo-feedback translation and to query translation described in the previous section. In few cases, document translation outperformed the query translation based on the same (similarly trained) translation model. For MeSH, all term-by-term translation systems performed significantly worse (between 48.4 and 81.5%) in terms of MAP than the baseline system. For UMLS$_{++}$, all systems performed worse than the baseline, in more than half of the cases significantly worse. It is notable that M1 and STATTHES performed slightly better than the other term-by-term translation models.

Combination with text-based retrieval

In the previous chapter, we observed that text-based retrieval could be significantly improved when a text representation and a concept representation obtained through feedback, were combined. In these cases, retrieval based on only the conceptual representation alone was also inferior to word-based retrieval. In additional experiments, which we will not exhaustively describe here, we observed that the concept-based query representations obtained from query translation could still be combined with a text based representation to achieve significant (but smaller) improvements over word-based retrieval alone (without word-based relevance feedback). The original interpolation between a concept-based representation based on pseudo-feedback combined with text achieved up to 9.5% improvement in MAP for MeSH and 9.9% for UMLS$_{++}$. In contrast, the translations obtained with term-by-term translation improved MAP up to 2.4% for MeSH and 4.9% for UMLS$_{++}$. The interpolation parameter (see Equation 4.9 on page 89) should be, however, more biased towards word-based retrieval.

Combining the matching of document translation with word-based retrieval did result in small and mostly insignificant improvements in retrieval effectiveness.

Discussion

Despite the relatively poor results in terms of retrieval performance, the experiments provide valuable information about the different translation models and the value of the used representation languages.

The experiments did show the value of having either a translation model based on a comparable corpus, or a statistical thesaurus for translation over using a translation model based on the thesaurus alone. Both approaches strongly outperformed the translation model

based on only the thesaurus. This corresponds to results in chapter 4, where classification systems based on a collection (MTI, KNN, CLM) performed better than systems using only the thesaurus information (ATM, EAGL, MetaMap).

The query translation experiments using the MeSH language again underlined its lack of specificity. Translating the textual query in a word-by-word fashion to a conceptual representation was far less effective than using a concept-based representation based on pseudo-feedback. Given the limited exhaustiveness of the MeSH-based document representations, it is unlikely that a precise translation of the original textual query will result in higher precision or recall. The translation based on pseudo-feedback was less precise, but these additional MeSH terms turned out to be quite useful for improving the recall of retrieval.

In contrast, the query translation experiments with the UMLS$_{++}$ language underline its value in precisely representing the original textual query. Despite the limited context taken into account, word-by-word translation performed slightly worse than feedback translation. Especially the IBM model 1 and statistical thesaurus translation models performed well. We think that this performance can be attributed to the (attempted) alignment between words and concepts, which results in cleaner translation models. The PTT and PMI methods are limited to determining the most strongly associated words and concept pairs and have not attempted to determine translations which distinguish concepts.

The experiments indicated a difference in quality and usefulness of the translation models in different directions. For traditional CLIR, Kraaij (2004) hypothesised that translation models between languages with a large vocabulary to languages with a small vocabulary can be more reliably estimated than vice versa[7]. For biomedical CLIR, the word vocabulary is considerably larger than the concept vocabulary, so following this line of reasoning one would expect that $P(c|t)$ can be more reliably estimated than $P(t|c)$ (based on a comparable corpus). Hence, this could explain the deteriorated results in document translation in comparison to query translation using the term-by-term translation models. It is quite counterintuitive that $P(t|c)$ cannot be reliably determined: given the concept it is quite clear to what words it should be translated. A more plausible explanation is the difference in *granularity* between the two representation vocabularies. For document translation, we noticed that many, specific concepts were found which translated with a high probability to more general words. For instance, the word 'mouse' had high translation probabilities for the UMLS$_{++}$ concepts [Mouse strains], [Mouse bioassay], and [Mouse cell line]. As a result, documents with these specific concepts were ranked inappropriately high. Hence, it is rather a problem related to an inappropriate translation and matching unit (a single word as a translation for a complex concept). Handling this challenge would be a very interesting direction for future work (see section 5.7).

5.6.2 Pruning representations

The effect of pruning a concept-based representation model obviously depends on how much is actually pruned. Table 5.5 lists the percentages of concepts pruned from the concept-based representation (based on KNN) using the different term-by-term translation models. It should be noted that this is the *least* restrictive pruning we can employ using

[7]It should be noted that these probabilities were determined on a sentence-aligned parallel corpus, where we built our translation models using a comparable corpus.

Table 5.3: Retrieval performance in terms of MAP based on only term-by-term query translation. [1], [2] and [3] indicate significant differences to the baseline at confidence levels 0.05, 0.01 and 0.001 respectively, determined with a paired sign test. The highest value of each column is printed in boldface.

(a) Word to MeSH translation

	MAP			
Model	2004	2005	2006	2007
KNN (MeSH)	**0.1889**	**0.1268**	**0.2518**	**0.1901**
M1 qt	0.1024 [3] -45.8%	0.0855 [2] -32.6%	0.1330 [2] -47.2%	0.1003 [2] -47.2%
PTT qt	0.0878 [3] -53.5%	0.0719 [2] -43.3%	0.1448 [2] -42.5%	0.1052 [2] -44.6%
PMI qt	0.1023 [3] -45.8%	0.0788 [2] -37.8%	0.1485 [2] -41.0%	0.1047 [2] -44.9%
THES qt	0.0301 [3] -84.1%	0.0303 [3] -76.1%	0.1224 [2] -51.4%	0.0536 [3] -71.8%

(b) Word to UMLS$_{++}$ translation

	MAP			
Model	2004	2005	2006	2007
KNN (UMLS$_{++}$)	**0.2799**	**0.1670**	0.3535	**0.2355**
M1 qt	0.2213 -20.9%	0.1441 -13.7%	0.3522 -0.4%	0.2165 -8.1%
PTT qt	0.2001 [2] -28.5%	0.1213 [2] -27.3%	0.3305 -6.5%	0.1867 -20.7%
PMI qt	0.1710 [3] -38.9%	0.1140 [2] -31.7%	0.3096 -12.4%	0.1745 -25.9%
STATTHES qt	0.2417 -13.6%	0.1280 -23.3%	**0.3708** +4.9%	0.2233 -5.2%
THES qt	0.1260 [3] -55.0%	0.0700 [3] -58.1%	0.2536 -28.2%	0.1511 [3] -35.8%

Table 5.4: Retrieval performance in terms of MAP based on only term-by-term document translation. See Table 5.3 for legend.

(a) Word to MeSH translation

	MAP			
Model	2004	2005	2006	2007
KNN (MeSH)	**0.1889**	**0.1268**	**0.2518**	**0.1901**
M1 dt	0.0557 [3] -70.5%	0.0599 [3] -52.7%	0.0849 [2] -66.3%	0.0463 [3] -75.6%
PTT dt	0.0504 [3] -73.3%	0.0654 [3] -48.4%	0.1032 [2] -59.0%	0.0526 [3] -72.3%
PMI dt	0.0418 [3] -77.9%	0.0492 [3] -61.2%	0.0685 [2] -72.8%	0.0410 [3] -78.4%
THES dt	0.0350 [3] -81.5%	0.0427 [3] -66.3%	0.0667 [2] -73.5%	0.0375 [3] -80.3%

(b) Word to UMLS$_{++}$ translation

	MAP			
Model	2004	2005	2006	2007
KNN (UMLS$_{++}$)	**0.2799**	**0.1670**	**0.3535**	**0.2355**
M1 dt	0.2027 [2] -27.6%	0.1247 [2] -25.3%	0.3373 -4.6%	0.2326 -1.2%
PTT dt	0.1523 [3] -45.6%	0.0877 [3] -47.5%	0.2551 -27.8%	0.1725 [2] -26.7%
PMI dt	0.1309 [3] -53.2%	0.0794 [3] -52.5%	0.2882 -18.5%	0.1301 [2] -44.7%
STATTHES dt	0.2040 [2] -27.1%	0.1050 [2] -37.1%	0.3146 -11.0%	0.1853 [2] -21.3%
THES dt	0.1146 [3] -59.1%	0.0426 [3] -74.5%	0.2221 -37.2%	0.1636 -30.5%

these translation models, that is without changing the pruning applied to the translation models themselves (as described in section 5.5). The table clearly indicates that even the most restrictive type of pruning removed many concepts: between 49.9% up to 91.5% of the concepts were removed. The translation models based on PMI and IBM model 1, resulted in the most restrictive pruning (between 49.9% and 79.1%); the models based on PTT and the thesauri (THES and STATTHES) resulted in stronger pruning (between 81.9% and 91.5%).

In Table 5.6 the effect of pruning on retrieval performance is listed. As a baseline, the (unpruned) representations based on retrieval feedback (KNN) are reported.

For the MeSH representation, pruning resulted in significant losses (between 18.4% and 65.3%) in retrieval effectiveness for the 2004 and 2005 topic sets. For the 2006 and 2007 topic sets results also deteriorated but fewer significant differences were observed. The models which were most restrictive in their pruning (M1 and PMI), outperformed models applying more rigorous pruning (as expected). Similar to the results on query translation, these results indicate that the MeSH representation language is not very useful in precisely representing information needs: many MeSH concepts are required to accurately fulfil an information need. Retrieval performance was hurt by reducing the representation to concepts which can be translated back into words found in the original query.

For the UMLS$_{++}$ representation, pruning did not result in the expected improvements in retrieval performance either. A small improvement (2.1%) over not pruning was observed, for only a single experiment (2006 topic set, pruned with the PMI translation model). Considering the large number of pruned concepts (between 61.5 and 91.5%), however, the

Table 5.5: Percentage of concept-terms pruned from the concept-based representation obtained from feedback using the different translation models.

(a) MeSH representation

Model	2004	2005	2006	2007
M1	55.0%	67.7%	67.6%	63.7%
PTT	86.8%	85.3%	86.7%	90.2%
PMI	49.9%	63.9%	64.7%	58.3%
THES	86.9%	89.5%	89.5%	90.1%

(b) MeSH representation

Model	2004	2005	2006	2007
M1	65.7%	75.9%	79.1%	77.9%
PTT	83.3%	81.9%	84.5%	88.7%
PMI	50.5%	63.3%	66.9%	61.5%
STATTHES	86.0%	86.4%	89.4%	91.5%
THES	84.5%	85.7%	87.4%	88.7%

losses in performance were only small (between 1.5 and 24.8%).

In the previous experiments we only investigated concept-only retrieval. Our goal, however, was to improve word-based retrieval using these concept-based representations. We also investigated whether these pruned concept-based representations could be used to improve word-based retrieval.

For MeSH, smaller (up to 6% rather than 10%), but still significant improvements over word-based retrieval were observed when these representations were interpolated[8] with word-based retrieval for the 2004 and 2005 topic sets. For the 2006 and 2007 topic sets no significant improvements were observed.

For the UMLS$_{++}$ representation, however, interpolating the pruned concept representations with the text representations turned out to be almost as effective as or even more effective than the unpruned representation. Irrespective of the type of translation model used for pruning, significant improvements (up to 10.5% in MAP) over the text-based baseline were observed. These improvements are listed in Table 5.7. For UMLS$_{++}$ pruning turned out to be useful: between 50.5% and 91.5% of the terms in the concept-based query could be pruned while improving the combined retrieval effectiveness.

5.6.3 Reweighting representations

The effect of reweighting word terms obviously depends on how much weight is actually transferred. As explained in subsection 5.4.3, this amount is controlled by the parameter α. Table 5.8 lists the average probability mass that was transferred as a result of reweighting based on the translation models. As intended, larger values of α resulted in the redistribution of more weight. No large differences were observed for the different translation

[8]The interpolation had to be skewed towards the word-based representation

Table 5.6: Retrieval performance after pruning the concept-based representation based on KNN using different translation models. See Table 5.3 for legend.

(a) MeSH representation

	MAP			
Model	2004	2005	2006	2007
KNN (MeSH)	**0.1889**	**0.1268**	**0.2518**	**0.1901**
M1 prune	0.1257 [3] -33.5%	0.1034 [3] -18.4%	0.2353 -6.6%	0.1384 -27.2%
PTT prune	0.0989 [3] -47.7%	0.0924 [3] -27.1%	0.2278 -9.5%	0.1141 [2] -40.0%
PMI prune	0.1132 [3] -40.1%	0.0934 [2] -26.3%	0.2277 -9.6%	0.1409 -25.9%
THES prune	0.0655 [3] -65.3%	0.0649 [3] -48.8%	0.1982 -21.3%	0.0943 [3] -50.4%

(b) UMLS$_{++}$ representation

	MAP			
Model	2004	2005	2006	2007
KNN (UMLS$_{++}$)	**0.2799**	**0.1670**	0.3535	**0.2355**
M1 prune	0.2556 -8.7%	0.1401 [2] -16.1%	0.3434 -2.8%	0.2201 -6.5%
PTT prune	0.2401 -14.2%	0.1449 -13.2%	0.3388 -4.1%	0.2014 -14.5%
PMI prune	0.2575 -8.0%	0.1484 -11.1%	**0.3608** +2.1%	0.2301 -2.3%
STATTHES prune	0.2687 -4.0%	0.1256 [2] -24.8%	0.3343 -5.4%	0.2163 -8.2%
THES prune	0.2594 -7.3%	0.1273 [2] -23.8%	0.3374 -4.6%	0.2319 -1.5%

Table 5.7: Retrieval effectiveness when the pruned concept representations are combined with a word-based representation. See Table 5.3 for legend.

	MAP			
Model	2004	2005	2006	2007
Text	0.3576	0.2219	0.3889	0.2796
Text + unpruned	**0.3929** [2] +9.9%	0.2285 +3.0%	0.4048 +4.1%	0.2981 +6.6%
Text + M1 prune	0.3854 [2] +7.8%	0.2275 +2.6%	0.4114 +5.8%	0.3062 [3] +9.5%
Text + PTT prune	0.3801 +6.3%	0.2293 +3.4%	0.4077 +4.9%	0.2947 [2] +5.4%
Text + PMI prune	0.3853 [2] +7.7%	**0.2319** [2] +4.5%	**0.4179** +7.5%	0.3005 [2] +7.5%
Text + STATTHES prune	0.3796 [2] +6.2%	0.2297 +3.5%	0.4122 +6.0%	0.3063 [2] +9.6%
Text + THES prune	0.3806 [2] +6.4%	0.2303 +3.8%	0.4089 +5.1%	**0.3089** [2] +10.5%

Table 5.8: Average probability mass redistributed as a result of the reweighting.

α	UMLS$_{++}$	MeSH
0.1	0.033	0.092
0.3	0.105	0.105
0.5	0.189	0.167
0.7	0.289	0.260
0.9	0.409	0.372

models used.

Table 5.9 lists the results of reweighting word terms based on the translation models obtained with α set to 0.5. As a baseline, the original combination of word and concept-based representations based on feedback was used. The translation models were subsequently used to reweigh the word-based query representation.

Reweighting word terms based on the MeSH representation and translation models performed poorly. In the best cases, no or small differences ($<$1%) were observed compared to the baseline. In many more cases, however, reweighting significantly decreased retrieval effectiveness. For different values of α (0.1, 0.3, 0.7 and 0.9), reweighting using the MeSH representation could also not beneficially affect performance either. The poor results can be explained by the fact that a MeSH representation is not capable of covering words completely. For instance, assuming that the MeSH concept [Parkinson's Disease] can cover the word 'parkinson' might be incorrect, since the MeSH concept has not been exhaustively assigned to documents about it. The difference in granularity between representations, encountered earlier during document translation, might have also lead to the false assumption that a query word is covered and therefore its weight can be decreased. For instance, based on the presence of the specific concept [Postencephalitic Parkinson Disease] (which clearly translates to the word 'parkinson'), removing weight from the word 'parkinson' is probably not beneficial to retrieval.

Reweighting based on the UMLS$_{++}$ representation turns out to be more effective. Improvements (up to 5.3%) could be observed for the 2005, 2006 and 2007 topic sets. Many of the results are insignificant however. For the 2004 topic set, no improvements could be observed, even with different values of α. The effect of reweighting turned out to be independent from the translation model used.

5.6.4 Structuring representations

Table 5.10 lists the impact of the structured query model described in subsection 5.4.4. It shows, for example, that using the IBM model 1 translation model and UMLS$_{++}$ representation vocabulary, on average resulted in 3.1 equivalence classes and that 5.7 concepts were grouped into one or more of these equivalence classes. Considering the original number of concepts in a query (50), the structuring does not result in very large changes to the original query.

Table 5.11 lists the results when using the structured queries for retrieval in comparison to a baseline using the original unstructured queries. Structuring the representations turned out to give strongly varying results, from significant deteriorations (up to 22.7% in MAP) to

Table 5.9: Retrieval effectiveness when reweighting word-based query language models based on the coverage of the concept-based translation determined with different term-by-term translation models. See Table 5.3 for legend.

(a) MeSH representation

Model	MAP 2004	2005	2006	2007
Text + KNN (MeSH)	**0.3868**	0.2429	**0.3736**	**0.2916**
M1 reweigh	0.3686 [2] -4.7%	0.2428 -0.0%	0.3400 [2] -9.0%	0.2228 [3] -23.6%
PTT reweigh	0.3608 [2] -6.7%	**0.2436** +0.3%	0.3171 [2] -15.1%	0.2147 [3] -26.4%
PMI reweigh	0.3697 -4.4%	0.2425 -0.1%	0.3463 [2] -7.3%	0.2243 [3] -23.1%
THES reweigh	0.3699 [2] -4.4%	0.2399 -1.2%	0.3509 [2] -6.1%	0.2273 [3] -22.1%

(b) UMLS$_{++}$ representation

Model	MAP 2004	2005	2006	2007
Text + KNN (UMLS$_{++}$)	**0.3929**	0.2285	0.4048	0.2981
M1 reweigh	0.3821 -2.7%	0.2300 +0.7%	0.4217 +4.2%	0.3045 +2.2%
PTT reweigh	0.3833 -2.5%	0.2320 +1.5%	0.4215 +4.1%	0.3025 +1.5%
PMI reweigh	0.3835 -2.4%	0.2341 +2.4%	0.4237 +4.7%	0.3051 +2.4%
STATTHES reweigh	0.3824 -2.7%	**0.2345** +2.6%	**0**.4226 +4.4%	0.3029 +1.6%
THES reweigh	0.3807 -3.1%	0.2332 +2.1%	**0.4262** +5.3%	**0.3070** +3.0%

Table 5.10: Average number of equivalence classes created (#e) and average number of concept terms (#c) grouped with word terms as a result of structuring the query using different translation models.

(a) MeSH representation

	2004		2005		2006		2007	
Model	#e	#c	#e	#c	#e	#c	#e	#c
M1 structure	2.2	4.7	2.4	4.1	2.4	3.7	1.7	3.4
PTT structure	1.3	2.1	1.9	2.9	1.8	2.5	1.1	1.8
PMI structure	1.8	3.4	2.0	3.1	1.8	2.6	1.2	2.3
THES structure	2.4	6.1	2.3	4.3	2.4	4.5	1.9	4.3

(b) UMLS$_{++}$ representation

	2004		2005		2006		2007	
Model	#e	#c	#e	#c	#e	#c	#e	#c
M1 structure	3.1	5.7	3.2	5.1	3.1	4.7	2.1	3.4
PTT structure	2.9	4.7	3.1	5.0	3.0	4.7	1.9	2.9
MI structure	2.8	5.1	2.8	4.9	2.7	4.3	2.0	3.4
STATTHES structure	3.1	6.2	3.0	5.5	3.0	4.5	2.1	3.6
THES structure	2.7	5.8	2.5	5.0	2.3	3.9	1.9	3.8

significant improvements (up to 6.4% in MAP). The decline in performance can to some extent be attributed to a difference in granularity of the word terms which have been grouped with more specific or too general concept terms. For instance, the UMLS$_{++}$ concept [nicotinic acetylcholine receptor location] is treated as a synonym of the word 'nicotin'. In other cases, clearly incorrect equivalence classes were formed. For example, the UMLS$_{++}$ concept [Device breakage] is grouped with the word 'break' in the context of 'DNA breaks'. In this case, the translation through feedback introduced these errors; by mapping these errors to original query words and treating them as equivalent, the impact of the erroneous translation was further emphasised. Improvements were observed when the words and concepts in the same equivalence class were clearly linked and were defined at the same granularity level.

5.7 Discussion

In this chapter we proposed and investigated monolingual biomedical information retrieval from a cross-lingual perspective. We distinguished a text-based and a concept-based language and proposed to view the integration of terminological resources in biomedical IR as a combination of translating and matching in either or both representations. We hypothesised that methods and techniques for traditional CLIR could also be beneficial for the effectiveness of monolingual biomedical IR. For brevity, we refer to this CLIR-enhanced monolingual biomedical IR as "biomedical CLIR".

Table 5.11: Retrieval effectiveness when structuring the word-based and concept-based query language model using different term-by-term translation models. See Table 5.3 for legend.

(a) MeSH representation

	MAP			
Model	2004	2005	2006	2007
Text + KNN (MeSH)	**0.3868**	0.2429	**0.3736**	**0.2916**
M1 structure	0.3737 -3.4%	0.2463 [2] +1.4%	0.3595 -3.8%	0.2308 [3] -20.9%
PTT structure	0.3788 -2.1%	**0.2501** +3.0%	0.3557 -4.8%	0.2253 [3] -22.7%
PMI structure	0.3765 -2.7%	0.2466 +1.5%	0.3574 -4.3%	0.2291 [3] -21.5%
THES structure	0.3732 -3.5%	0.2421 -0.3%	0.3391 -9.2%	0.2269 [3] -22.2%

(b) UMLS$_{++}$ representation

	MAP			
Model	2004	2005	2006	2007
Text + KNN (UMLS$_{++}$)	**0.3929**	0.2285	0.4048	**0.2981**
M1 structure	0.3782 -3.7%	0.2372 +3.8%	0.4234 +4.6%	0.2951 -1.0%
PTT structure	0.3845 -2.1%	**0.2431** [2] +6.4%	0.4224 +4.3%	0.2965 -0.5%
PMI structure	0.3852 -2.0%	0.2327 +1.9%	**0.4240** +4.7%	0.2929 -1.7%
STATTHES structure	0.3850 -2.0%	0.2367 [2] +3.6%	0.4215 +4.1%	0.2912 -2.3%
THES structure	0.3770 -4.1%	0.2361 +3.3%	0.4263 +5.3%	0.2978 -0.1%

Translation models for biomedical CLIR

Translation models are required to allow for translation between languages or representation types and subsequent cross-lingual matching. We asked the following question.

RQ3.1: *How can we build translation models for biomedical CLIR?*

Analogous to translation models used for traditional CLIR, we identified three types of translation models for biomedical CLIR: 1) translation based on a comparable corpus of documents in both a text and concept-based representation; 2) translation based on term-by-term translation models trained on a comparable corpus of documents; and 3) translation based on a thesaurus.

We investigated six different implementations of these types of translation models. KNN (type 1) uses a collection of comparable documents to determine a translation on pseudo-feedback (relevance models): the translation of a text-based query is based on concepts occurring in closely associated documents. Three translation models of type 2 were investigated: based on IBM Model 1 (M1), Pointwise Mutual Information (PMI) and Parsimonious Term Translation (PTT). These models use a comparable corpus of documents to estimate word-to-concept and concept-to-word translation probabilities. A major difference between M1 and the other two models is that M1 attempts to align the words and concepts in a comparable document. PMI and PTT only rely on document co-occurrence of words and concepts. The last two models estimate word-to-concept and concept-to-word translation probabilities on using entries in a thesaurus (type 3). THES

only uses information in the thesaurus itself. STATTHES also takes into account how frequently words are used to refer to concepts based on a tagged document collection.

A major difference between KNN and the other translation models is the amount of context taken into account during translation. KNN translates the query as a whole (that is, all query terms at the same time), whereas the other translation models can only translate single words or concepts. KNN can therefore take more query context into account than the other models. Given the ambiguity of biomedical terminology (discussed in chapter 2), one would expect the term-by-term translation models to perform poorly in comparison to KNN.

We therefore formulated the following research question.

RQ3.3: *How does context affect translation quality for biomedical CLIR?*

We investigated this effect using two CLIR retrieval models based on term-by-term translation: retrieval based on query translation and retrieval based on document translation. During query translation, a textual query was translated to a concept-based representation which was matched to the concept-based representation of the documents. During document translation, the concept-based document translation was translated to a textual representation which was matched to the textual query representation. Experiments with these retrieval models also provided insight into the quality of the investigated translation models.

The query translation experiments illustrated the impact of using a limited context (a single word) for translation. Especially, the translation to MeSH suffered from this lack of context: the term-by-term translation models (M1, PMI, PTT, and THES) performed significantly worse than the translation obtained from KNN. From these results we conclude that for accurate translation to MeSH a larger query context (than single query words) should be taken into account.

Surprisingly, the word-to-concept query translation to a UMLS$_{++}$ representation performed relatively well in comparison to the translation based on pseudo-feedback. In some cases increased performance was observed because the query translation results in a well-balanced query: each word contributes equally to the translation of the query as a whole. The translation based on KNN can suffer from query drift, resulting in an unbalanced concept-based query. Especially for queries which require multiple aspects to be present, such a balanced query was important. From these results we conclude that either the problem of ambiguity for biomedical IR is not that large after all, since a translation based on a very limited context is almost as effective as a translation taking more context into account. Or, that the translation based on pseudo-feedback fails to effectively benefit from the additional context taken into account during translation.

The retrieval effectiveness of using the translation models for document translation were disappointing in comparison to those using query translation. The primary explanation was the difference in granularity between word and concept representations. We will discuss this difference in more detail later.

The query and document translation experiments in subsection 5.6.1 demonstrated the usefulness of translation models trained on a comparable corpus and a statistical thesaurus over the use of a translation model based on the thesaurus alone. Incorporating the co-occurrence of words and concepts in a comparable corpus proved to be useful for building translation models for biomedical CLIR.

The experiments also provided insight into the limitations of the used translation models. The translation model based on IBM model 1 (and the translation model based on the statistical thesaurus) outperformed the two other translation model based on a comparable corpus (PMI and PTT). This illustrates the limitations of PMI and PTT which are solely based on co-occurrence of words and concepts in documents. In contrast, IBM model 1 also determines discriminative translations between words and concepts by determining the most likely alignment between them.

Improving word-based biomedical IR

In chapter 4, we concluded that retrieval based solely on a concept-based representation could not outperform word-based retrieval but that the representation could be used to improve word-based retrieval.

For biomedical CLIR, we asked the following question.

RQ3.2: *How effective are these translation models for improving word-based retrieval?*

The query translation experiments illustrated that, despite of the limited context taken into account, word-to-concept translation models can be useful for improving the retrieval effectiveness of word-based retrieval. Combining document translation with word-based retrieval did, however, only lead to very small improvements in retrieval effectiveness to word-based retrieval.

In comparison to the other translation models, the query translation based on pseudo-feedback (KNN) performed well. This can be partially explained by the larger context taken into account for translation, but even more because of its expansion effect: not only precise translations of the original query are returned, but also related concepts. Drawbacks of the approach are, however, that the obtained concept-based representation is particularly large (up to 50 concepts), may contain noise and may overemphasise particular query aspects. As a result, the KNN approach can suffer from query drift.

We therefore hypothesised that the KNN translation can be improved by combining it with the other translation models. We formulated the following questions.

RQ3.4: *Can translation for biomedical CLIR be improved by combining translation models?*

RQ3.5: *Can translation models be used to prevent query drift?*

To investigate whether this was possible we proposed three retrieval models in which term-by-term translation models were combined with pseudo-feedback translation (KNN) to improve word-based retrieval. The experiments with combining translation models (for pruning, reweighting and structuring queries) demonstrated that, similar to traditional CLIR, biomedical CLIR can benefit from combining multiple translation resources.

The goal of the pruning experiments was to remove superfluous and noisy concepts from the concept translation based on pseudo-feedback using concept-to-word translation models. For the MeSH-based representation, this approach was shown to be detrimental for its effectiveness as an individual representation for retrieval. Pruning also decreased its value to enhance word-based retrieval. From these results we conclude that a MeSH-based representation can be primarily used as a recall enhancing device. For a MeSH-based representation to be effective, however, many terms indirectly related to the information

need are required. The retrieval effectiveness of a UMLS$_{++}$-based query also slightly dropped as a result of pruning. However, when combined with a text-based representation this strongly reduced (between 61.5 and 91.5% fewer concepts) representation could still improve word-based retrieval. From these results we conclude that the UMLS$_{++}$ representation can be used to precisely represent information needs and can be used as a precision enhancing device.

The goal of the reweighting experiments was to "balance" the combined text and concept-based query: we hypothesised that query drift could be prevented by emphasising query words which were not covered by the concept-based representation. We determined this coverage using the term-by-term translation models. For MeSH, reweighting resulted in deteriorated retrieval effectiveness. We can explain these deteriorated results by the fact that the MeSH-based document representation is not exhaustive enough. Despite the fact that MeSH query terms according to the translation model covered the query words, lowering the weight of these words resulted in a loss of recall. For UMLS$_{++}$, small improvements in retrieval effectiveness were observed. We conclude that UMLS$_{++}$ terms can be effectively used to cover words from the original information need.

The goal of the structuring experiments was to prevent query drift by grouping words and concepts covering the same aspect of the query. For both MeSH and UMLS$_{++}$ small improvements in retrieval effectiveness were observed. Structuring errors, and differences in granularity of grouped words and concepts turned out to hurt retrieval effectiveness.

Future work

A major issue we encountered during these experiments was the difference in granularity between the translated words and concepts. Especially for document translation this demonstrated to strongly influence retrieval effectiveness: specific concepts were translated to general words, resulting in specific concepts being inappropriately important when matching documents to these words. This issue is strongly related to the translation unit chosen in our translation models: single words are difficult to translate to single concepts and vice versa. The experimental results showed, however, that even such simple translation models can be used to enhance monolingual word-based biomedical IR. An interesting direction for future work is to enhance these translation models with more sophisticated word-based translation units, such as phrases and word combinations in a short window of text. An iterative algorithm such as IBM model 1 can then be used to learn discriminative translation models between concepts and such a *text*-based representation.

5.8 Chapter summary

In this chapter we proposed and investigated a cross-lingual framework for biomedical IR. In this framework, we distinguished between a word and concept-based representation language. We modelled the integration of a concept-based representation in monolingual biomedical IR as a translation and matching process. We hypothesised that such an approach to the integration of a concept-based representation in biomedical IR could benefit from methods and techniques used in established CLIR.

Analogous to translation models used for traditional CLIR, we identified three types of translation models for biomedical CLIR: 1) translation based on a comparable corpus

of documents in both a text and concept-based representation; 2) translation based on term-by-term translation models trained on a comparable corpus of documents; and 3) translation based on a thesaurus.

We used these translation models in a number of different cross-lingual retrieval models. The first two, based on query and document translation, were intended to compare the quality of word-to-concept and concept-to-word translation models. Despite the limited context taken into account during translation, word-to-concept translation could still be used to improve word-based retrieval. Translation based on pseudo-feedback using a comparable corpus in both a word and concept-based representation (again) proved to perform best. In the other three retrieval models we investigated, we evaluated whether, similar to traditional CLIR, translation between text and concepts could be improved by combining translation models. Despite the simplicity of the term-by-term translation models, the results showed that a combination of translation models could improve retrieval effectiveness when combined with a word-based representation.

We conclude that the proposed cross-lingual framework offers a transparent view on the integration of a concept-based representation for monolingual biomedical IR. Based on the promising results with relatively simple translation and retrieval models, we have high expectations of more sophisticated translation and retrieval models.

Chapter 6

Summary and Conclusions

In this thesis, we investigated how to cope with the challenges for biomedical information retrieval caused by inconsistent, complex, and ambiguous terminology. Handling these challenges relieves end-users of biomedical IR systems of the burden of precisely and exhaustively describing their information needs in complex queries. Moreover, automated biomedical knowledge discovery applications may benefit from retrieval systems in which these challenges are automatically handled. We investigated how to make word-based IR more robust, how biomedical IR could benefit from a concept-based representation, and we proposed a framework for the integration of concept-based representation in a transparent manner.

In this last chapter, we will summarise the work in this thesis and reflect on the research themes identified in the introduction. In section 6.2, we will indicate directions for future research.

6.1 Research themes

In the following subsections we will discuss the three research themes identified in chapter 1.

6.1.1 RT1: Robust word-based retrieval

The first research theme we addressed was robust word-based retrieval. Effective retrieval models commonly use a word-based representation for retrieval. Choices in text preprocessing determine how these representations are obtained and what index vocabulary is used for representing documents and information needs. Dealing with the many spelling variations is a challenge for word-based biomedical IR. The way in which these variations are handled, was expected to influence retrieval effectiveness. We posed the following research question:

> **RQ1:** *How can the effectiveness of word-based biomedical information retrieval be improved using document preprocessing heuristics?*

In chapter 3, after an investigation of the characteristics of biomedical text, we investigated different document preprocessing heuristics to obtain word-based representations for biomedical IR. This investigation included stop-word removal, stemming, different

approaches to breakpoint normalisation, and n-gramming. Stop-word removal turned out to be primarily useful for improving the retrieval effectiveness of verbose information needs; removing words from shorter, manual queries only led to minor changes in retrieval effectiveness. Stemming, that is conflating words to a root form, turned out to be useful for improving retrieval effectiveness. We ascribe this increase to the many biomedical concepts which are referred to both in nouns and in conjugated variants of verbs. Breakpoint normalisation was intended to effectively handle the many compound terms encountered in biomedical text, by automatically determining word parts in compound terms and normalising these parts to index terms. Breakpoint normalisation was confirmed to strongly affect retrieval performance. Converting biomedical compound terms into multiple, overlapping index terms (as a result a single piece of text can be tokenised into multiple index terms) turned out to be particularly effective. This normalisation was observed to be of more importance for citation retrieval than for the retrieval of full-text journal articles.

Character n-gramming, a preprocessing technique frequently used for retrieval in languages without explicit word boundaries, performed poorly for biomedical IR.

Based on these experiments, a combination of document preprocessing heuristics was chosen to obtain word-based representations for biomedical IR. This method was used in the remainder of this thesis.

6.1.2 RT2: Concept-based retrieval

In chapter 4, the topic of *concept-based* representations for biomedical IR was introduced and investigated. Theoretically, a concept-based representation has the added value of being capable of representing information in a normalised, unambiguous fashion. In the ideal case, such a representation deals with the challenges of complex multi-word terms, synonymy, and ambiguous terminology. In practice, however, such a concept-based representation is also limited, because, for example, it is incomplete, or defines concepts at the incorrect level of granularity. Our research question was therefore as follows.

> **RQ2:** *What is the added value of a concept-based representation based on terminological resources for biomedical IR?*

Two concept-based representation vocabularies were investigated. Firstly, the Medical Subject Headings thesaurus (MeSH), a manually maintained vocabulary actively used to manually index biomedical documents. Secondly, the Unified Medical Language System (UMLS) metathesaurus, a large vocabulary database with biomedical and health related concepts, extended with a number of gene and protein dictionaries (referred to as UMLS$_{++}$).

We compared different classification systems to automatically obtain concept-based document and query representations. We proposed two classification methods based on statistical language models, one based on K-Nearest Neighbours (*KNN*) and one based on Concept Language Models (*CLM*). KNN classifies text based on similar, pre-classified documents. The method based on CLM classifies text by ranking language models which have been built for each concept. The systems were compared to a number of out-of-the-box classification systems.

The value of automatic document classification

For a selection of classification systems we carried out a document classification experiment using the MeSH representation vocabulary. In this experiment, we investigated to what extent the classification systems could reproduce manually created concept-based document representations. The proposed KNN system performed surprisingly well in comparison to the out-of-the-box systems: on average 4.5 out of the top 10 suggested MeSH concepts corresponded to manual classification. Manual analysis indicated that many highly ranked concepts which did not correspond to manual classification (between 34 and 58%) were in fact relevant to the documents. The results illustrate the improved exhaustiveness of automatic classification over manual classification.

The value of concept representations for queries

In a query classification experiment, we investigated the usefulness of a concept-based representation for retrieval. The investigated classification systems showed strongly varying performance in effectively mapping a text-based query to a concept-based representation for retrieval. Retrieval based on only concepts was demonstrated to be significantly less effective than word-based retrieval. This deteriorated performance could be explained by 1) errors in the classification process, in particular erroneous classification of specific concepts; 2) limited concept vocabularies: some information needs could not be accurately represented in terms of the concept vocabulary; 3) limited exhaustiveness of the concept-based document representations.

Retrieval based on a combination word-based and automatically obtained concept-based query representation did significantly improve word-only retrieval. Despite these limitations (and depending on the classification method used), the combination of a word-based and automatically obtained concept-based query representation significantly improved word-only retrieval. Small and mostly insignificant improvements in early precision (up to 7.4%) were observed. Larger and significant improvements were measured in terms of mean average precision (up to 9.9%), indicating a recall-enhancing effect of the concept representations.

In an artificial setting, we compared the optimal retrieval performance which could be obtained with word-based and concept-based representations. Contrary to our intuition, on average a single word-based query performed better than a single concept-based representation, even when the best concept term precisely represented part of the information need.

In general, we conclude that in practice a concept-based representation is very limited in expressiveness in comparison to a word-based representation. On its own, it cannot completely and precisely represent information needs. However, when combined with a text-based representation it can bring significant improvements to retrieval. Obtaining a concept-based representation through pseudo-relevance feedback (KNN) was shown to be especially effective.

The value of language models for predicting concept relatedness

In a final experiment, we investigated to what extent the relatedness between pairs of concepts as indicated by human judgements could be automatically reproduced. Results

on a small test set indicated that a method based on comparing concept language models performed particularly well in comparison to systems based on taxonomy structure, information content and (document) association. It was noted, however, that future work is necessary to make the relatedness measures useful for IR.

6.1.3 RT3: A framework for concept-based retrieval

As a final research theme of this thesis, we investigated the challenge of incorporating a concept-based representation into biomedical IR from a more fundamental perspective. Our underlying research question was as follows.

> **RQ3:** *Is it possible to cast the integration of knowledge from terminological resources in biomedical IR into a retrieval framework?*

In chapter 5, we suggested that monolingual biomedical IR should be viewed as a cross-lingual information retrieval (CLIR) problem. We distinguished between a text-based and concept-based language and viewed the integration of terminological resources in biomedical IR as a combination of translation and matching in either or both representations. Such a cross-lingual perspective gives the opportunity of adopting a large set of established CLIR methods and techniques for this domain. We hypothesised that monolingual word-based biomedical IR could benefit from translation and retrieval models available in conventional CLIR. (For brevity, we refer to this CLIR-enhanced monolingual biomedical IR as "biomedical CLIR".)

We identified three types of translation models for biomedical CLIR, analogous to translation models for conventional CLIR: 1) translation based on a comparable corpus of documents in both a text and concept-based representation; 2) translation based on term-by-term translation models trained on a comparable corpus of documents; and 3) translation based on a thesaurus. We investigated six different implementations of these types of translation models. The implementations varied in the way in which the comparable corpus and thesaurus were used for training translation probabilities. Moreover, they varied in the amount of context they took into account during translation.

These translation models were compared to and used in a number of retrieval models. The first two models were directly borrowed from conventional cross-lingual information retrieval and used term-by-term translation models to translate between the two languages. The second set of three retrieval models were driven by the hypothesis that translation can be improved by combining multiple translation models. In particular, we used the term-by-term translation models to improve translation based on a comparable corpus. The term-by-term translation models were used to prune, structure and reweigh the text and concept-based query.

The importance of context for translation

Experiments with these retrieval models indicated the importance of context for translation: term-by-term translation models, which take no context into account, were compared to a translation model which takes more query context into account by translating the complete query in a single translation. Context turned out to be especially important when no precise equivalent concept representation was available. The word-by-word query translation to

MeSH concepts was significantly less useful than a single translation of the whole query. In contrast, word-by-word translation to a more precise representation vocabulary, UMLS$_{++}$, performed only slightly worse than a single translation. Surprisingly, despite the limited query context taken into account by term-by-term translation models, word-based retrieval could still be improved when combined with these automatically translated representations.

The value of comparable corpora for translation

The experiments demonstrated the usefulness of translation models trained on a comparable corpus and a statistical thesaurus over the use of a translation model based on the thesaurus alone. Incorporating the co-occurrence of words and concepts in a comparable corpus proved to be useful for building more effective translation models for biomedical CLIR.

The value of alignment for translation

The experiments also illustrated important differences between the translation models. The term-by-term translation models trained on a parallel corpus, which were based on an (estimated) alignment between words and concepts performed better than translation models based on word and concept (document) co-occurrence. This alignment was shown to be important to create discriminative translation models: the alignment allows the system to distinguish concepts which frequently co-occur in the same documents.

The value of combination of translation models

The experiments with combining translation models (for pruning, reweighting and structuring queries) demonstrated that, similar to conventional CLIR, biomedical CLIR can benefit from combining multiple translation resources.

Pruning the translation obtained directly from a parallel corpus using a term-by-term translation model proved to be useful for the UMLS$_{++}$ concept representation. Up to an average of 91.5% of the translated concepts could be pruned without losing its added value of improving word-based retrieval effectiveness. Applying the same pruning operation to the translated MeSH representation turned out to hurt its complementing value for word-based retrieval: word-based retrieval could still be improved, but the improvements were smaller. The results showed a clear difference between the added value of the MeSH and UMLS$_{++}$ concept vocabularies for word-based retrieval. The UMLS$_{++}$ representation vocabulary can be used as a precision enhancing device: its added value comes from precisely covering the information need. The MeSH representation vocabulary can be used for enhancing recall of word-based retrieval: the representation of an information need requires many carefully weighted MeSH concepts related to the information need to serve this purpose.

The goal of the reweighting experiments was to “balance” the combined text and concept-based query: we hypothesised that query drift could be prevented by emphasising query words which were not covered by the concept-based representation. We determined this coverage using the term-by-term translation models. For MeSH, reweighting resulted in deteriorated retrieval effectiveness. We can explain these deteriorated results by the fact that the MeSH-based document representation is not exhaustive. Despite the fact that according to the translation model the MeSH query concepts covered the query words, lowering the weight of these words resulted in a loss of recall. For UMLS$_{++}$, small improvements in

retrieval effectiveness were observed. We conclude that UMLS$_{++}$ terms can be effectively used to cover words from the original information need.

The goal of the structuring experiments was to prevent query drift by grouping words and concepts covering the same aspect of the query. For both MeSH and UMLS$_{++}$, small improvements in retrieval effectiveness were observed. Structuring errors, and differences in granularity of grouped words and concepts sometimes appeared to be detrimental to retrieval effectiveness.

Conclusions about the framework

We conclude that the proposed cross-lingual framework offers a transparent view on the integration of a concept-based representation for monolingual biomedical IR. We noticed that the more advanced models based on an estimated alignment of a comparable or a parallel corpus performed better than the translation models based on only word and concept co-occurrence. Based on the promising results with these relatively basic models and given the fact that multi-word terms are frequently used in biomedical text, we have high expectations for more sophisticated translation and retrieval models.

6.2 Directions for future work

Based on the work in this thesis, we identified three directions for future work.

Concepts for communication

One major added value of having a concept-based representation language as opposed to a word-based representation language that we have not addressed in this thesis, is the possibility of using a concept-based representation language for communication between the retrieval system and the user (Fonseca et al., 2005). We expect that concepts can be particularly useful for summarising retrieved documents and suggesting better or additional query terms. How such feedback should be presented, whether it is appreciated by the user and how subsequent concept-based feedback from the user should be incorporated in the retrieval system are open questions. Approaches to conventional interactive CLIR (Oard et al., 2008) and interactive query expansion (Joho et al., 2004) can be used as a starting point for extending the CLIR-enhanced monolingual biomedical IR proposed in this thesis.

Sophisticated CLIR translation models

The investigated translation models in chapter 5 were quite basic: primarily term-to-term translation models were investigated. Despite their simplicity, they can already be used to improve word-based retrieval. Especially translating concepts from and to single words is rather unsophisticated in a domain where multi-word terms are so frequently used. We expect more sophisticated translation models between concept and *text*-based representations to be even more beneficial for word-based retrieval. For instance, by building translation models which translate between concepts and unordered word combinations rather than between concepts and single words.

Extending concept-based CLIR to other domains

A third direction future work is the investigation of the concept-based CLIR framework outside the biomedical domain. Other domains with an additional representation vocabulary might also benefit from a similar translation and matching approach. For example, the retrieval of information on intellectual property might benefit from a similar integration of the International Patent Classification language, a controlled vocabulary used to index patents. Or, retrieval of news might benefit from International Press Telecommunications Council (IPTC) NewsCodes taxonomy.

Appendix A

TREC Genomics topic sets

This appendix lists the topic sets used in the experiments reported in chapter 3 to chapter 5.

A.1 TREC Genomics 2004 topic set

ID	Section	Description
1	Title	Ferroportin-1 in humans
	Need	Find articles about Ferroportin-1, an iron transporter, in humans.
	Context	Ferroportin1 (also known as SLC40A1; Ferroportin 1; FPN1; HFE4; IREG1; Iron regulated gene 1; Iron-regulated transporter 1; MTP1; SLC11A3; and Solute carrier family 11 (proton-coupled divalent metal ion transporters), member 3) may play a role in iron transport.
2	Title	Generating transgenic mice
	Need	Find protocols for generating transgenic mice.
	Context	Determine protocols to generate transgenic mice having a single copy of the gene of interest at a specific location.
3	Title	Time course for gene expression in mouse kidney
	Need	What is the time course of gene expression in the murine developing kidney?
	Context	Relevant articles describe genes involved in kidney development.
4	Title	Gene expression profiles for kidney in mice
	Need	What mouse genes are specific to the kidney?
	Context	What genes are expressed only in the mouse kidney and not in other tissues?
5	Title	Protocols for isolating cell nuclei
	Need	Articles are relevant if they describe methods for subcellular fractionation of nuclei.
	Context	Laboratory preparations can be enriched for certain kinds of proteins if the cellular compartment in which they reside is purified away from the rest of the cell contents.
6	Title	FancD2
	Need	Find articles about function of FancD2.
	Context	There are many genes involved in Fanconi Anemia and the downstream pathways of FancD2 in flies. The FancD2 is monoubiquitylated and there are 2 components of the FancD2 pathway. The researcher studies the FancD2 pathway in flies.
7	Title	DNA repair and oxidative stress
	Need	Find correlation between DNA repair pathways and oxidative stress.
	Context	Researcher is interested in how oxidative stress effects DNA repair.
8	Title	Correlation between DNA repair pathways and skin cancer
	Need	Genes and proteins (pathways) common to DNA repair, oxidative diseases, skin-carcinogenesis, and UV-carcinogenesis.
	Context	Are there genes and mechanisms that are utilized by more than one of these fields? A relevant article mentions a gene or pathway, DNA repair, and one or more oxidative or cancerous diseases.
9	Title	mutY
	Need	Find articles about the function of mutY in humans.
	Context	mutY is particularly challenging, because it is also known as hMYH. This is further complicated by the fact that myoglobin genes are also typically located in search results.

ID	Section	Description
10	Title	NEIL1
	Need	Find articles about the role of NEIL1 in repair of DNA.
	Context	Interested in role that NEIL1 plays in DNA repair.
11	Title	Carcinogenesis and hairless mice
	Need	Find articles regarding carcinogenesis induced in hairless mice.
	Context	Researching genes and proteins (pathways) common to DNA repair, oxidative diseases, skin-carcinogenesis, and UV-carcinogenesis.
12	Title	Genes regulated by Smad4
	Need	Find articles describing genes that are regulated by the signal transducing molecule Smad4.
	Context	Project is to characterize Smad4 knockout mouse in skin (specifically skin) to establish signaling network. Identify all Smad4 targets to compare gene expression patterns of the knockout mouse to the normal mouse.
13	Title	Role of TGFB in angiogenesis in skin
	Need	Documents regarding the role of TGFB in angiogenesis in skin with respect to homeostasis and development.
	Context	TGFB plays a crucial role in regulating angiogenesis, a biological process that occurs during development and homeostasis, as well as during inflammatory perturbation.
14	Title	Expression or Regulation of TGFB in HNSCC cancers
	Need	Documents regarding TGFB expression or regulation in HNSCC cancers.
	Context	The laboratory wants to identify components of the the TGFB signaling pathway in HNSCC, and determine new targets to study HNSCC.
15	Title	ATPase and apoptosis
	Need	Find information on role of ATPases in apoptosis
	Context	The laboratory wants to know more about the role of ATPases in apoptosis.
16	Title	AAA proteins
	Need	How do AAA proteins mediate interaction with lipids or DNA and what is their functional impact?
	Context	A relevant document is one that discusses protein interactions interactions involving members of the AAA protein family that can help to determine their functional importance.
17	Title	DO1 antibody
	Need	Determine binding affinity of anti-p53 monoclonal antibody DO1.
	Context	One aspect of determining how an antibody works is to determine its binding affinity. A relevant document is one which discusses the binding affinity of DO1.
18	Title	Gis4
	Need	Properties of Gis4 with respect to cell cycle and/or metabolism.
	Context	It is possible that Gis4 plays a role between cell cycle and yeast carbon pathways and that there is a link between cell cycle and metabolism. A relevant document is one that supports or refutes this hypothesis with regard to the properties of Gis4 in one or both processes.
19	Title	Comparison of Promoters of GAL1 and SUC1
	Need	What similarities and differences exist between the upstream promoter regions of GAL1 and SUC1? Are there co-repressors or co-activators? If so, are they regulated by SNF1?
	Context	Gis4 may play a role between the cell cycle and yeast carbon pathways. SNF1 is an upstream kinase of Gis 4.
20	Title	Substrate modification by ubiquitin
	Need	Which biological processes are regulated by having constituent proteins modified by covalent attachment to ubiquitin or ubiquitin-like proteins?
	Context	Ubiquitin and ubiquitin-like proteins have important roles in controlling cell division, signal transduction, embryonic development, endocytic trafficking, and the immune response.
21	Title	Role of p63 and p73 in relation to DNA damage
	Need	Do p63 and p73 cause cell cycle arrest or apoptosis related to DNA damage?
	Context	DNA damage may cause cell cycle arrest or apoptosis. p63 and p73 may play a role in mediating these sequelae of DNA damage.
22	Title	Relative response of p53 family members to agents causing single-stranded versus double-stranded DNA breaks
	Need	Does p53 respond differently to different DNA-damaging agents? Do they respond differently to single-strand versus double-strand breaks?
	Context	DNA damage may cause cell cycle arrest or apoptosis. p53 plays a role in mediating these sequelae of DNA damage.
23	Title	Saccharomyces cerevisiae proteins involved in ubiquitin system
	Need	Which Saccharomyces cerevisiae proteins are involved in the ubiquitin proteolytic pathway?
	Context	The researcher identified a protein in another yeast species and wants to compare it to the same one in Saccharomyces cerevisiae.

ID	Section	Description
24	Title	Mouse peptidoglycan recognition proteins (PGRP)
	Need	Find all reports describing mouse peptidoglycan recognition proteins (PGRP).
	Context	A research group is preparing a manuscript about four poorly characterized mouse PGRP genes. Their findings include new information about gene regulation. They report longer DNA and protein sequences than those found in GenBank, and sub-cellular location discrepancies.
25	Title	Cause of scleroderma
	Need	Identify studies that include genome-wide scans and microarray analysis in the investigation of scleroderma.
	Context	New information about experiments and genes involved in scleroderma.
26	Title	Function of BUB2/BFA1 in the process of cytokinesis
	Need	Retrieval of information regarding the role of BUB2 and BFA1 in cytokinesis in yeast.
	Context	Information gathering for the purpose of supplementing the information from a local protocol.
27	Title	Role of autophagy in apoptosis
	Need	Experiments establishing positive or negative interconnection between autophagy and apoptosis.
	Context	New information about experiments and genes involved in autophagic cell death.
28	Title	Proteases that function in both apoptosis and autophagy cell death
	Need	Studies that investigate similarities in morphological changes among apoptosis and autophagy processes.
	Context	Collection of information regarding the potential relationship between apoptosis and autophagy.
29	Title	Phenotypes of gyrA mutations
	Need	Documents containing the sequences and phenotypes of E. coli gyrA mutations.
	Context	The laboratory has isolated some gyrA mutations in E. coli. They want to compare their mutant gyrA with the wild-type and other mutant sequences.
30	Title	Regulatory targets of the Nkx gene family members
	Need	Documents identifying genes regulated by Nkx gene family members.
	Context	The laboratory needs markers to follow Nkx family-member expression and activity.
31	Title	TOR signaling in neurofibromatosis
	Need	Reports that provide possible links between neurofibromatosis and TOR signaling.
	Context	TOR is a serine-threonine kinase in a pathway involved in the control of cell growth and proliferation, and it is the target of the signaling inhibitor rapamycin.
32	Title	Xenograft animal models of tumorogenesis
	Need	Find reports that describe xenograft models of human cancers.
	Context	A xenograft animal model of cancer is one in which foreign tumor tissue is grafted into animals, usually rodents, providing a means to test various compounds for their ability to slow or halt tumor growth.
33	Title	Mice, mutant strains, and Histoplasmosis
	Need	Identify research on mutant mouse strains and factors which increase susceptibility to infection by Histoplasma capsulatum.
	Context	The ultimate goal of this initial research study, is to identify mouse genes that will influence the outcome of blood borne pathogen infections.
34	Title	Gene products of Cryptococcus important to fungal survival
	Need	Articles reporting experiments allowing annotation of gene products of Cryptococcus.
	Context	Information needed to contribute to the development of a standardized annotated database of Cryptococcus neoformans genome.
35	Title	WD40 repeat-containing proteins
	Need	What is the function of proteins containing WD40 repeats?
	Context	Need to understand the variety of functions that involve this domain.
36	Title	RAB3A
	Need	Background information on RAB3A.
	Context	Further information about a gene is needed after it is identified through a gene expression profile. The genes are related to synaptic plasticity in learning and memory.
37	Title	PAM
	Need	What research is being done on peptide amidating enzyme, PAM?
	Context	Need to put specific PAM research in the context of other researchers work.
38	Title	Risk factors for stroke
	Need	Information concerning genetic loci that are associated with increased risk of stroke, such as apolipoprotein E4 or factor V mutations.
	Context	Candidate gene testing within a large Scottish case-control study of genetic risk factors for stroke. Future research includes investigations into other ethnically distinct populations.

ID	Section	Description
39	Title	Hypertension
	Need	Identify genes as potential genetic risk factors candidates for causing hypertension.
	Context	A relevant document is one which discusses genes that could be considered as candidates to test in a randomized controlled trial which studies the genetic risk factors for stroke.
40	Title	Antigens expressed by lung epithelial cells
	Need	To identify the antigens expressed by lung epithelial cells and the antibodies available.
	Context	Information gathering to design assays to determine the nature of donor cells in tissues of chimaeric animals.
41	Title	Mutations in the Cystic Fibrosis conductance regulator gene
	Need	What phenotypes have been described resulting from mutations in the Cystic Fibrosis conductance regulator gene?
	Context	Comparing protein mutations detected utilizing mass spectrometry.
42	Title	Genes altered by chromosome translocations
	Need	What genes show altered behavior due to chromosomal rearrangements?
	Context	Information is required on the disruption of functions from genomic DNA rearrangements.
43	Title	Sleeping Beauty
	Need	Studies of Sleeping Beauty transposons.
	Context	A relevant document is one that discusses studies on Sleeping Beauty. Interviewee's group studies a related element and want to know what others are doing in a similar field.
44	Title	Proteins involved in the nerve growth factor pathway
	Need	Create a list of all the nerve growth factor pathway proteins.
	Context	Need to identify genes that are most likely to be involved in the nerve growth factor pathway.
45	Title	Mental Health Wellness-1
	Need	What genetic loci, such as Mental Health Wellness 1 (MWH1) are implicated in mental health?
	Context	Want to identify genes involved in mental disorders.
46	Title	RSK2
	Need	What human biological processes is RSK2 known to be involved in?
	Context	After being identified via microarrays, the biological processes the genes are involved in needs to be discovered.
47	Title	Human gene BCL-2 antagonists and inhibitors
	Need	Research the human gene BCL-2 to determine if there are antagonists and inhibitors inside of a cell.
	Context	Early research goals included learning more about BCL2-interacting molecules, which facilitated identifying new inhibitors during preliminary testing.
48	Title	Human homologues of C. elegans UNC genes
	Need	What is the focus of studies involving the members of the human UNC gene family?
	Context	The interviewee wished to determine the interests and focus of a fellow scientist that was investigating similar topics to their own.
49	Title	Glyphosate tolerance gene sequence
	Need	Find reports and glyphosate tolerance gene sequences in the literature.
	Context	A DNA sequence isolated in the laboratory is often sequenced only partially, until enough sequence is generated to identify the gene. In these situations, the rest of the sequence is inferred from matching clones in the public domain. When there is difficulty in the laboratory manipulating the DNA segment using sequence-dependent methods, the laboratory isolate must be re-examined.
50	Title	Low temperature protein expression in E. coli
	Need	Find research on improving protein expressions at low temperature in Escherichia coli bacteria.
	Context	The researcher is not satisfied with the yield of expressing a protein in E. coli when grown at low temperature and is searching for a better solution. The researcher is willing to try a different organism and/or method.

A.2 TREC Genomics 2005 topic set

ID	Query
100	Describe the procedure or methods for how to "open up" a cell through a process called "electroporation."
101	Describe the procedure or methods for exact reactions that take place when you do glutathione S-transferase (GST) cleavage during affinity chromatography.
102	Describe the procedure or methods for different quantities of different components to use when pouring a gel to make it more or less porous.
103	Describe the procedure or methods for green fluorescent protein (GFP) tagged proteins to do experiments with tagged proteins.
104	Describe the procedure or methods for how to do a microsomal budding assay, i.e., budding of vesicles from microsomes in vitro.
105	Describe the procedure or methods for purification of rat IgM.
106	Describe the procedure or methods for chromatin IP (Immuno Precipitations) to isolate proteins that are bound to DNA in order to precipitate the proteins out of the DNA.
107	Describe the procedure or methods for normalization procedures that are used for microarray data.
108	Describe the procedure or methods for identifying in vivo protein-protein interactions in time and space in the living cell.
109	Describe the procedure or methods for fluorogenic 5'-nuclease assay.
110	Provide information about the role of the gene Interferon-beta in the disease Multiple Sclerosis.
111	Provide information about the role of the gene PRNP in the disease Mad Cow Disease.
112	Provide information about the role of the gene IDE gene in the disease Alzheimer's Disease.
113	Provide information about the role of the gene MMS2 in the disease Cancer.
114	Provide information about the role of the gene APC (adenomatous polyposis coli) in the disease Colon Cancer.
115	Provide information about the role of the gene Nurr-77 in the disease Parkinson's Disease.
116	Provide information about the role of the gene Insulin receptor gene in the disease Cancer.
117	Provide information about the role of the gene Apolipoprotein E (ApoE) in the disease Alzheimer's Disease.
118	Provide information about the role of the gene Transforming growth factor-beta1 (TGF-beta1) in the disease Cerebral Amyloid Angiopathy (CAA).
119	Provide information about the role of the gene GSTM1 in the disease Breast Cancer.
120	Provide information on the role of the gene nucleoside diphosphate kinase (NM23) in the process of tumor progression.
121	Provide information on the role of the gene BARD1 in the process of BRCA1 regulation.
122	Provide information on the role of the gene APC (adenomatous polyposis coli) in the process of actin assembly.
123	Provide information on the role of the gene COP2 in the process of transport of CFTR out of the endoplasmic reticulum.
124	Provide information on the role of the gene casein kinase II in the process of ribosome assembly.
125	Provide information on the role of the gene Nurr-77 in the process of preventing auto-immunity by deleting reactive T-cells before they migrate to the spleen or the lymph nodes.
126	Provide information on the role of the gene P53 in the process of apoptosis.
127	Provide information on the role of the gene alpha7 nicotinic receptor subunit gene in the process of ethanol metabolism.
128	Provide information on the role of the gene gamma-aminobutyric acid receptors (GABABRs) in the process of inhibitory synaptic transmission.
129	Provide information on the role of the gene Interferon-beta in the process of viral entry into host cell.
130	Provide information about the genes BRCA1 regulation of ubiquitin in cancer.
131	Provide information about the genes L1 and L2 in the HPV11 virus in the role of L2 in the viral capsid.
132	Provide information about the genes APC (adenomatous polyposis coli) and wnt in colon cancer.
133	Provide information about the genes phospholipase A2 (PLA2) and SAR1 in Endoplasmic reticulum transport (i.e. vesicle budding from the ER).
134	Provide information about the genes CFTR and Sec61 in degradation of CFTRwhich leads to cystic fibrosis.
135	Provide information about the genes Bop and Pes in cell growth.
136	Provide information about the genes alpha7 nicotinic receptor gene and ApoE gene in the neurotoxic effects of ethanol.
137	Provide information about the genes Insulin-like GF and insulin receptor gene in the function in skin.
138	Provide information about the genes HNF4 and COUP-TF I in the suppression in the function of the liver.
139	Provide information about the genes Ret and GDNF in kidney development.
140	Provide information about BRCA1 185delAG mutation and its/their role in ovarian cancer.
141	Provide information about Huntingtin mutations and its/their role in Huntington's Disease.
142	Provide information about Sonic hedgehog mutations and its/their role in developmental disorders.
143	Provide information about Mutations of NM23 and its/their impact on tracheal development.
144	Provide information about Mutations in metazoan Pes and its/their effect on cell growth.
145	Provide information about Mutations of hypocretin receptor 2 and its/their role in narcolepsy.
146	Provide information about Mutations of presenilin-1 gene and its/their biological impact in Alzheimer's disease.
147	Provide information about Mutations of alpha7 nAChR gene and its/their biological impact in alcoholism.
148	Provide information about Mutation of familial hemiplegic migraine type 1 (FHM1) and its/their neuronal Ca2+ influx in hippocampal neurons.
149	Provide information about Mutations of the alpha 4-GABAA receptor and its/their impact on behavior.

A.3 TREC Genomics 2006 topic set

ID	Query
160	What is the role of PrnP in mad cow disease?
161	What is the role of IDE in Alzheimer's disease
162	What is the role of MMS2 in cancer?
163	What is the role of APC (adenomatous polyposis coli) in colon cancer?
164	What is the role of Nurr-77 in Parkinson's disease?
165	How do Cathepsin D (CTSD) and apolipoprotein E (ApoE) interactions contribute to Alzheimer's disease?
166	What is the role of Transforming growth factor-beta1 (TGF-beta1) in cerebral amyloid angiopathy (CAA)?
167	How does nucleoside diphosphate kinase (NM23) contribute to tumor progression?
168	How does BARD1 regulate BRCA1 activity?
169	How does APC (adenomatous polyposis coli) protein affect actin assembly
170	How does COP2 contribute to CFTR export from the endoplasmic reticulum?
171	How does Nurr-77 delete T cells before they migrate to the spleen or lymph nodes and how does this impact autoimmunity?
172	How does p53 affect apoptosis?
173	How do alpha7 nicotinic receptor subunits affect ethanol metabolism?
174	How does BRCA1 ubiquitinating activity contribute to cancer?
175	How does L2 interact with L1 to form HPV11 viral capsids?
176	How does Sec61-mediated CFTR degradation contribute to cystic fibrosis?
177	How do Bop-Pes interactions affect cell growth?
178	How do interactions between insulin-like GFs and the insulin receptor affect skin biology?
179	How do interactions between HNF4 and COUP-TF1 suppress liver function?
180	How do Ret-GDNF interactions affect liver development?
181	How do mutations in the Huntingtin gene affect Huntington's disease?
182	How do mutations in Sonic Hedgehog genes affect developmental disorders?
183	How do mutations in the NM23 gene affect tracheal development?
184	How do mutations in the Pes gene affect cell growth?
185	How do mutations in the hypocretin receptor 2 gene affect narcolepsy?
186	How do mutations in the Presenilin-1 gene affect Alzheimer's disease?
187	How do mutations in familial hemiplegic migraine type 1 (FHM1) gene affect calcium ion influx in hippocampal neurons?

A.4 TREC Genomics 2007 topic set

ID	Query
200	What serum [PROTEINS] change expression in association with high disease activity in lupus?
201	What [MUTATIONS] in the Raf gene are associated with cancer?
202	What [DRUGS] are associated with lysosomal abnormalities in the nervous system?
203	What [CELL OR TISSUE TYPES] express receptor binding sites for vasoactive intestinal peptide (VIP) on their cell surface?
204	What nervous system [CELL OR TISSUE TYPES] synthesize neurosteroids in the brain?
205	What [SIGNS OR SYMPTOMS] of anxiety disorder are related to coronary artery disease?
206	What [TOXICITIES] are associated with zoledronic acid?
207	What [TOXICITIES] are associated with etidronate?
208	What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to zoledronic acid?
209	What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to etidronate?
210	What [MOLECULAR FUNCTIONS] are attributed to glycan modification?
211	What [ANTIBODIES] have been used to detect protein PSD-95?
212	What [GENES] are involved in insect segmentation?
213	What [GENES] are involved in Drosophila neuroblast development?
214	What [GENES] are involved axon guidance in C.elegans?
215	What [PROTEINS] are involved in actin polymerization in smooth muscle?
216	What [GENES] regulate puberty in humans?
217	What [PROTEINS] in rats perform functions different from those of their human homologs?
218	What [GENES] are implicated in regulating alcohol preference?
219	In what [DISEASES] of brain development do centrosomal genes play a role?
220	What [PROTEINS] are involved in the activation or recognition mechanism for PmrD?
221	Which [PATHWAYS] are mediated by CD44?
222	What [MOLECULAR FUNCTIONS] is LITAF involved in?
223	Which anaerobic bacterial [STRAINS] are resistant to Vancomycin?
224	What [GENES] are involved in the melanogenesis of human lung cancers?
225	What [BIOLOGICAL SUBSTANCES] induce clpQ expression?
226	What [PROTEINS] make up the murine signal recognition particle?
227	What [GENES] are induced by LPS in diabetic mice?
228	What [GENES] when altered in the host genome improve solubility of heterologously expressed proteins?
229	What [SIGNS OR SYMPTOMS] are caused by human parvovirus infection?
230	What [PATHWAYS] are involved in Ewing's sarcoma?
231	What [TUMOR TYPES] are found in zebrafish?
232	What [DRUGS] inhibit HIV type 1 infection?
233	What viral [GENES] affect membrane fusion during HIV infection?
234	What [GENES] make up the NFkappaB signaling pathway?
235	Which [GENES] involved in NFkappaB signaling regulate iNOS?

Appendix B

Word-based Biomedical IR

B.1 Optimal smoothing values

Table B.1 lists the smoothing values (parameter λ in Equation 2.2) used for the experiments reported in chapter 3.

Table B.1: Document smoothing values which result in the highest mean average precision for that collection.

	Original queries					Manual queries			
	2004	2005	2006	2007		2004	2005	2006	2007
base	0.65	0.80	0.50	0.85	base	0.50	0.05	0.10	0.65
basestop	0.60	0.80	0.25	0.65	basestop	0.55	0.10	0.15	0.55
basestem	0.60	0.80	0.50	0.85	basestem	0.50	0.15	0.15	0.60
join1	0.70	0.90	0.45	0.85	join1	0.55	0.35	0.05	0.80
join2	0.70	0.90	0.45	0.85	join2	0.60	0.35	0.05	0.80
split1	0.75	0.90	0.55	0.85	split1	0.70	0.30	0.05	0.75
split3	0.55	0.80	0.55	0.85	split3	0.50	0.15	0.30	0.65
js1	0.80	0.90	0.55	0.90	js1	0.65	0.30	0.10	0.65
js2	0.80	0.90	0.50	0.85	js2	0.65	0.30	0.05	0.75
js3	0.70	0.65	0.25	0.85	js3	0.70	0.15	0.10	0.65
jse1	0.75	0.90	0.50	0.90	jse1	0.60	0.35	0.05	0.55
jse2	0.80	0.90	0.65	0.90	jse2	0.60	0.20	0.05	0.75
jse3	0.65	0.60	0.20	0.90	jse3	0.55	0.05	0.05	0.65
ngram4	0.35	0.30	0.05	0.45	ngram4	0.25	0.05	0.05	0.25
ngram5	0.30	0.40	0.05	0.35	ngram5	0.10	0.05	0.05	0.20
ngram6	0.25	0.35	0.05	0.15	ngram6	0.10	0.20	0.05	0.10
combined	0.60	0.70	0.05	0.65	combined	0.55	0.10	0.05	0.55

Appendix C

Concept-based biomedical IR

C.1 Example classifications

Table C.1 lists the output of the tested MeSH classification systems (section 4.4) when used to classify the title of a MEDLINE citation. The MEDLINE column shows the manual classification determined by human indexers.

C.2 Optimal cut-off values

A number of the tested systems for document classification (section 4.4) return a ranked list of concepts rather than a discrete set. For calculation of the macro and micro F-measure a discrete number of terms should be assigned to the input text. Hence, the ranked list of MeSH terms needs to be cut off after a particular number of terms. In a real-world scenario, this parameter can be based on a training set. Following the approach used by Lam et al. (1999), we set this cutoff to the value which gives the best performance. A single cutoff value is determined for each system. Using this approach we determine the upper bound of the system's performance, independent of the ability to train the right parameter for such a system.

Table C.2 and Table C.3 show the macro and micro F-measures, respectively, of the different systems at different cutoff levels. For example, when using only the title as input (Table C.2(a)), EAGL performs optimally when only the top 15 terms are taken into account, achieving an F-measure of 0.2413.

C.3 Annotations for the false positive analysis

Table C.4 lists the scale used for analysing false positives returned by the MeSH classification systems. Each label is illustrated with an example.

C.4 Fusion of word and concept-based retrieval

In chapters 4 and 5 word and concept-based retrieval were combined by linear interpolation of retrieval status values (the negated cross entropy between query and document language

models). We also investigated a number of alternative well-known fusion techniques to combine lists of retrieved documents including round robin fusion, CombMNZ, CombMax, and CombSum (Fox and Shaw, 1993).

Suppose we have n result sets (R_1 to R_n)[1], each consisting of m ranked documents (d_{i1} to d_{im}) and corresponding retrieval status values (rsv_{i1} to rsv_{im}). Each result set looks as follows.

$$R_i = (d_{i1}, rsv_{i1}), \ldots, (d_{im}, rsv_{im}) \tag{C.1}$$

The CombMNZ, CombMax and CombSum fusion methods normalise these retrieval status values to document scores between 1 (for the highest ranked document) and 0 (for the lowest ranked document) as follows.

$$s_{ij} = \frac{rsv_{ij} - rsv_{im}}{rsv_{i1} - rsv_{im}} \tag{C.2}$$

rsv_{i1} is the highest retrieval status value and rsv_{im} is the lowest retrieval status value for a single set of retrieved documents.

We will use the following notation to describe the fusion methods.

$$\begin{aligned} rsv(i,d) &\text{ The retrieval status value of document } d \text{ in } R_i \\ s(i,d) &\text{ The normalised score of document } d \text{ in } R_i \\ s_{\text{method}}(d) &\text{ The document score in the fused result list} \end{aligned} \tag{C.3}$$

The fusion methods are defined as follows.

Interpolation The interpolation method simply sums the weighted retrieval status values of different runs.

$$s_{\text{interpolation}}(d) = \sum_i w_i \times rsv(i,d) \tag{C.4}$$

w_i is the weight assigned to the ranked list of documents R_i. The sum of the weights should be equal to 1: $\sum_i w_i = 1$.

Round robin During round robin fusion, documents are merged according to rank in the ranked lists of documents R_1 to R_n. The merged result list is assembled by iterating over the ranked lists and adding the highest ranked document from this result list which was not yet in the merged list.

CombSum CombSum ranks documents according to their summed normalised score.

$$s_{\text{combSum}}(d) = \sum_i s(i,d) \tag{C.5}$$

CombMNZ CombMNZ multiplies the CombSum with the number of ranked lists which assign a non-zero score to the document.

$$s_{\text{combMNZ}}(d) = s_{\text{combSum}}(d) \times |\{i | s(i,d) > 0\}| \tag{C.6}$$

[1] In this case we only have 2 result sets: one from word-based retrieval and one from concept-based retrieval

CombMax CombMax uses the maximum normalised score assigned by one of the ranked lists as the fused document score.

$$s_{\text{combMax}}(d) = max_i s(i, d) \tag{C.7}$$

Table C.5 lists the mean average precision obtained using the different fusion methods. For the interpolation a weight of 0.5 was used for both the text and concept result lists. Round robin performed worst in fusing the results. Except from the 2006 and 2007 topic sets, retrieval effectiveness in terms of MAP became worse. The approaches based on normalised scores (CombSum, CombMNZ and CombMax) performed well for the UMLS$_{++}$ representation, but performed poorly for MeSH. We expect this was caused by the fact that retrieval performance based on only MeSH was poor in comparison to text-based retrieval. As a result, fusion based on normalised scores put too much emphasis on documents retrieved with MeSH concepts. Interpolation turned out to perform well across the two concept representations and the different topic sets. Based on these results was decided to use interpolation of unnormalised retrieval scores throughout this thesis.

C.5 Relatedness correlation plots

To illustrate how well the scores from the different relatedness measures investigated in section 4.7 agree with the judgements from the human annotators, the system's scores have been plotted against the annotators' judgements.

Figure C.1 and Figure C.2 show the correlation plots on the test set from Caviedes (test set 1). Figure C.3 and Figure C.4 show the plots for the test set from Pedersen (test set 2). Note that for test set 1 a judgment of 1 indicates a strong relatedness, whereas for test set 2 the same judgement indicates a weaker relatedness.

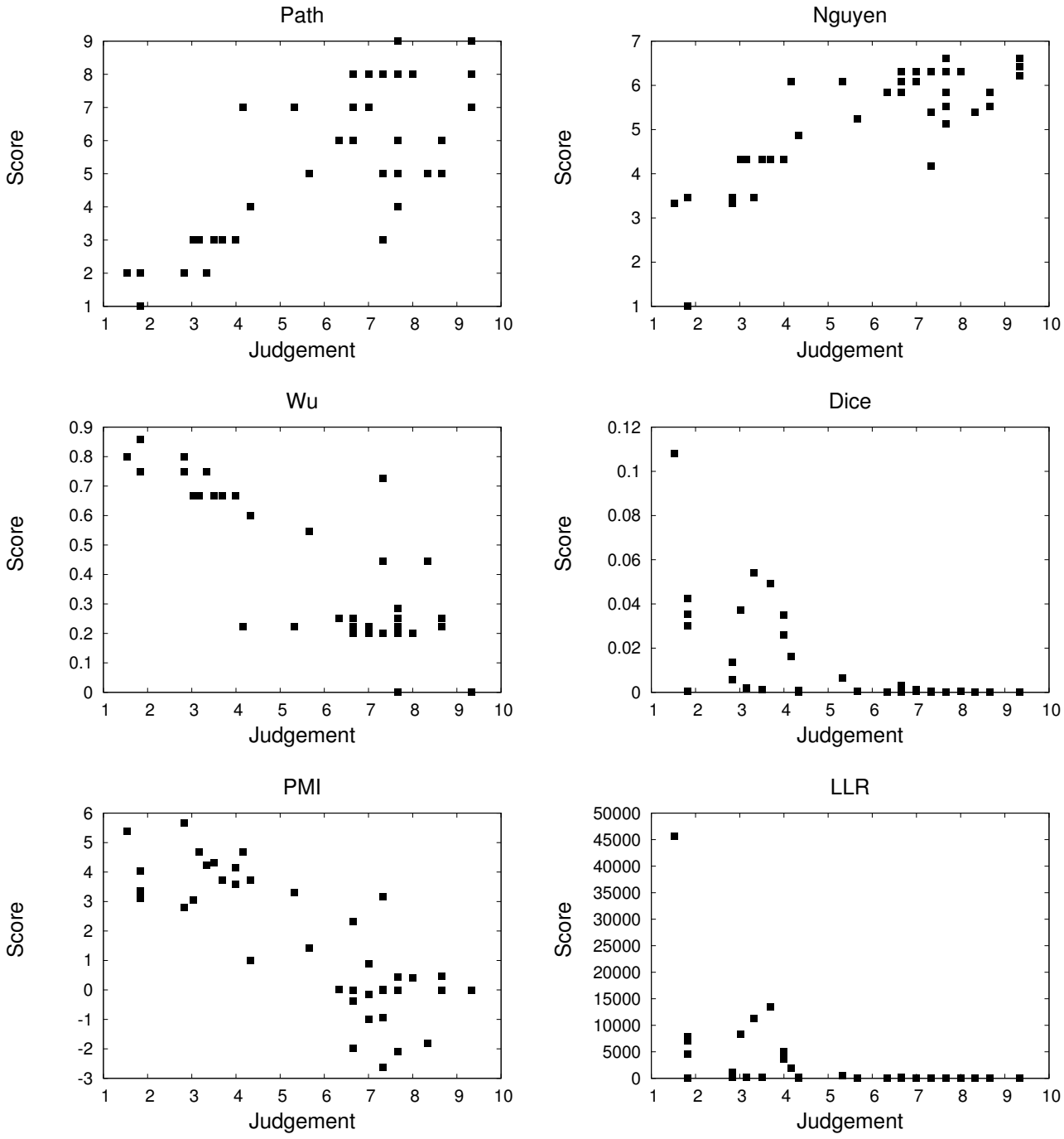


Figure C.1: Caviedes (test set 1): plots for Path, Nguyen, Wu, Dice, PMI, and LLR relatedness measures.

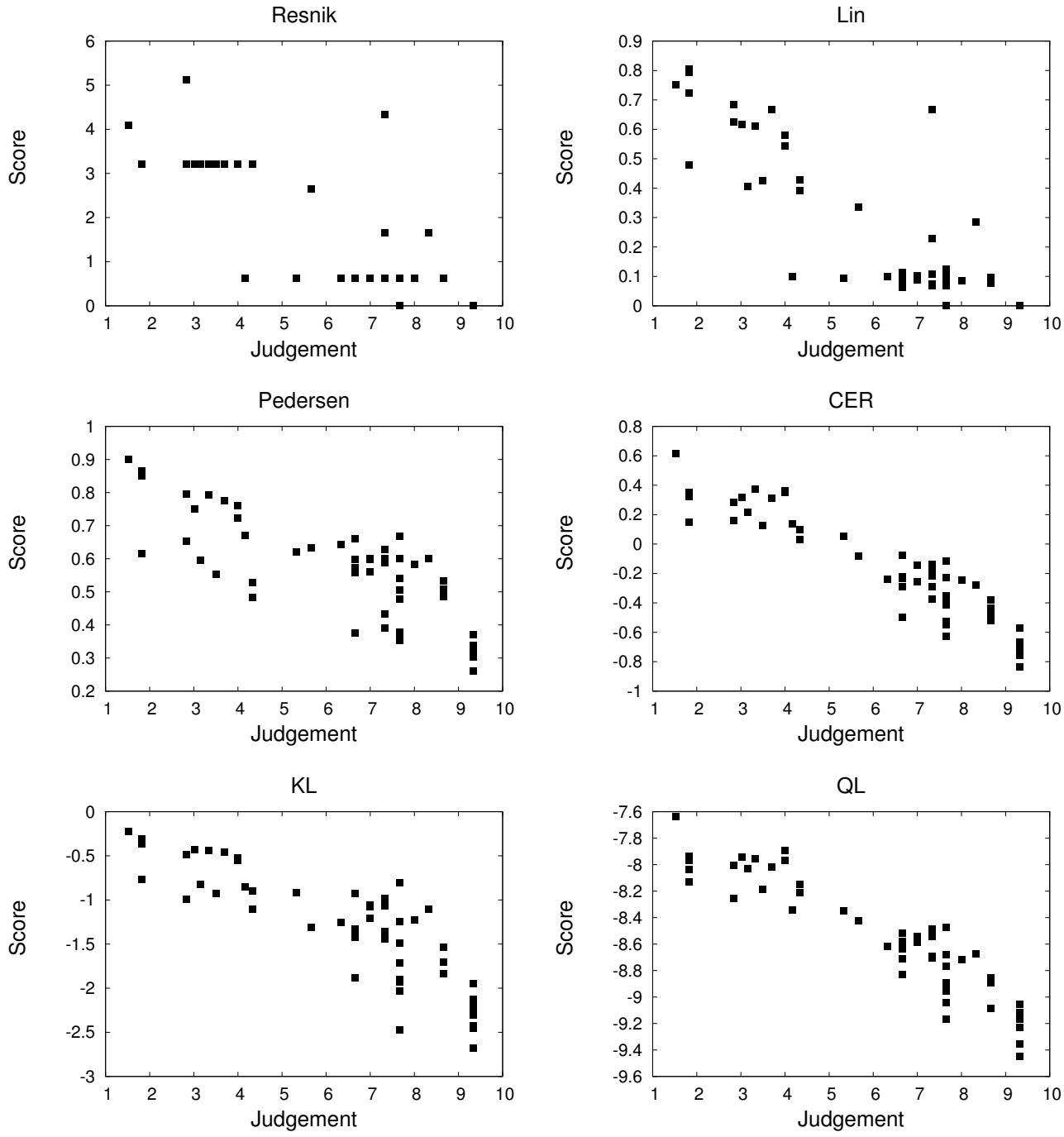


Figure C.2: Caviedes (test set 1): plots for Resnik, Lin, Pedersen, CER, KL and QL relatedness measures.

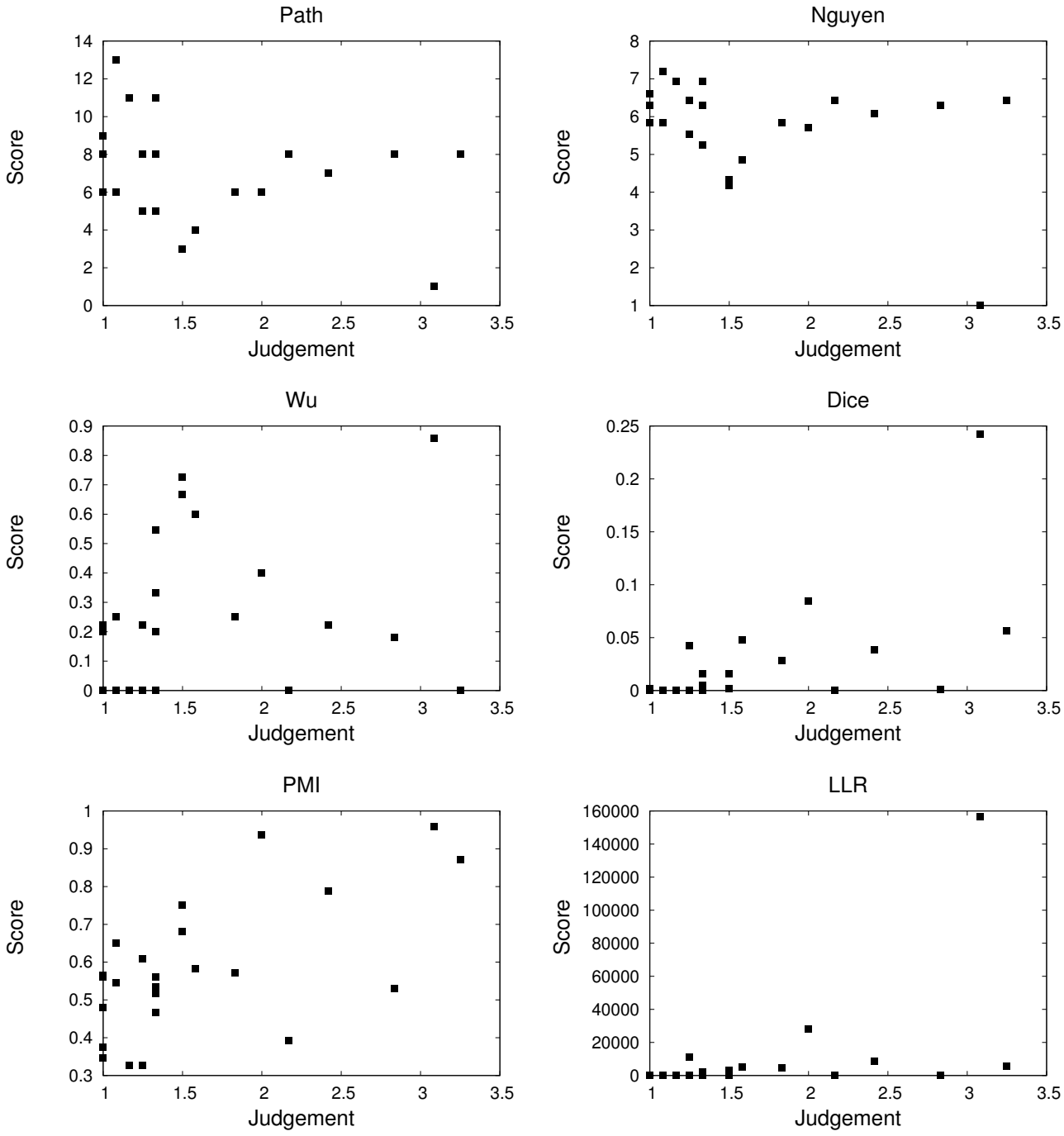


Figure C.3: Pedersen (test set 2): plots for Path, Nguyen, Wu, Dice, PMI, and LLR relatedness measures.

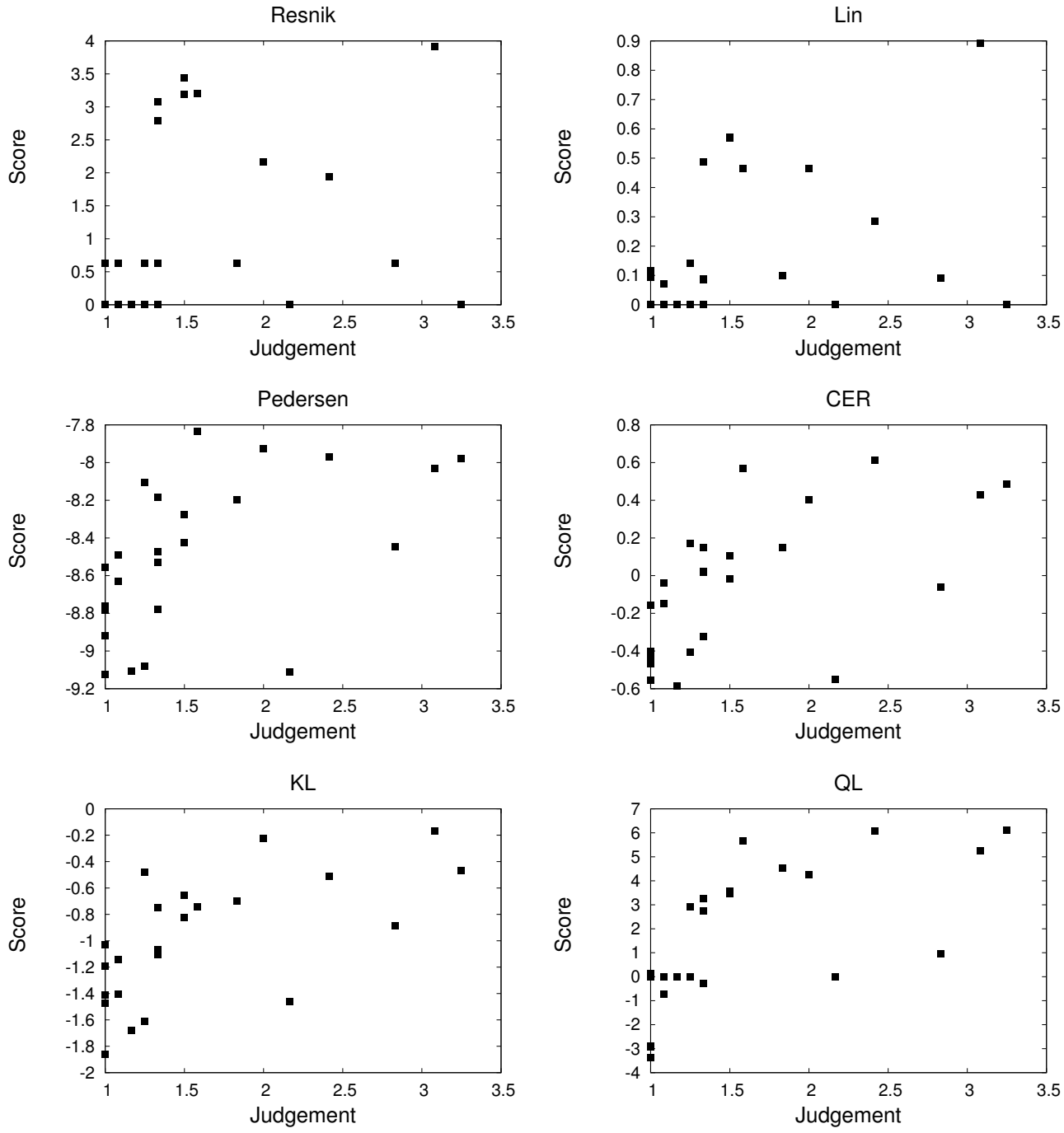


Figure C.4: Pedersen (test set 2): plots for Resnik, Lin, Pedersen, CER, KL and QL relatedness measures.

Table C.1: Classifications obtained from the different systems for the input text: "Reactive oxygen species and the regulation of cell death by the Bcl-2 gene family" [PMID 7599227]. The MEDLINE column shows the manual classifications used as a ground truth. Classifications marked in bold face agree with the manual classification.

MEDLINE (manual)	MetaMap	EAGL
[**Multigene Family**]	[Family]	[**Reactive Oxygen Species**]
[**Oncogenes**]	[Sensitivity and Specificity]	[Genes, bcl-2]
[**Proto-Oncogene Proteins**]	[**Cell Death**]	[**Cell Death**]
[**Proto-Oncogenes**]	[**Reactive Oxygen Species**]	[Social Control, Formal]
[**Cell Death**]	[Social Control, Formal]	[Oxygen]
[**Reactive Oxygen Species**]	[Death]	[Family]
[**Amino Acid Sequence**]	[Cells]	[Death]
[**Animals**]	[Genes, bcl-2]	[Genes]
[**Chromosomes, Human, Pair 14**]	[Oxygen]	[Reactive Nitrogen Species]
[**Chromosomes, Human, Pair 18**]	[Oxygen Inhalation Therapy]	[Oxygenators]
[**Homeostasis**]	[Genes]	[Species Specificity]
[**Humans**]		[Hybridization, Genetic]
[**Lymphoma**]		[Extinction, Biological]
[**Molecular Sequence Data**]		[Cells]
[**Translocation, Genetic**]		[Up-Regulation]
[**Apoptosis**]		[Genes, Regulator]
[**Sequence Homology, Amino Acid**]		[Binding Sites]
[**Proto-Oncogene Proteins c-bcl-2**]		[Reactive Inhibition]
		[Hyperemia]

CLM	KNN	MTI
[**Proto-Oncogene Proteins c-bcl-2**]	[**Proto-Oncogene Proteins c-bcl-2**]	[**Cell Death**]
[Genes, bcl-2]	[**Animals**]	[**Reactive Oxygen Species**]
[bcl-X Protein]	[**Apoptosis**]	[Genes, bcl-2]
[bcl-2-Associated X Protein]	[**Reactive Oxygen Species**]	[Genetics]
[bcl-2 Homologous Antagonist-Killer Protein]	[Mice]	
[BH3 Interacting Domain Death Agonist Protein]	[**Proto-Oncogene Proteins**]	
[bcl-Associated Death Protein]	[**Humans**]	
[Apoptosis Inducing Factor]	[**Cell Death**]	
[**Reactive Oxygen Species**]	[Gene Expression Regulation]	
[Reactive Nitrogen Species]	[Antioxidants]	
[Caspase 2]	[Mice, Transgenic]	
[**Apoptosis**]	[**Homeostasis**]	
[**Cell Death**]	[bcl-2-Associated X Protein]	
[Apoptosis Regulatory Proteins]	[Cell Transformation, Neoplastic]	
[**Proto-Oncogene Proteins**]	[Cell Line]	
[Receptors, Tumor Necrosis Factor, Member 25]	[Oxidation-Reduction]	
[Apoptotic Protease-Activating Factor 1]	[In Situ Nick-End Labeling]	
[Mitochondrial Membranes]	[bcl-X Protein]	
[Vitamin K 3]	[Tumor Cells, Cultured]	

Table C.2: F-measures at different cut-off values.

(a) Title used as input

Cut-off	MTI	MetaMap	EAGL	CLM	KNN
5	0.2572	0.1893	0.2115	0.2361	0.2552
10	**0.2663**	0.2280	0.2364	0.2937	0.3432
15	0.2663	0.2329	**0.2413**	0.3217	**0.3693**
20	0.2663	**0.2330**	0.2406	0.3299	0.3684
25	0.2663	0.2330	0.2377	**0.3326**	0.3586
30	0.2663	0.2330	0.2343	0.3324	0.3504
35	0.2663	0.2330	0.2324	0.3280	0.3429
40	0.2663	0.2330	0.2291	0.3212	0.3350
45	0.2663	0.2330	0.2256	0.3149	0.3264
50	0.2663	0.2330	0.2234	0.3091	0.3223
55	0.2663	0.2330	0.2234	0.3027	0.3168

(b) Title and abstract used as input

Cut-off	MTI	MetaMap	CLM	EAGL	KNN
5	0.2673	0.1669	0.2348	0.2304	0.2565
10	0.4102	0.2348	0.2990	0.2737	0.3615
15	0.4402	0.2723	0.3261	0.2859	0.3994
20	0.4471	0.2921	0.3407	0.2951	**0.4074**
25	**0.4503**	0.3049	**0.3429**	0.2973	0.3997
30	0.4498	0.3132	0.3418	**0.2987**	0.3903
35	0.4498	0.3169	0.3385	0.2974	0.3805
40	0.4498	0.3183	0.3318	0.2948	0.3708
45	0.4498	0.3184	0.3242	0.2917	0.3619
50	0.4498	**0.3187**	0.3190	0.2877	0.3554
55	0.4498	0.3187	0.3135	0.2877	0.3503

Table C.3: Micro F-measure at different cut-off values.

(a) Title used as input

Cut-off	MTI	MetaMap	EAGL	CLM	KNN
5	0.2759	0.2242	0.2412	0.2160	0.4300
10	**0.2859**	0.2622	**0.2588**	0.2617	**0.4758**
15	0.2859	0.2659	0.2542	0.2809	0.4565
20	0.2859	**0.2660**	0.2460	0.2855	0.4244
25	0.2859	0.2660	0.2356	**0.2877**	0.3949
30	0.2859	0.2660	0.2259	0.2868	0.3705
35	0.2859	0.2660	0.2178	0.2833	0.3485
40	0.2859	0.2660	0.2098	0.2775	0.3305
45	0.2859	0.2660	0.2015	0.2726	0.3159
50	0.2859	0.2660	0.1956	0.2677	0.3057
55	0.2859	0.2660	0.1956	0.2627	0.2971

(b) Title and abstract used as input

Cut-off	MTI	MetaMap	CLM	EAGL	KNN
5	0.2919	0.2072	0.2166	0.2484	0.4403
10	0.4073	0.2635	0.2672	0.2884	**0.4963**
15	0.4334	0.2843	0.2872	0.2964	0.4812
20	0.4402	0.2920	0.2972	**0.2977**	0.4510
25	**0.4415**	0.2961	**0.2982**	0.2920	0.4189
30	0.4410	**0.2968**	0.2969	0.2850	0.3916
35	0.4410	0.2962	0.2940	0.2770	0.3684
40	0.4410	0.2948	0.2881	0.2684	0.3490
45	0.4410	0.2939	0.2819	0.2601	0.3334
50	0.4410	0.2934	0.2771	0.2530	0.3217
55	0.4410	0.2934	0.2720	0.2530	0.3126

Label	Description	Example
“Strongly relevant”	A term that should be used for annotation	The MeSH term [Urinary Incontinence, Stress] (D014550) for the input text “Urinary stress incontinence” (PMID 3664398)
“Relevant”	An appropriate term but a better term is available	The MeSH term [Urination Disorders] [D014555] for the input text “Urinary stress incontinence” (PMID 3664398)
“Undecided”	1) no direct evidence available in the text to support the term, or 2) the term is too specific	The MeSH term [Influenza A Virus, H2N2 Subtype] [D053121] for the input text “Nonrandom association of parental genes in influenza A virus recombinants” (PMID 442543)
“Irrelevant”	when the complete MeSH term is irrelevant, but part of its phrase can be related to the text	The MeSH term [Child, Preschool] [D002675] (defined to be age 2-5) for the input text “Accidental infant death and stroller-prams.” (PMID 8709876, describing cases of a three-month-old boy and an eight-month-old boy)
“Incorrect”	A term that is clearly wrong	The MeSH term [Social Control, Formal] [D012926] for the input text “Reactive oxygen species and the regulation of cell death by the Bcl-2 gene family” (PMID 7599227)

Table C.4: Judgement scale description and examples used for analysing false positives.

Table C.5: Retrieval effectiveness when combining word-based and concept-based retrieval using different fusion models. [1], [2] and [3] indicate significant differences to the baseline (word-only) at confidence levels 0.05, 0.01 and 0.001 respectively, determined with a paired sign test. The highest value of each column is printed in boldface.

(a) MeSH representation

	MAP			
Model	2004	2005	2006	2007
Text	0.3576	0.2219	**0.3889**	0.2796
KNN (MeSH)	0.1889	0.1268	0.2518	0.1901
Interpolate	**0.3868** [2] +8.2%	**0.2429** [1] +9.5%	0.3736 -5.7%	**0.2916** +4.3%
Round robin	0.3095 [3] -13.4%	0.2007 [2] -9.5%	0.3667 [3] -5.7%	0.2638 [2] -5.6%
CombSum	0.3606 +0.9%	0.2260 +1.9%	0.3606 -7.3%	0.2865 +2.5%
CombMNZ	0.3577 +0.0%	0.2266 +2.2%	0.3534 [2] -9.1%	0.2797 +0.0%
CombMax	0.3127 [2] -12.6%	0.1965 [2] -11.5%	0.3883 [2] -0.1%	0.2758 -1.3%

(b) UMLS$_{++}$ representation

	MAP			
Model	2004	2005	2006	2007
Text	0.3576	0.2219	0.3889	0.2796
KNN (UMLS$_{++}$)	0.2799	0.1670	0.3535	0.2355
Interpolate	**0.3929** [2] +9.9%	0.2285 +3.0%	0.4048 +4.1%	**0.2981** +6.6%
Round robin	0.3499 -2.1%	0.2145 -3.3%	0.3999 +2.8%	0.2889 +3.4%
CombSum	0.3875 [2] +8.4%	**0.2290** +3.2%	0.4033 +3.7%	0.2910 +4.1%
CombMNZ	0.3828 +7.0%	0.2257 +1.7%	**0.4050** +4.1%	0.2935 +5.0%
CombMax	0.3646 +2.0%	0.2105 -5.1%	0.3984 +2.5%	0.2822 +1.0%

Appendix D

A Cross-Lingual Framework for Biomedical IR

D.1 Pruning examples

Figure D.1 and Figure D.2 illustrate the effect of pruning a concept representation obtained through pseudo-relevance feedback with a translation model based on PMI.

D.2 Reweighting examples

Table D.1 to Table D.3 illustrate the effect of reweighting words in the word-based query language model, based on the coverage of the words ($P_{cov}(w|\phi_Q)$) by the concept-based representation obtained through pseudo-feedback. The first column shows the original query term probabilities; the second column shows the coverage of these query terms by the concept-based representation and the third column indicates the probabilities in the updated query language model. The last column indicates which concepts covered the original word terms. For these examples, $P(w|c)$ was estimated using the naive thesaurus translation model (THES).

Table D.1: Reweighted query terms for topic 170 "How does COP2 contribute to CFTR export from the endoplasmic reticulum?". Reweighting resulted in an improvement of 15.3% in average precision.

Word	Original $P(w\|\theta_Q)$	Coverage	Updated $P(w\|\theta'_Q)$	Concepts
contribut	0.125	0	0.145	
cop2	0.125	0	0.145	
2	0.125	0	0.145	
endoplasm	0.125	0.104	0.128	[Endoplasmic Reticulum], [LMAN1]
export	0.125	0.462	0.072	[Export]
cop	0.125	0.014	0.142	[CARD16]
reticulum	0.125	0.412	0.080	[Reticulum], [Endoplasmic Reticulum], [LMAN1]
cftr	0.125	0.009	0.143	[CFTR]

Before pruning 0.121 [Animals], 0.106 [**Prions**], 0.065 [Humans], 0.036 [Slow Virus Diseases], 0.036 [Models, Molecular], 0.036 [Spectrum Analysis], 0.036 [Structure-Activity Relationship], 0.036 [Zoonoses], 0.028 [Mammals], 0.024 [**Encephalopathy, Bovine Spongiform**], 0.024 [**Cattle**], 0.023 [Mice], 0.019 [Molecular Sequence Data], 0.017 [Brain], 0.017 [Fungi], 0.017 [Fungal Proteins], 0.016 [**Prion Diseases**], 0.016 [**Creutzfeldt-Jakob Syndrome**], 0.016 [**PrPC Proteins**], 0.015 [Mutation], 0.013 [Sequence Alignment], 0.013 [Evolution, Molecular], 0.013 [Base Sequence], 0.013 [Phylogeny], 0.012 [Mice, Inbred C57BL], 0.011 [Point Mutation], 0.011 [Fetus], 0.011 [Organ Specificity], 0.011 [Tissue Extracts], 0.011 [Precipitin Tests], 0.011 [Protein Isoforms], 0.011 [Female], 0.011 [Viscera], 0.011 [Sheep], 0.011 [Immunoenzyme Techniques], 0.011 [Epitopes], 0.011 [Mice, Transgenic], 0.011 [Golgi Apparatus], 0.011 [Sequence Analysis, Protein], 0.011 [Neurons], 0.007 [Amino Acid Sequence], 0.007 [Amyloid], 0.007 [Protein Precursors], 0.006 [DNA Transposable Elements], 0.006 [Models, Genetic], 0.006 [Sequence Analysis, DNA], 0.006 [DNA Footprinting], 0.006 [Gene Order], 0.006 [Likelihood Functions], 0.006 [Computational Biology].

After pruning 0.528 [Prions], 0.119 [Cattle], 0.119 [Encephalopathy, Bovine Spongiform], 0.078 [Creutzfeldt-Jakob Syndrome], 0.078 [Prion Diseases], 0.078 [PrPC Proteins].

Figure D.1: The result of pruning the KNN concept representation of the query "What is the role of PrnP in mad cow disease?" (topic 160) using the PMI translation model. Concepts which were not pruned are displayed in bold face.

Table D.2: Reweighted word query terms for topic 169 "How does APC (adenomatous polyposis coli) protein affect actin assembly?". Reweighting resulted in an improvement of 19.5% in average precision.

Word	Original $P(w\|\theta_Q)$	Coverage	Updated $P(w\|\theta'_Q)$	Concepts
assembl	0.143	0.070	0.156	[Assembly (construction)]
actin	0.143	0	0.169	
protein	0.143	0	0.169	
apc	0.143	0.110	0.149	[adenomatous polyposis coli], [Apc2], [MAPRE1], [MAPRE2]
polyposi	0.143	0.463	0.084	[adenomatous polyposis coli], [Multiple polyps], [Apc2], [MAPRE1]
coli	0.143	0.149	0.142	[adenomatous polyposis coli], [Apc2], [MAPRE1]
adenomat	0.143	0.209	0.131	[adenomatous polyposis coli], [Apc2], [MAPRE1]

Before pruning 0.074 [Humans], 0.065 [**Huntington Disease**], 0.058 [**Nuclear Proteins**], 0.057 [**Animals**], 0.050 [**Nerve Tissue Proteins**], 0.045 [**Mice**], 0.043 [Peptides], 0.037 [**Trinucleotide Repeat Expansion**], 0.029 [**Oligonucleotide Array Sequence Analysis**], 0.028 [Male], 0.027 [Corpus Striatum], 0.026 [**Molecular Sequence Data**], 0.023 [**Gene Expression Profiling**], 0.022 [Neurons], 0.022 [Disease Models, Animal], 0.021 [Brain], 0.021 [Female], 0.018 [**RNA, Messenger**], 0.016 [**Trinucleotide Repeats**], 0.015 [Subcellular Fractions], 0.015 [Antibodies], 0.015 [**Amino Acid Sequence**], 0.014 [**Cell Nucleus**], 0.014 [**Cytoplasm**], 0.013 [**Mice, Transgenic**], 0.013 [**Conserved Sequence**], 0.013 [**Blotting, Western**], 0.013 [Rats], 0.013 [**Mutation**], 0.011 [Immunoenzyme Techniques], 0.011 [**Blotting, Northern**], 0.011 [**Gene Expression Regulation**], 0.011 [**Base Sequence**], 0.008 [**Mice, Knockout**], 0.008 [**DNA Repair**], 0.008 [**MutS Homolog 2 Protein**], 0.008 [**Proto-Oncogene Proteins**], 0.008 [**DNA-Binding Proteins**], 0.008 [Microscopy, Confocal], 0.008 [Immunohistochemistry], 0.008 [**Cell Line**], 0.008 [**Cell Nucleolus**], 0.008 [Fluorescent Antibody Technique, Indirect], 0.008 [Rabbits], 0.008 [Aged], 0.008 [**Age of Onset**], 0.008 [Middle Aged], 0.008 [Adult], 0.008 [Peptide Fragments], 0.008 [Nerve Degeneration]

After pruning 0.106 [Huntington Disease], 0.094 [Nuclear Proteins], 0.092 [Animals], 0.081 [Nerve Tissue Proteins], 0.072 [Mice], 0.060 [Trinucleotide Repeat Expansion], 0.047 [Oligonucleotide Array Sequence Analysis], 0.042 [Molecular Sequence Data], 0.038 [Gene Expression Profiling], 0.029 [RNA, Messenger], 0.026 [Trinucleotide Repeats], 0.024 [Amino Acid Sequence], 0.023 [Cell Nucleus], 0.023 [Cytoplasm], 0.022 [Mice, Transgenic], 0.022 [Conserved Sequence], 0.021 [Blotting, Western], 0.021 [Mutation], 0.018 [Gene Expression Regulation], 0.018 [Base Sequence], 0.018 [Blotting, Northern], 0.013 [MutS Homolog 2 Protein], 0.013 [Mice, Knockout], 0.013 [DNA Repair], 0.013 [Proto-Oncogene Proteins], 0.013 [DNA-Binding Proteins], 0.013 [Cell Line], 0.013 [Cell Nucleolus], 0.012 [Age of Onset].

Figure D.2: The result of pruning the KNN concept representation of the query "How do mutations in the Huntingtin gene affect Huntington's disease?" (topic 181) using the PMI translation model. Concepts which were not pruned are displayed in bold face.

Table D.3: Reweighting word query terms for topic 162 "What is the role of MMS2 in cancer?". Reweighting resulted in an deterioration of 20.2% in average precision.

Word	Original $P(w\|\theta_Q)$	Coverage	Updated $P(w\|\theta'_Q)$	Concepts
2	0.25	0	0.296	
mm	0.25	0.810	0.148	[UBE2V2]
mms2	0.25	0.190	0.261	[UBE2V2]
cancer	0.25	0	0.296	

0.075 interact, 0.075 es, 0.075 **#wsyn(0.842 growth, 0.079 [Growth], 0.078 [Tissue Growth])**, 0.075 p, 0.075 **#wsyn(0.699 bop, 0.301 [BOP1])**, 0.075 cell, 0.075 bopp, 0.064 [Cells], 0.026 [Genes], 0.020 [equus asinus asinus], 0.020 [Mutant], 0.016 [Plant Leaves], 0.014 [Pancreas], 0.013 [Anabolism], 0.013 [Process], 0.013 [Processing (action)], 0.013 [Analysis], 0.011 [binding], 0.010 [Ribosomes], 0.010 [Homo sapiens], 0.009 [FIG], 0.009 [Figs], 0.009 [Protein Domain], 0.009 [Rattus], 0.008 [FLC1], 0.008 [Plants], 0.008 [receptor], 0.008 [Tissue membrane], 0.008 [Typing Classification], 0.007 [development aspects], 0.007 [DICOM Study], 0.007 [Clinical Trials], 0.007 [Scientific Study], 0.007 [BOP2], 0.007 [Affinity], 0.007 [Malignant Neoplasms], 0.007 [primary malignant neoplasm], 0.007 [premature cardiac complex], 0.007 [P53], 0.007 [Wild Type], 0.007 [Hamsters], 0.006 [p53], 0.006 [protein expression], 0.006 [mRNA Expression], 0.006 [Saccharomyces cerevisiae], 0.006 [TP53], 0.005 [Neoplasms], 0.005 [Mutation Abnormality], 0.005 [Organ], 0.005 [Laboratory culture], 0.005 [Culture], 0.005 [Forms], 0.005 [Arabidopsis], 0.005 [WDR12].

Figure D.3: Topic 177 ("How Bop-Pes interactions affect cell growth?") after structuring using a term-by-term translation model. As a result, average precision increased from 0.287 to 0.526.

0.085 **#wsyn(0.899 gene, 0.101 [Genes])**, 0.085 mutat, 0.085 2, 0.085 **#wsyn(0.956 hypocretin, 0.044 [hcrt])**, 0.085 **#wsyn(0.896 narcolepsi, 0.104 [Narcolepsy])**, 0.085 receptor, 0.057 [HCRT], 0.032 [Asleep], 0.032 [Sleep], 0.023 [receptor], 0.022 [Rattus], 0.015 [equus asinus asinus], 0.015 [Neurons], 0.014 [Homo sapiens], 0.013 [Discharge (release)], 0.013 [Release (procedure)], 0.012 [Brain], 0.012 [Cells], 0.011 [process of secretion], 0.010 [Patients], 0.009 [Clinical Trials], 0.009 [Scientific Study], 0.009 [DICOM Study], 0.009 [Regulation], 0.009 [Pituitary Gland], 0.009 [thyroid stimulating hormone measurement], 0.008 [Hypothalamic structure], 0.007 [regulatory], 0.007 [Adrenal Glands], 0.007 [Analysis], 0.007 [NQO1], 0.007 [Plasma], 0.006 [regulation of biological process], 0.006 [Male gender], 0.006 [Stimulation (motivation)], 0.006 [Disease], 0.006 [Feeding patient], 0.006 [PTGDS], 0.006 [human study subject], 0.005 [Others], 0.005 [CD200R1], 0.005 [Study Subject], 0.005 [Time], 0.005 [Female], 0.005 [Behavior], 0.005 [Activation action], 0.005 [Activities], 0.005 [FIG], 0.005 [Figs], 0.004 [mRNA Expression], 0.004 [Cell Nucleus], 0.004 [Specimen], 0.004 [Sampling].

Figure D.4: Topic 185 ("How do mutations in the hypocretin receptor 2 gene affect narcolepsy?") after structuring using a term-by-term translation model. Average precision dropped from 0.4479 to 0.4012 as a result of this structuring.

D.3 Structuring examples

Figure D.3 and Figure D.4 illustrate the effect of structuring mixed word and concept queries based on a term-by-term translation model. Words and concepts grouped in "#wsyn" are equivalence classes.

D.4 Example of a comparable document

Table D.4 lists an example document in three comparable representations (text, MeSH and UMLS$_{++}$). A corpus of documents in such a parallel representation was used for training the translation models described in chapter 5. Note that the text-based representation has not been tokenised (for readability). The concepts in the UMLS$_{++}$ representation have been sorted in alphabetical order and duplicates have been removed.

Table D.4: Example of a document in three parallel representations (PMID: 10050890, judged relevant for topic 111).

Text	Fatal familial insomnia: a new Austrian family. We present clinical, pathological and molecular features of the first Austrian family with fatal familial insomnia. Detailed clinical data are available in five patients and autopsy in four patients. Age at onset of disease ranged between 20 and 60 years, and disease duration between 8 and 20 months. Severe loss of weight was an early symptom in all five patients. Four patients developed insomnia and/or autonomic dysfunction, and all five patients developed motor abnormalities. Analysis of the prion protein (PrP) gene revealed the codon 178 point mutation and methionine homozygosity at position 129. In all brains, neuropathology showed widespread cortical astrogliosis, widespread brainstem nuclei and tract degeneration, and olivary 'pseudohypertrophy' with vacuolated neurons, in addition to neuropathological features described previously, such as thalamic and olivary degeneration. Western blotting of one brain and immunocytochemistry in four brains revealed quantitative and regional dissociation between PrP(res)(the protease resistant form of PrP) deposition and histopathology. In the cerebellar cortex of one patient, PrP(res) deposits were prominent in the molecular layer and displayed a peculiar patchy and strip-like pattern with perpendicular orientation to the surface. In another patient, a single vacuolated neuron in the inferior olivary nuclei contained prominent intravacuolar granular PrP(res) deposits, resembling changes of brainstem neurons in bovine spongiform encephalopathy.
MeSH	[Adult] [Austria] [Brain] [Female] [Humans] [Sleep Initiation and Maintenance Disorders] [Male] [Middle Aged] [Pedigree] [Prions] [Blotting, Western] [Fatal Outcome] [PrPSc Proteins]
UMLS++	[Abnormality] [Adrenal Cortex] [Age] [Aging] [Analysis] [Astrogliosis] [Austrians] [Autonomic dysfunction] [Autopsy] [Bos taurus] [Brain Stem] [Brain] [Cattle] [Cell Nucleus] [Cerebellum] [Cerebral cortex] [Codon Genus] [Congenital Abnormality] [Cytoplasmic Granules] [Disease] [Dissociation] [Encephalopathies] [Encephalopathy] [Entire brainstem] [Family] [Fives] [Forms] [Functional disorder] [Genes] [Grade 5] [Histopathology] [Homozygote] [Immunocytochemistry] [Mutation Abnormality] [Neurons] [Neuropathology] [Notodontidae] [Olivary Nucleus] [PRNP] [Pathology] [Patients] [Point Mutation] [Prion Diseases] [Protein Domain] [Proteolytic Enzyme] [Pseudohypertrophy] [Severe] [Sleeplessness] [Stripping] [Symptoms] [Tissue Degeneration] [Tract] [Unmarried person] [Weights] [Western Blot] [Western Blotting] [abnormalities] [anatomical layer] [autonomic nervous system] [body weight decreased] [cerebellar cortex structure] [entire cerebellar cortex] [entire inferior olivary nucleus] [equus asinus asinus] [histopathology] [inferior olivary nucleus] [inferiority] [insomnia adverse event] [neuropathology disease] [physiopathology] [positioning patient (procedure)] [psychological orientation] [structure of cortex of kidney] [symptoms] [vac]

References

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

Allan J., Aslam J., Belkin N., Buckley C., Callan J., Croft B., Dumais S., Fuhr N., Harman D., Harper D.J., Hiemstra D., Hofmann T., Hovy E., Kraaij W., Lafferty J., Lavrenko V., Lewis D., Liddy L., Manmatha R., McCallum A., Ponte J., Prager J., Radev D., Resnik P., Robertson S., Rosenfeld R., Roukos S., Sanderson M., Schwartz R., Singhal A., Smeaton A., Turtle H., Voorhees E., Weischedel R., Xu J., and Zhai C. (2003). Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, University of Massachusetts Amherst, September 2002. *SIGIR Forum*, 37(1):31–47.

Ando R., Dredze M., and Zhang T. (2005). TREC 2005 Genomics Track Experiments at IBM Watson. In *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005*. NIST, Gaithersburg, MD, USA.

Aronson A.R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *AMIA Symposium*, pages 17–21.

Aronson A.R., Mork J.G., Mork J.G., Gay C.W., Humphrey S.M., and Rogers W.J. (2004). The NLM Indexing Initiative's Medical Text Indexer. In *Proceedings of MEDINFO 2004*, pages 268–272.

Aronson A.R. and Rindflesch T.C. (1997). Query expansion using the UMLS Metathesaurus. *Proceedings of the AMIA Annual Fall Symposium*, pages 485–489.

Baeza-Yates R. and Ribeiro-Neto B. (1999). *Modern Information Retrieval*. ACM Press, New York, NY, USA.

Baeza-Yates R.A., Ziviani N., Marchionini G., Moffat A., and Tait J., editors (2005). *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*. ACM Press, New York, NY, USA.

Bai J. and Nie J.Y. (2008). Adapting information retrieval to query contexts. *Information Processing and Management*, 44(6):1901–1922.

Bai J., Nie J.Y., Cao G., and Bouchard H. (2007). Using query contexts in information retrieval. In Kraaij et al. (2007), pages 15–22.

Bai J., Song D., Bruza P., Nie J.Y., and Cao G. (2005). Query expansion using term relationships in language models for information retrieval. In *CIKM '05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 688–695. ACM, New York, NY, USA.

Ballesteros L. and Croft W.B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In Belkin et al. (1997), pages 84–91.

Ballesteros L. and Croft W.B. (1998). Resolving ambiguity for cross-language retrieval. In Croft et al. (1998), pages 64–71.

Beaulieu M., Baeza-Yates R., Myaeng S.H., and Järvelin K., editors (2002). *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*. ACM Press, New York, NY, USA.

Belkin N.J., Narasimhalu A.D., and Willet P., editors (1997). *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97)*. ACM Press, New York, NY, USA.

Berger A. and Lafferty J. (1999). Information retrieval as statistical translation. In Hearst et al. (1999), pages 222–229.

Bian G.W. and Chen H.H. (1998). Integrating Query Translation and Document Translation in a Cross-language Information Retrieval System. In *AMTA '98: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, pages 250–265. Springer-Verlag, London, UK.

Boughanem M., Chrisment C., and Nassr N. (2002). Investigation on Disambiguation in CLIR: Aligned Corpus and Bi-directional Translation-Based Strategies. In *CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 158–168. Springer-Verlag, London, UK.

Braun L. (2008). *Pro-Active Medical Information Retrieval*. Ph.D. thesis, University of Maastricht.

Brown P.F., Pietra S.A.D., Pietra V.J.D., and Mercer R.L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Buckley C., Salton G., and Allan J. (1994). The effect of adding relevance information in a relevance feedback environment. In Croft and van Rijsbergen (1994), pages 292–300.

Buckley C. and Voorhees E.M. (2004). Retrieval evaluation with incomplete information. In Sanderson et al. (2004), pages 25–32.

Büttcher S., Clarke C.L.A., and Cormack G.V. (2004). Domain-Specific Synonym Expansion and Validation for Biomedical Information Retrieval. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*. NIST, Gaithersburg, MD, USA.

Callan J., Cormaxk G., Clarke C., Hawking D., and Smeaton A., editors (2003). *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*. ACM Press, New York, NY, USA.

Camous F. (2007). *Ontology-based Document Representation for Biomedical Information Retrieval*. Ph.D. thesis, Dublin City University.

Carpenter B. (2004). Phrasal Queries with LingPipe and Lucene: Ad Hoc Genomics Text Retrieval. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*. NIST, Gaithersburg, MD, USA.

Caviedes J.E. and Cimino J.J. (2004). Towards the development of a conceptual distance metric for the UMLS. *Journal of Biomedical Informatics*, 37(2):77–85.

Chen L., Liu H., and Friedman C. (2005). Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21(2):248–56.

Chen L. and Thiel U. (2004). Language Modeling for Effective Construction of Domain Specific Thesauri. *Natural Language Processing and Information Systems*, pages 552–564.

Chen S. and Goodman J. (1998). An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University.

Chirita P.A., Firan C.S., and Nejdl W. (2007). Personalized query expansion for the web. In Kraaij et al. (2007), pages 7–14.

Church K.W. and Hanks P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Cleverdon C. (1967). The Cranfield tests on index language devices. In *ASLIB proceedings*, pages 173–192.

Cleverdon C.W., Mills J., and Keen M. (1966). Factors determining the performance of indexing systems. *ASLIB Cranfield project, Cranfield*.

Coletti M.H. and Bleich H.L. (2001). Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association*, 8(4):317–323.

Croft W., Harper D., D.H.Kraft, and Zobel J., editors (2001). *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*. ACM Press, New York, NY, USA.

Croft W., Moffat A., van Rijsbergen C., Wilkinson R., and Zobel J., editors (1998). *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*. ACM Press, New York, NY, USA.

Croft W. and van Rijsbergen C., editors (1994). *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*. ACM Press, New York, NY, USA.

Croft W.B. (1993). Knowledge-Based and Statistical Approaches to Text Retrieval. *IEEE Expert: Intelligent Systems and Their Applications*, 8(2):8–12.

Darwish K. and Oard D.W. (2003). Probabilistic structured query methods. In Callan et al. (2003), pages 338–344.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., and Harshman R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Dorff K.C., Wood M.J., and Campagne F. (2006). Twease at TREC 2006: Breaking and fixing BM25 scoring with query expansion. In *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006*. NIST, Gaithersburg, MD, USA.

Duda R.O., Hart P.E., and Stork D.G. (2000). *Pattern Classification*. Wiley-Interscience, 2nd edition edition.

Dumais S., Cutrell E., Cadiz J., Jancke G., Sarin R., and Robbins D.C. (2003). Stuff I've seen: a system for personal information retrieval and re-use. In Callan et al. (2003), pages 72–79.

Efthimiadis E.N., Dumais S.T., Hawking D., and Järvelin K., editors (2006). *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*. ACM Press, New York, NY, USA.

Fisher R.A. (1935). *The Design of Experiments*. Oliver and Boyd, first edition.

Fix E. and Hodges J. (1951). Nonparametric discrimination: consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas.

Fonseca B.M., Golgher P., Pôssas B., Ribeiro-Neto B., and Ziviani N. (2005). Concept-based interactive query expansion. In *CIKM '05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 696–703. ACM, New York, NY, USA.

Fox E.A. and Shaw J.A. (1993). Combination of Multiple Searches. In *Proceedings of the Second Text REtrieval Conference, TREC-2*, pages 243–252. NIST, Gaithersburg, MD, USA.

Frei H.P., Harman D., Schäuble P., and Wilkinson R., editors (1996). *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96)*. ACM Press, New York, NY, USA.

Fung P., Xiaohu L., and Shun C.C. (1999). Mixed language query disambiguation. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 333–340. Association for Computational Linguistics, Morristown, NJ, USA.

Furnas G.W., Landauer T.K., Gomez L.M., and Dumais S.T. (1987). The Vocabulary Problem in Human-System Communication: an Analysis and a Solution. *Communications of the ACM*, 30(11):964–971.

Gao J., Nie J.Y., and Zhou M. (2006). Statistical query translation models for cross-language information retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5(4):323–359.

Gollins T. and Sanderson M. (2001). Improving cross language retrieval with triangulated translation. In Croft et al. (2001), pages 90–95.

Greenberg S.J. and Gallagher P.E. (2009). The great contribution: Index Medicus, Index-Catalogue, and IndexCat. *Journal of the Medical Library Association*, 97(2):108–113.

Guo Y., Harkema H., and Gaizauskas R. (2004). Sheffield University and the TREC 2004 Genomics Track: Query Expansion Using Synonymous Terms. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*. NIST, Gaithersburg, MD, USA.

Haynes R.B., McKibbon K.A., Walker C.J., Ryan N., Fitzgerald D., and Ramsden M.F. (1990). Online access to MEDLINE in clinical settings. A study of use and usefulness. *Annals of Internal Medicine*, 112(1):78–84.

Heaps H.S. (1978). *Information Retrieval: Computational and Theoretical Aspects*. Academic Press Inc., Orlando, FL, USA.

Hearst M., Gey F., and Tong R., editors (1999). *Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. ACM Press, New York, NY, USA.

Hearst M.A. (1999). Untangling Text Data Mining. In *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*. University of Maryland.

Hersh W. and Bhupatiraju R. (2003). TREC Genomics Track Overview. In *Proceedings of the Twelfth Text REtrieval Conference TREC 2003*. NIST, Gaithersburg, MD, USA.

Hersh W., Bhupatiraju R., and Price S. (2003). Phrases, boosting, and query expansion using external knowledge resources for genomic information retrieval. In *Proceedings of the Twelfth Text REtrieval Conference TREC 2003*. NIST, Gaithersburg, MD, USA.

Hersh W., Bhupatiraju R., Ross L., Cohen A., Kraemer D., and Johnson P. (2004). TREC 2004 Genomics Track Overview. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*. NIST, Gaithersburg, MD, USA.

Hersh W., Buckley C., Leone T.J., and Hickam D. (1994a). OHSUMED: an interactive retrieval evaluation and new large test collection for research. In Croft and van Rijsbergen (1994), pages 192–201.

Hersh W., Cohen A., Roberts P., and Rekapalli H. (2006). TREC 2006 Genomics Track Overview. In *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006*. NIST, Gaithersburg, MD, USA.

Hersh W., Cohen A., Ruslen L., and Roberts P. (2007). TREC 2007 Genomics Track Overview. In *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007*. NIST, Gaithersburg, MD, USA.

Hersh W., Cohen A., Yang J., Bhupatiraju R., Roberts P., and Hearst M. (2005). TREC 2005 Genomics Track Overview. In *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005*. NIST, Gaithersburg, MD, USA.

Hersh W. and Hickam D. (1995). Information retrieval in medicine: the SAPHIRE experience. *Journal of the American Society for Information Science*, 8 Pt 2:1433–1437.

Hersh W., Price S., and Donohoe L. (2000). Assessing thesaurus-based query expansion using the UMLS Metathesaurus. *Proceedings of the AMIA Symposium*, pages 344–348.

Hersh W.R. (2009). *Information Retrieval: A Health and Biomedical Perspective*. Health Informatics. Springer-Verlag, New York, 3rd edition.

Hersh W.R., Hickam D.H., Haynes R.B., and McKibbon K.A. (1994b). A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association*, 1(1):51–60.

Hiemstra D. and Kraaij W. (1999). Twenty-One at TREC-7: ad-hoc and cross-language track. In *Proceedings of the Seventh Text REtrieval Conference, TREC-7*, pages 227–238. NIST, Gaithersburg, MD, USA.

Hiemstra D., Robertson S., and Zaragoza H. (2004). Parsimonious language models for information retrieval. In Sanderson et al. (2004), pages 178–185.

Hirst G. and St Onge D. (1998). *Lexical Chains as representation of context for the detection and correction malapropisms*, chapter 13, pages 305–332. The MIT Press.

Huang X., Hu B., and Rohian H. (2006). York University at TREC 2006: Genomics Track. In *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006*. NIST, Gaithersburg, MD, USA.

Hull D. (1993). Using statistical testing in the evaluation of retrieval experiments. In Korfhage et al. (1993), pages 329–338.

Hull D.A. and Grefenstette G. (1996). Querying across languages: A Dictionary-Based approach to

Multilingual Information Retrieval. In Frei et al. (1996), pages 49–57.

Ide N.C., Loane R.F., and Demner-Fushman D. (2007). Essie: a concept-based search engine for structured biomedical text. *Journal of the American Medical Informatics Association*, 14(3):253–263.

Jacob E.K. (2004). Classification and categorization : a difference that makes a difference. *Library Trends*, 52(3):515–540.

Jang M.G., Myaeng S.H., and Park S.Y. (1999). Using mutual information to resolve query translation ambiguities and query term weighting. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 223–229. Association for Computational Linguistics, Morristown, NJ, USA.

Jelinek F. and Mercer R.L. (1980). Interpolated Estimation of Markov Source Parameters from Sparse Data. In *Workshop Pattern Recognition in Practice*. Amsterdam, The Netherlands.

Jiang J. and Zhai C. (2007). An empirical study of tokenization strategies for biomedical information retrieval. *Information Retrieval*, 10(4-5):341–363.

Jiang J.J. and Conrath D.W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*, pages 9008+.

Jing Y. and Croft (1994). An Association Thesaurus for Information Retrieval. In *Proceedings of RIAO '94*.

Joachims T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137–142.

Joho H., Sanderson M., and Beaulieu M. (2004). A Study of User Interaction with a Concept-Based Interactive Query Expansion Support Tool. *Advances in Information Retrieval*, 2997:42–56.

Katz S.M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401.

Kim W., Aronson A.R., and Wilbur W.J. (2001). Automatic MeSH term assignment and quality assessment. In S. Bakken, editor, *Proc AMIA Symp*, pages 319–323. Washington DC, USA.

Korfhage R., Rasmussen E., and Willett P., editors (1993). *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93)*. ACM Press, New York, NY, USA.

Korfhage R.R. (1984). Query enhancement by user profiles. In C.J. van Rijsbergen, editor, *Research and Development in Information Retrieval, Proceedings of the Third Joint BCS/ACM Symposium on Research and Development in Information Retrieval*. Cambridge University Press.

Kraaij W. (2004). *Variations on Language Modeling for Information Retrieval*. Ph.D. thesis, University of Twente.

Kraaij W., de Vries A.P., Clarke C.L.A., Fuhr N., and Kando N., editors (2007). *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*. ACM Press, New York, NY, USA.

Kraaij W., Weeber M., Raaijmakers S., and Jelier R. (2004). MeSH based feedback, concept recognition and stacked classification for curation tasks. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*. NIST, Gaithersburg, MD, USA.

Kraaij W., Westerveld T., and Hiemstra D. (2002). The Importance of Prior Probabilities for Entry Page Search. In Beaulieu et al. (2002), pages 27–34.

Krallinger M. and Valencia A. (2005). Text-mining and information-retrieval services for molecular biology. *Genome Biology*, 6(7):224.

Krauthammer M. and Nenadic G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526.

Krovetz R. (1997). Homonymy and polysemy in information retrieval. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 72–79.

Association for Computational Linguistics, Morristown, NJ, USA.

Kurland O. and Lee L. (2009). Clusters, language models, and ad hoc information retrieval. *ACM Transactions on Information Systems*, 27(3):1–39.

Kwok K.L. (2000). Exploiting a Chinese-English bilingual wordlist for English-Chinese cross language information retrieval. In *IRAL '00: Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, pages 173–179. ACM, New York, NY, USA.

Lafferty J. and Zhai C. (2001). Document language models, query models, and risk minimization for information retrieval. In Croft et al. (2001), pages 111–119.

Lam W. and Ho C.Y. (1998). Using a generalized instance set for automatic text categorization. In Croft et al. (1998), pages 81–89.

Lam W., Ruiz M., Ruiz M., and Srinivasan P. (1999). Automatic Text categorization and its application to text retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 11:865–879.

Lancaster F.W. (1969). MEDLARS: Report on the Evaluation of Its Operating Efficiency. *American Documentation*, 20(2):119–142.

Lavrenko V., Choquette M., and Croft W.B. (2002). Cross-lingual relevance models. In Beaulieu et al. (2002), pages 175–182.

Lavrenko V. and Croft W.B. (2001). Relevance based language models. In Croft et al. (2001), pages 120–127.

Lesk M. (2008). Recycling Information: Science Through Data Mining. *International Journal of Digital Curation*, 3(1).

Lewis D.D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nédellec and C. Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, 1398, pages 4–15. Springer Verlag, Heidelberg, DE, Chemnitz, DE.

Lewis D.D., Schapire R.E., Callan J.P., and Papka R. (1996). Training algorithms for linear text classifiers. In Frei et al. (1996), pages 298–306.

Li Y., Bandar Z., and Mclean D. (2002). Measuring Semantic Similarity Between Words Using Lexical Knowledge and Neural Networks. *Intelligent Data Engineering and Automated Learning – IDEAL 2002*, pages 67–74.

Li Y., Bandar Z.A., and McLean D. (2003). An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882.

Lin D. (1998a). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th conference on Computational linguistics*, pages 768–774. Association for Computational Linguistics, Morristown, NJ, USA.

Lin D. (1998b). An Information-Theoretic Definition of Similarity. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Lin J. and Wilbur W.J. (2007). PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, 8(1):423.

Liu X.Y., Zhou Y.M., and Zheng R.S. (2007). Measuring Semantic Similarity in Wordnet. In *Machine Learning and Cybernetics, 2007 International Conference on*, volume 6, pages 3431–3435.

Liu Y., Jin R., and Chai J.Y. (2005). A maximum coherence model for dictionary-based cross-language information retrieval. In Baeza-Yates et al. (2005), pages 536–543.

Lovins J. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31.

Lu Z., Kim W., and Wilbur W.J. (2009). Evaluation of query expansion using MeSH in PubMed. *Information Retrieval*, 12(1):69–80.

Luhn H.P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317.

Maglott D., Ostell J., Pruitt K.D., and Tatusova T. (2007). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 35(Database issue):D26–31.

Manning C.D., Raghavan P., and Schütze H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Manning C.D. and Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, USA.

McCarley J.S. (1999). Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 208–214. Association for Computational Linguistics, Morristown, NJ, USA.

McCray A.T. (1998). The nature of lexical knowledge. *Methods of Information in Medicine*, 37:353–60.

McCray A.T. and Miller R.A. (1998). Making the conceptual connections: the Unified Medical Language System (UMLS) after a decade of research and development. *Journal of the American Medical Informatics Association*, 5(1):129–30.

McNamee P. (2008). *Textual Representations for Corpus-Based Bilingual Retrieval*. Ph.D. thesis, University of Maryland.

Meij E., Trieschnigg R.B., de Rijke M., and Kraaij W. (2010). Conceptual language models for domain-specific retrieval. *Information Processing and Management*, 47(22).

Metzler D. and Croft B.W. (2005). A Markov random field model for term dependencies. In Baeza-Yates et al. (2005).

Miller D.R.H., Leek T., and Schwartz R.M. (1999). A Hidden Markov Model Information Retrieval System. In Hearst et al. (1999), pages 214–221.

Miller G.A. and Charles W.G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Mitra M., Singhal A., and Buckley C. (1998). Improving automatic query expansion. In Croft et al. (1998).

Moreau N. (2009). Best practices in language resources for multilingual information access. Treble-CLEF: Evaluation, Best Practices & Collaboration for Multilingual Information Access.

Morgan A.A., Lu Z., Wang X., Cohen A.M., Fluck J., Ruch P., Divoli A., Fundel K., Leaman R., Hakenberg J., Sun C., Liu H.h., Torres R., Krauthammer M., Lau W.W., Liu H., Hsu C.N., Schuemie M., Cohen K.B., and Hirschman L. (2008). Overview of BioCreative II gene normalization. *Genome Biology*, 9 Suppl 2:S3.

Myaeng S.H., Oard D.W., Sebastiani F., Chua T.S., and Leong M.K., editors (2008). *Proceedings of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*. ACM, New York, NY, USA.

Na S.H., Kang I.S., and Lee J.H. (2007). Parsimonious translation models for information retrieval. *Information Processing and Management*, 43(1):121–145.

Nelson S.J., Johnston W.D., and Humphreys B.L. (2001). Relationships in Medcial Subject Headings. `http://www.nlm.nih.gov/mesh/meshrels.html`.

Nenadic G., Spasic I., and Ananiadou S. (2005). Mining Biomedical Abstracts: What's in a Term? *Natural Language Processing IJCNLP 2004*, pages 797–806.

Nguyen H. and Al-Mubaid H. (2006). New ontology-based semantic similarity measure for the biomedical domain. In *2006 IEEE Int. Conf. on Granular Computing*, pages 623–628.

Oard D.W. (1997). Alternative approaches for cross-language text retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence*, pages 131–139.

Oard D.W., He D., and Wang J. (2008). User-assisted query translation for interactive cross-language information retrieval. *Information Processing and Management*, 44(1):181–211.

Pearson H. (2001). Biology's name game. *Nature*, 411(6838):631–632.

Pedersen T., Pakhomov S.V.S., Patwardhan S., and Chute C.G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299.

Pirkola A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In Croft et al. (1998), pages 55–63.

Pirkola A. (2005). TREC 2005 Genomics Track Experiments at UTA. In *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005*. NIST, Gaithersburg, MD, USA.

Pirkola A., Hedlund T., Keskustalo H., and Järvelin K. (2001). Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Information Retrieval*, 4(3-4):209–230.

Pirkola A. and Leppänen E. (2003). TREC 2003 Genomics Track Experiments at UTA: Query Expansion with Predefinded High Frequency Terms. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*, pages 796–805. NIST, Gaithersburg, MD, USA.

Ponte J.M. (2001). Is Information Retrieval Anything More Than Smoothing. In *Proceedings of the Workshop on Language Models for Information Retrieval (LMIR2001)*.

Ponte J.M. and Croft W.B. (1998). A Language Modeling Approach to Information Retrieval. In Croft et al. (1998), pages 275–281.

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

Qiu Y. (1995). *Automatic Query expansion Based on a Similarity Thesaurus*. Ph.D. thesis, ETH Zürich.

Qiu Y. and Frei H.P. (1993). Concept based query expansion. In Korfhage et al. (1993), pages 160–169.

Rada R., Mili H., Bicknell E., and Blettner M. (1989). Development and application of a metric on semantic nets. *IEEE International Conference on Systems, Man, and Cybernetics*, 19(1):17–30.

Rak R., Kurgan L.A., and Reformat M. (2007). Multilabel associative classification categorization of MEDLINE articles into MeSH keywords. *IEEE engineering in medicine and biology magazine*, 26(2):47–55.

Resnik P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *IJCAI 1995*, pages 448–453.

Rocchio J. (1971). Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall.

Rubenstein H. and Goodenough J.B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Ruch P. (2006). Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, 22(6):658–664.

Ruiz M. (2005). Experiments on Genomics Ad Hoc Retrieval. In *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005*. NIST, Gaithersburg, MD, USA.

Ruiz M., Diekema A., and Sheridan P. (1999). CINDOR Conceptual Interlingua Document Retrieval: TREC-8 Evaluation. In *Proceedings of the Eighth Text REtrieval Conference, TREC-8*, pages 597–606. NIST, Gaithersburg, MD, USA.

Ruiz M.E. and Srinivasan P. (2002). Hierarchical text categorization using neural networks. *Information Retrieval*, 5:87–118.

Salton G. (1968). *Automatic Information Organisation and Retrieval*. McGraw-Hill, New York.

Salton G. (1971). Information analysis and dictionary construction. In *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall.

Salton G. (1972). A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the American Society for Information Science*, 23(2):75–84.

Sanderson M. (1994). Word sense disambiguation and information retrieval. In Croft and van Rijsbergen (1994), pages 142–151.

Sanderson M. (2000). Retrieving with Good Sense. *Information Retrieval*, 2(1):49–69.

Sanderson M., Järvelin K., Allan J., and Bruza P., editors (2004). *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*. ACM Press, New York, NY, USA.

Schijvenaars B., Mons B., Weeber M., Schuemie M., van Mulligen E., Wain H., and Kors J. (2005). Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics*, 6(1):149.

Schuemie M., Jelier R., and Kors J. (2007a). Peregrine: lightweight gene name normalization by dictionary lookup. In *Second BioCreative Workshop*, pages 131–133. Madrid.

Schuemie M., Trieschnigg D., and Kraaij W. (2007b). Cross Language Information Retrieval for Biomedical Literature. In *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007*. NIST, Gaithersburg, MD, USA.

Schuemie M.J., Kors J.A., and Mons B. (2005). Word sense disambiguation in the biomedical domain: an overview. *Journal of Computational Biology*, 12(5):554–65.

Schuemie M.J., Mons B., Weeber M., and Kors J.A. (2007c). Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. *Journal of Biomedical Informatics*, 40(3):316–324.

Shatkay H. (2005). Hairpins in bookstacks: information retrieval from biomedical text. *Briefings in Bioinformatics*, 6(3):222–38.

Shatkay H. and Feldman R. (2003). Mining the biomedical literature in the genomic era: an overview. *Journal of Computational Biology*, 10(6):821–55.

Si L., Lu J., and Callan J. (2006). Combining Multiple Resources, Evidences and Criteria for Genomic Information Retrieval. In *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006*. NIST, Gaithersburg, MD, USA.

Smucker M.D., Allan J., and Carterette B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *CIKM '07: Proceedings of the 16th ACM International Conference on Information and Knowledge Management*, pages 623–632. ACM, New York, NY, USA.

Sohn S., Kim W., Comeau D.C., and Wilbur W.J. (2008). Optimal training sets for Bayesian prediction of MeSH assignment. *Journal of the American Medical Informatics Association*, 15(4):546–553.

Spärck Jones K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Spärck Jones K. and Jackson D.M. (1970). The Use of Automatically-Obtained Keyword Classifications for Information Retrieval. *Information Processing and Management*, 5(1):175–201.

Spärck Jones K. and Van Rijsbergen C. (1975). Report on the Need for and Provision of an Ideal Information Retrieval Test Collection. British Library Research and Development report 5266, Cambridge University Computer Laboratory.

Srinivasan P. (1996a). Optimal document-indexing vocabulary for MEDLINE. *Information Processing and Management*, pages 503–514.

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

Stokes N., Li Y., Cavedon L., and Zobel J. (2007). Exploring Abbreviation Expansion for Genomic Information Retrieval. In *Proceedings of the Australasian Language Technology Workshop 2007*, pages 100–108. Melbourne, Australia.

Stokes N., Li Y., Cavedon L., and Zobel J. (2009). Exploring criteria for successful query expansion in the genomic domain. *Information Retrieval*, 12(1):17–50.

Swanson D.R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18.

Teevan J., Dumais S.T., and Horvitz E. (2005). Personalizing search via automated analysis of interests and activities. In Baeza-Yates et al. (2005), pages 449–456.

Tjong Kim Sang E.F. and De Meulder F. (2003). Introduction to the CoNLL-2003 shared task:

language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 142–147. Association for Computational Linguistics, Morristown, NJ, USA.

Tomlinson S. (2003). Robust, Web and Genomic Retrieval with Hummingbird SearchServer at TREC 2003. In *Proceedings of the Twelfth Text REtrieval Conference TREC 2003*, pages 254–267. NIST, Gaithersburg, MD, USA.

Trieschnigg D. (2008). Biomedical cross-language information retrieval. In Myaeng et al. (2008), page 897.

Trieschnigg D., Hiemstra D., de Jong F., and Kraaij W. (2010). A Cross-lingual Framework for Monolingual Biomedical Information Retrieval. In *CIKM '10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA. Accepted for publication.

Trieschnigg D., Kraaij W., and de Jong F. (2007). The influence of basic tokenization on biomedical document retrieval. In Belkin et al. (1997), pages 803–804.

Trieschnigg D., Meij E., de Rijke M., and Kraaij W. (2008). Measuring concept relatedness using language models. In Myaeng et al. (2008), pages 823–824.

Trieschnigg D., Pezik P., Lee V., Kraaij W., de Jong F., and Rebholz-Schuhmann D. (2009). MeSH Up: Effective MeSH Text Classification and Improved Document Retrieval. *Bioinformatics*, 25(11):1412–1418.

Trieschnigg D., Schuemie M., and Kraaij W. (2006). Concept Based Document Retrieval for Genomics Literature. In *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006*. NIST, Gaithersburg, MD, USA.

Tuason O., Chen L., Liu H., Blake J.A., and Friedman C. (2004). Biological nomenclatures: a source of lexical knowledge and ambiguity. *Pacific Symposium on Biocomputing*, pages 238–49.

Urbain J., Goharian N., and Frieder O. (2006). IIT TREC 2006: Genomics Track. In *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006*. NIST, Gaithersburg, MD, USA.

van Rijsbergen C.J. (1979). *Information Retrieval*. Butterworths, London, second edition.

Voorhees E.M. (1994). Query expansion using lexical-semantic relations. In Croft and van Rijsbergen (1994).

Voorhees E.M. and Harman D.K., editors (2005). *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.

Wang J. and Oard D.W. (2006). Combining bidirectional translation and synonymy for cross-language information retrieval. In Efthimiadis et al. (2006), pages 202–209.

White R.W. and Marchionini G. (2007). Examining the effectiveness of real-time query expansion. *Information Processing and Management*, 43(3):685–704.

Wu Z. and Palmer M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, Morristown, NJ, USA.

Xu J. and Croft W.B. (1996). Query Expansion Using Local and Global Document Analysis. In Frei et al. (1996), pages 4–11.

Yang D. and Powers D.M.W. (2005). Measuring semantic similarity in the taxonomy of WordNet. In *ACSC '05: Proceedings of the Twenty-eighth Australasian conference on Computer Science*, pages 315–322. Australian Computer Society, Inc., Darlinghurst, Australia, Australia.

Zhai C. (2008). *Statistical Language Models for Information Retrieval*. Synthesis Lectures Series on Human Language Technologies. Morgan and Claypool Publishers.

Zhai C. and Lafferty J. (2001). Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01: Proceedings of the tenth International Conference on Information and Knowledge Management*, pages 403–410. ACM, New York, NY, USA.

Zhai C. and Lafferty J. (2004). A study of smoothing methods for language models applied to

information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214.

Zhai C. and Lafferty J. (2006). A risk minimization framework for information retrieval. *Information Processing and Management*, 42(1):31–55.

Zhou W., Torvik V.I., and Smalheiser N.R. (2006a). ADAM: another database of abbreviations in MEDLINE. *Bioinformatics*, 22(22):2813–2818.

Zhou W. and Yu C.T. (2006). TREC Genomics Track at UIC. In *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006*. NIST, Gaithersburg, MD, USA.

Zhou W., Yu C.T., Smalheiser N.R., Torvik V.I., and Hong J. (2007). Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In Kraaij et al. (2007), pages 655–662.

Zhou X., Hu X., Zhang X., Lin X., and Song I.Y. (2006b). Context-sensitive semantic smoothing for the language modeling approach to genomic IR. In Efthimiadis et al. (2006), pages 170–177.

Zipf G.K. (1949). *Human behaviour and the principle of the least effort*. Addison Wesley.

Zobel J. (1998). How reliable are the results of large-scale information retrieval experiments? In Croft et al. (1998), pages 307–314.

Summary

In this thesis we investigate the possibility to integrate domain-specific knowledge into biomedical information retrieval (IR). Recent decades have shown a fast growing interest in biomedical research, reflected by an exponential growth in scientific literature. Biomedical IR is concerned with the disclosure of these vast amounts of written knowledge. Biomedical IR is not only important for end-users, such as biologists, biochemists, and bioinformaticians searching directly for relevant literature but also plays an important role in more sophisticated knowledge discovery. An important problem for biomedical IR is dealing with the complex and inconsistent terminology encountered in biomedical publications. Multiple synonymous terms can be used for single biomedical concepts, such as genes and diseases. Conversely, single terms can be ambiguous, and may refer to multiple concepts. Dealing with the terminology problem requires domain knowledge stored in terminological resources: controlled indexing vocabularies and thesauri. The integration of this knowledge in modern word-based information retrieval is, however, far from trivial. This thesis investigates the problem of handling biomedical terminology based on three research themes.

The first research theme deals with robust word-based retrieval. Effective retrieval models commonly use a word-based representation for retrieval. As so many spelling variations are present in biomedical text, the way in which these word-based representations are obtained affect retrieval effectiveness. We investigated the effect of choices in document preprocessing heuristics on retrieval effectiveness. This investigation included stop-word removal, stemming, different approaches to breakpoint identification and normalisation, and character n-gramming. In particular breakpoint identification and normalisation (that is determining word parts in biomedical compounds) showed a strong effect on retrieval performance. A combination of effective preprocessing heuristics was identified and used to obtain word-based representations from text for the remainder of this thesis.

The second research theme deals with concept-based retrieval. We investigated two representation vocabularies for concept-based indexing, one based on the Medical Subject Headings thesaurus, the other based on the Unified Medical Language System metathesaurus extended with a number of gene and protein dictionaries. We investigated the following five topics.

1. How documents are represented in a concept-based representation.
2. To what extent such a document representation can be obtained automatically.
3. To what extent a text-based query can be automatically mapped onto a concept-based representation and how this affects retrieval performance.
4. To what extent a concept-based representation is effective in representing information needs.

5. How the relationship between text and concepts can be used to determine the relatedness of concepts.

We compared different classification systems to obtain concept-based document and query representations automatically. We proposed two classification methods based on statistical language models, one based on K-Nearest Neighbours (KNN) and one based on Concept Language Models (CLM).

For a selection of classification systems we carried out a document classification experiment in which we investigated to what extent automatic classification could reproduce manual classification. The proposed KNN system performed well in comparison to the out-of-the-box systems. Manual analysis indicated the improved exhaustiveness of automatic classification over manual classification. Retrieval based on only concepts was demonstrated to be significantly less effective than word-based retrieval. This deteriorated performance could be explained by errors in the classification process, limitations of the concept vocabularies and limited exhaustiveness of the concept-based document representations. Retrieval based on a combination of word-based and automatically obtained concept-based query representations did significantly improve word-only retrieval. In an artificial setting, we compared the optimal retrieval performance which could be obtained with word-based and concept-based representations. Contrary to our intuition, on average a single word-based query performed better than a single concept-based representation, even when the best concept term precisely represented part of the information need.

We investigated to what extent the relatedness between pairs of concepts as indicated by human judgements could be automatically reproduced. Results on a small test set indicated that a method based on comparing concept language models performed particularly well in comparison to systems based on taxonomy structure, information content and (document) association.

In the third and last research theme of this thesis we propose a framework for concept-based retrieval. We approached the integration of domain knowledge in monolingual information retrieval as a cross-lingual information retrieval (CLIR) problem. Two languages were identified in this monolingual setting: a word-based representation language based on free text, and a concept-based representation language based on a terminological resource. Similar to what is common in traditional CLIR, queries and documents are translated into the same representation language and matched. The cross-lingual perspective gives us the opportunity to adopt a large set of established CLIR methods and techniques for this domain. In analogy to established CLIR practise, we investigated translation models based on a parallel corpus containing documents in multiple representations and translation models based on a thesaurus. Surprisingly, even the integration of very basic translation models showed improvements in retrieval effectiveness over word-only retrieval. A translation model based on pseudo-feedback translation was shown to perform particularly well. We proposed three extensions to a basic cross-lingual retrieval model which, similar to previous approaches in established CLIR, improved retrieval effectiveness by combining multiple translation models. Experimental results indicate that, even when using very basic translation models, monolingual biomedical IR can benefit from a cross-lingual approach to integrate domain knowledge.

Directions for future work are using these concepts for communication between user and retrieval system, extending upon the translation models and extending CLIR-enhanced concept-based retrieval outside the biomedical domain.

Curriculum Vitae

Dolf Trieschnigg was born on April 20th 1981 in Heino. In 1999 he finished secondary school at the Marianum in Groenlo. He continued his education at the Department of Computer Science at the University of Twente in Enschede. During his Computer Science studies he worked for several small companies as a software developer. During a four months internship at a large company in India, he developed an intranet search system In 2004, he started his Master thesis project at TNO in Delft (a research institute founded by the Dutch government) where he worked on hierarchical topic detection in digital news archives. He carried out his PhD. project in the contetxt of the NBIC BioRange program, focusing on biomedical information retrieval. In 2008, he was awarded an EBI fellowship from the Netherlands Genomics Initiative which allowed for a six months research visit at the Text Mining Group of the EMBL's European Bioinformatics Institute. After finishing his PhD. thesis in 2010, he joined the Database Group of the University of Twente as a postdoctoral researcher. His current research focuses on the integration of domain knowledge in information retrieval and distributed information retrieval.

SIKS Dissertation Series

Since 1998, all dissertations written by PhD students who have conducted their research under auspices of a senior research fellow of the SIKS research school are published in the SIKS Dissertation Series. This thesis is the 280th in the series.

2010-35 Dolf Trieschnigg (UT), *Proof of Concept: Concept-based Biomedical Information Retrieval*.
2010-34 Teduh Dirgahayu (UT), *Interaction Design in Service Compositions*.
2010-33 Robin Aly (UT), *Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval*.
2010-32 Marcel Hiel (UvT), *An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems*.
2010-31 Victor de Boer (UVA), *Ontology Enrichment from Heterogeneous Sources on the Web*.
2010-30 Marieke van Erp (UvT), *Accessing Natural History – Discoveries in data cleaning, structuring, and retrieval*.
2010-29 Stratos Idreos (CWI), *Database Cracking: Towards Auto-tuning Database Kernels*.
2010-28 Arne Koopman (UU), *Characteristic Relational Patterns*.
2010-27 Marten Voulon (UL), *Automatisch contracteren*.
2010-26 Ying Zhang (CWI), *XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines*.
2010-25 Zulfiqar Ali Memon (VU), *Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective*.
2010-24 Dmytro Tykhonov (TUD), *Designing Generic and Efficient Negotiation Strategies*.
2010-23 Bas Steunebrink (UU), *The Logical Structure of Emotions*.
2010-22 Michiel Hildebrand (CWI), *End-user Support for Access to Heterogeneous Linked Data*.
2010-21 Harold van Heerde (UT), *Privacy-aware data management by means of data degradation*.
2010-20 Ivo Swartjes (UT), *Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative*.
2010-19 Henriette Cramer (UvA), *People's Responses to Autonomous and Adaptive Systems*.
2010-18 Charlotte Gerritsen (VU), *Caught in the Act: Investigating Crime by Agent-Based Simulation*.
2010-17 Spyros Kotoulas (VU), *Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications*.
2010-16 Sicco Verwer (TUD), *Efficient Identification of Timed Automata, theory and practice*.
2010-15 Lianne Bodenstaff (UT), *Managing Dependency Relations in Inter-Organizational Models*.
2010-14 Sander van Splunter (VU), *Automated Web Service Reconfiguration*.
2010-13 Gianluigi Folino (RUN), *High Performance Data Mining using Bio-inspired techniques*.
2010-12 Susan van den Braak (UU), *Sensemaking software for crime analysis*.
2010-11 Adriaan Ter Mors (TUD), *The world according to MARP: Multi-Agent Route Planning*.
2010-10 Rebecca Ong (UL), *Mobile Communication and Protection of Children*.
2010-09 Hugo Kielman (UL), *Politiële gegevensverwerking en Privacy, Naar een effectieve waarborging*.
2010-08 Krzysztof Siewicz (UL), *Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments*.
2010-07 Wim Fikkert (UT), *Gesture interaction at a Distance*.
2010-06 Sander Bakkes (UvT), *Rapid Adaptation of Video Game AI*.
2010-05 Claudia Hauff (UT), *Predicting the Effectiveness of Queries and Retrieval Systems*.
2010-04 Olga Kulyk (UT), *Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments*.
2010-03 Joost Geurts (CWI), *A Document Engineering Model and Processing Framework for Multimedia documents*.
2010-02 Ingo Wassink (UT), *Work flows in Life Science*.
2010-01 Matthijs van Leeuwen (UU), *Patterns that Matter*.
2009-46 Loredana Afanasiev (UvA), *Querying XML: Benchmarks and Recursion*.
2009-45 Jilles Vreeken (UU), *Making Pattern Mining Useful*.
2009-44 Roberto Santana Tapia (UT), *Assessing Business-IT Alignment in Networked Organizations*.
2009-43 Virginia Nunes Leal Franqueira (UT), *Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients*.
2009-42 Toine Bogers (UvT), *Recommender Systems for Social Bookmarking*.
2009-41 Igor Berezhnyy (UvT), *Digital Analysis of Paintings*.
2009-40 Stephan Raaijmakers (UvT), *Multinomial Language Learning: Investigations into the Geometry of Language*.
2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin), *Service Substitution – A Behavioral Approach Based on Petri Nets*.
2009-38 Riina Vuorikari (OU), *Tags and self-organisation: a metadata ecology for learning resources in a multilingual context*.
2009-37 Hendrik Drachsler (OU), *Navigation Support for Learners in Informal Learning Networks*.
2009-36 Marco Kalz (OU), *Placement Support for Learners in Learning Networks*.
2009-35 Wouter Koelewijn (UL), *Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling*.
2009-34 Inge van de Weerd (UU), *Advancing in Software Product Management: An Incremental Method Engineering Approach*.
2009-33 Khiet Truong (UT), *How Does Real Affect Affect Affect Recognition In Speech?*.

2009-32 Rik Farenhorst (VU) and Remco de Boer (VU), *Architectural Knowledge Management: Supporting Architects and Auditors*.
2009-31 Sofiya Katrenko (UVA), *A Closer Look at Learning Relations from Text*.
2009-30 Marcin Zukowski (CWI), *Balancing vectorized query execution with bandwidth-optimized storage*.
2009-29 Stanislav Pokraev (UT), *Model-Driven Semantic Integration of Service-Oriented Applications*.
2009-28 Sander Evers (UT), *Sensor Data Management with Probabilistic Models*.
2009-27 Christian Glahn (OU), *Contextual Support of social Engagement and Reflection on the Web*.
2009-26 Fernando Koch (UU), *An Agent-Based Model for the Development of Intelligent Mobile Services*.
2009-25 Alex van Ballegooij (CWI), "*RAM: Array Database Management through Relational Mapping*".
2009-24 Annerieke Heuvelink (VU), *Cognitive Models for Training Simulations*.
2009-23 Peter Hofgesang (VU), *Modelling Web Usage in a Changing Environment*.
2009-22 Pavel Serdyukov (UT), *Search For Expertise: Going beyond direct evidence*.
2009-21 Stijn Vanderlooy (UM), *Ranking and Reliable Classification*.
2009-20 Bob van der Vecht (UU), *Adjustable Autonomy: Controling Influences on Decision Making*.
2009-19 Valentin Robu (CWI), *Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets*.
2009-18 Fabian Groffen (CWI), *Armada, An Evolving Database System*.
2009-17 Laurens van der Maaten (UvT), *Feature Extraction from Visual Data*.
2009-16 Fritz Reul (UvT), *New Architectures in Computer Chess*.
2009-15 Rinke Hoekstra (UVA), *Ontology Representation – Design Patterns and Ontologies that Make Sense*.
2009-14 Maksym Korotkiy (VU), *From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)*.
2009-13 Steven de Jong (UM), *Fairness in Multi-Agent Systems*.
2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin), *Operating Guidelines for Services*.
2009-11 Alexander Boer (UVA), *Legal Theory, Sources of Law & the Semantic Web*.
2009-10 Jan Wielemaker (UVA), *Logic programming for knowledge-intensive interactive applications*.
2009-09 Benjamin Kanagwa (RUN), *Design, Discovery and Construction of Service-oriented Systems*.
2009-08 Volker Nannen (VU), *Evolutionary Agent-Based Policy Analysis in Dynamic Environments*.
2009-07 Ronald Poppe (UT), *Discriminative Vision-Based Recovery and Recognition of Human Motion*.
2009-06 Muhammad Subianto (UU), *Understanding Classification*.
2009-05 Sietse Overbeek (RUN), *Bridging Supply and Demand for Knowledge Intensive Tasks – Based on Knowledge, Cognition, and Quality*.
2009-04 Josephine Nabukenya (RUN), *Improving the Quality of Organisational Policy Making using Collaboration Engineering*.
2009-03 Hans Stol (UvT), *A Framework for Evidence-based Policy Making Using IT*.
2009-02 Willem Robert van Hage (VU), *Evaluating Ontology-Alignment Techniques*.
2009-01 Rasa Jurgelenaite (RUN), *Symmetric Causal Independence Models*.
2008-35 Ben Torben Nielsen (UvT), *Dendritic morphologies: function shapes structure*.
2008-34 Jeroen de Knijf (UU), *Studies in Frequent Tree Mining*.
2008-33 Frank Terpstra (UVA), *Scientific Workflow Design; theoretical and practical issues*.
2008-32 Trung H. Bui (UT), *Toward Affective Dialogue Management using Partially Observable Markov Decision Processes*.
2008-31 Loes Braun (UM), *Pro-Active Medical Information Retrieval*.
2008-30 Wouter van Atteveldt (VU), *Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content*.
2008-29 Dennis Reidsma (UT), *Annotations and Subjective Machines – Of Annotators, Embodied Agents, Users, and Other Humans*.
2008-28 Ildiko Flesch (RUN), *On the Use of Independence Relations in Bayesian Networks*.
2008-27 Hubert Vogten (OU), *Design and Implementation Strategies for IMS Learning Design*.
2008-26 Marijn Huijbregts (UT), *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled*.
2008-25 Geert Jonker (UU), *Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-signed Currency*.
2008-24 Zharko Aleksovski (VU), *Using background knowledge in ontology matching*.
2008-23 Stefan Visscher (UU), *Bayesian network models for the management of ventilator-associated pneumonia*.
2008-22 Henk Koning (UU), *Communication of IT-Architecture*.
2008-21 Krisztian Balog (UVA), *People Search in the Enterprise*.
2008-20 Rex Arendsen (UVA), *Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met de overheid op de administratieve lasten van bedrijven*.
2008-19 Henning Rode (UT), *From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search*.
2008-18 Guido de Croon (UM), *Adaptive Active Vision*.
2008-17 Martin Op 't Land (TUD), *Applying Architecture and Ontology to the Splitting and Allying of Enterprises*.
2008-16 Henriëtte van Vugt (VU), *Embodied agents from a user's perspective*.
2008-15 Martijn van Otterlo (UT), *The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains*.
2008-14 Arthur van Bunningen (UT), *Context-Aware Querying; Better Answers with Less Effort*.
2008-13 Caterina Carraciolo (UVA), *Topic Driven Access to Scientific Handbooks*.
2008-12 József Farkas (RUN), *A Semiotically Oriented Cognitive Model of Knowledge Representation*.
2008-11 Vera Kartseva (VU), *Designing Controls for Network Organizations: A Value-Based Approach*.
2008-10 Wauter Bosma (UT), *Discourse oriented summarization*.
2008-09 Christof van Nimwegen (UU), *The paradox of the guided user: assistance can be counter-effective*.
2008-08 Janneke Bolt (UU), *Bayesian Networks: Aspects of Approximate Inference*.
2008-07 Peter van Rosmalen (OU), *Supporting the tutor in the design and support of adaptive e-learning*.
2008-06 Arjen Hommersom (RUN), *On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective*.
2008-05 Bela Mutschler (UT), *Modeling and simulating causal dependencies on process-aware information systems from a cost perspective*.
2008-04 Ander de Keijzer (UT), *Management of Uncertain Data – towards unattended integration*.
2008-03 Vera Hollink (UVA), *Optimizing hierarchical menus: a usage-based approach*.

2008-02 Alexei Sharpanskykh (VU), *On Computer-Aided Methods for Modeling and Analysis of Organizations*.
2008-01 Katalin Boer-Sorbán (EUR), *Agent-Based Simulation of Financial Markets: A modular, continuous-time approach*.
2007-25 Joost Schalken (VU), *Empirical Investigations in Software Process Improvement*.
2007-24 Georgina Ramírez Camps (CWI), *Structural Features in XML Retrieval*.
2007-23 Peter Barna (TUE), *Specification of Application Logic in Web Information Systems*.
2007-22 Zlatko Zlatev (UT), *Goal-oriented design of value and process models from patterns*.
2007-21 Karianne Vermaas (UU), *Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005*.
2007-20 Slinger Jansen (UU), *Customer Configuration Updating in a Software Supply Network*.
2007-19 David Levy (UM), *Intimate relationships with artificial partners*.
2007-18 Bart Orriëns (UvT), *On the development an management of adaptive business collaborations*.
2007-17 Theodore Charitos (UU), *Reasoning with Dynamic Networks in Practice*.
2007-16 Davide Grossi (UU), *Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems*.
2007-15 Joyca Lacroix (UM), *NIM: a Situated Computational Memory Model*.
2007-14 Niek Bergboer (UM), *Context-Based Image Analysis*.
2007-13 Rutger Rienks (UT), *Meetings in Smart Environments; Implications of Progressing Technology*.
2007-12 Marcel van Gerven (RUN), *Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty*.
2007-11 Natalia Stash (TUE), *Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System*.
2007-10 Huib Aldewereld (UU), *Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols*.
2007-09 David Mobach (VU), *Agent-Based Mediated Service Negotiation*.
2007-08 Mark Hoogendoorn (VU), *Modeling of Change in Multi-Agent Organizations*.
2007-07 Nataša Jovanović (UT), *To Whom It May Concern – Addressee Identification in Face-to-Face Meetings*.
2007-06 Gilad Mishne (UVA), *Applied Text Analytics for Blogs*.
2007-05 Bart Schermer (UL), *Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance*.
2007-04 Jurriaan van Diggelen (UU), *Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach*.
2007-03 Peter Mika (VU), *Social Networks and the Semantic Web*.
2007-02 Wouter Teepe (RUG), *Reconciling Information Exchange and Confidentiality: A Formal Approach*.
2007-01 Kees Leune (UvT), *Access Control and Service-Oriented Architectures*.
2006-28 Börkur Sigurbjörnsson (UVA), *Focused Information Access using XML Element Retrieval*.
2006-27 Stefano Bocconi (CWI), *Vox Populi: generating video documentaries from semantically annotated media repositories*.
2006-26 Vojkan Mihajlović (UT), *Score Region Algebra: A Flexible Framework for Structured Information Retrieval*.
2006-25 Madalina Drugan (UU), *Conditional log-likelihood MDL and Evolutionary MCMC*.
2006-24 Laura Hollink (VU), *Semantic Annotation for Retrieval of Visual Resources*.
2006-23 Ion Juvina (UU), *Development of Cognitive Model for Navigating on the Web*.
2006-22 Paul de Vrieze (RUN), *Fundaments of Adaptive Personalisation*.
2006-21 Bas van Gils (RUN), *Aptness on the Web*.
2006-20 Marina Velikova (UvT), *Monotone models for prediction in data mining*.
2006-19 Birna van Riemsdijk (UU), *Cognitive Agent Programming: A Semantic Approach*.
2006-18 Valentin Zhizhkun (UVA), *Graph transformation for Natural Language Processing*.
2006-17 Stacey Nagata (UU), *User Assistance for Multitasking with Interruptions on a Mobile Device*.
2006-16 Carsten Riggelsen (UU), *Approximation Methods for Efficient Learning of Bayesian Networks*.
2006-15 Rainer Malik (UU), *CONAN: Text Mining in the Biomedical Domain*.
2006-14 Johan Hoorn (VU), *Software Requirements: Update, Upgrade, Redesign – towards a Theory of Requirements Change*.
2006-13 Henk-Jan Lebbink (UU), *Dialogue and Decision Games for Information Exchanging Agents*.
2006-12 Bert Bongers (VU), *Interactivation – Towards an e-cology of people, our technological environment, and the arts*.
2006-11 Joeri van Ruth (UT), *Flattening Queries over Nested Data Types*.
2006-10 Ronny Siebes (VU), *Semantic Routing in Peer-to-Peer Systems*.
2006-09 Mohamed Wahdan (UM), *Automatic Formulation of the Auditor's Opinion*.
2006-08 Eelco Herder (UT), *Forward, Back and Home Again – Analyzing User Behavior on the Web*.
2006-07 Marko Smiljanic (UT), *XML schema matching – balancing efficiency and effectiveness by means of clustering*.
2006-06 Ziv Baida (VU), *Software-aided Service Bundling – Intelligent Methods & Tools for Graphical Service Modeling*.
2006-05 Cees Pierik (UU), *Validation Techniques for Object-Oriented Proof Outlines*.
2006-04 Marta Sabou (VU), *Building Web Service Ontologies*.
2006-03 Noor Christoph (UVA), *The role of metacognitive skills in learning to solve problems*.
2006-02 Cristina Chisalita (VU), *Contextual issues in the design and use of information technology in organizations*.
2006-01 Samuil Angelov (TUE), *Foundations of B2B Electronic Contracting*.
2005-21 Wijnand Derks (UT), *Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics*.
2005-20 Cristina Coteanu (UL), *Cyber Consumer Law, State of the Art and Perspectives*.
2005-19 Michel van Dartel (UM), *Situated Representation*.
2005-18 Danielle Sent (UU), *Test-selection strategies for probabilistic networks*.
2005-17 Boris Shishkov (TUD), *Software Specification Based on Re-usable Business Components*.
2005-16 Joris Graaumans (UU), *Usability of XML Query Languages*.
2005-15 Tibor Bosse (VU), *Analysis of the Dynamics of Cognitive Processes*.
2005-14 Borys Omelayenko (VU), *Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics*.
2005-13 Fred Hamburg (UL), *Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen*.
2005-12 Csaba Boer (EUR), *Distributed Simulation in Industry*.

2005-11 Elth Ogston (VU), *Agent Based Matchmaking and Clustering – A Decentralized Approach to Search*.
2005-10 Anders Bouwer (UVA), *Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments*.
2005-09 Jeen Broekstra (VU), *Storage, Querying and Inferencing for Semantic Web Languages*.
2005-08 Richard Vdovjak (TUE), *A Model-driven Approach for Building Distributed Ontology-based Web Applications*.
2005-07 Flavius Frasincar (TUE), *Hypermedia Presentation Generation for Semantic Web Information Systems*.
2005-06 Pieter Spronck (UM), *Adaptive Game AI*.
2005-05 Gabriel Infante-Lopez (UVA), *Two-Level Probabilistic Grammars for Natural Language Parsing*.
2005-04 Nirvana Meratnia (UT), *Towards Database Support for Moving Object data*.
2005-03 Franc Grootjen (RUN), *A Pragmatic Approach to the Conceptualisation of Language*.
2005-02 Erik van der Werf (UM)), *AI techniques for the game of Go*.
2005-01 Floor Verdenius (UVA), *Methodological Aspects of Designing Induction-Based Applications*.
2004-20 Madelon Evers (Nyenrode), *Learning from Design: facilitating multidisciplinary design teams*.
2004-19 Thijs Westerveld (UT), *Using generative probabilistic models for multimedia retrieval*.
2004-18 Vania Bessa Machado (UvA), *Supporting the Construction of Qualitative Knowledge Models*.
2004-17 Mark Winands (UM), *Informed Search in Complex Games*.
2004-16 Federico Divina (VU), *Hybrid Genetic Relational Search for Inductive Learning*.
2004-15 Arno Knobbe (UU), *Multi-Relational Data Mining*.
2004-14 Paul Harrenstein (UU), *Logic in Conflict. Logical Explorations in Strategic Equilibrium*.
2004-13 Wojciech Jamroga (UT), *Using Multiple Models of Reality: On Agents who Know how to Play*.
2004-12 The Duy Bui (UT), *Creating emotions and facial expressions for embodied agents*.
2004-11 Michel Klein (VU), *Change Management for Distributed Ontologies*.
2004-10 Suzanne Kabel (UVA), *Knowledge-rich indexing of learning-objects*.
2004-09 Martin Caminada (VU), *For the Sake of the Argument; explorations into argument-based reasoning*.
2004-08 Joop Verbeek (UM), *Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politiâle gegevensuitwisseling en digitale expertise*.
2004-07 Elise Boltjes (UM), *Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes*.
2004-06 Bart-Jan Hommes (TUD), *The Evaluation of Business Process Modeling Techniques*.
2004-05 Viara Popova (EUR), *Knowledge discovery and monotonicity*.
2004-04 Chris van Aart (UVA), *Organizational Principles for Multi-Agent Architectures*.
2004-03 Perry Groot (VU), *A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving*.
2004-02 Lai Xu (UvT), *Monitoring Multi-party Contracts for E-business*.
2004-01 Virginia Dignum (UU), *A Model for Organizational Interaction: Based on Agents, Founded in Logic*.
2003-18 Levente Kocsis (UM), *Learning Search Decisions*.
2003-17 David Jansen (UT), *Extensions of Statecharts with Probability, Time, and Stochastic Timing*.
2003-16 Menzo Windhouwer (CWI), *Feature Grammar Systems – Incremental Maintenance of Indexes to Digital Media Warehouses*.
2003-15 Mathijs de Weerdt (TUD), *Plan Merging in Multi-Agent Systems*.
2003-14 Stijn Hoppenbrouwers (KUN), *Freezing Language: Conceptualisation Processes across ICT-Supported Organisations*.
2003-13 Jeroen Donkers (UM), *Nosce Hostem – Searching with Opponent Models*.
2003-12 Roeland Ordelman (UT), *Dutch speech recognition in multimedia information retrieval*.
2003-11 Simon Keizer (UT), *Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks*.
2003-10 Andreas Lincke (UvT), *Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture*.
2003-09 Rens Kortmann (UM), *The resolution of visually guided behaviour*.
2003-08 Yongping Ran (UM), *Repair Based Scheduling*.
2003-07 Machiel Jansen (UvA), *Formal Explorations of Knowledge Intensive Tasks*.
2003-06 Boris van Schooten (UT), *Development and specification of virtual environments*.
2003-05 Jos Lehmann (UVA), *Causation in Artificial Intelligence and Law – A modelling approach*.
2003-04 Milan Petković (UT), *Content-Based Video Retrieval Supported by Database Technology*.
2003-03 Martijn Schuemie (TUD), *Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy*.
2003-02 Jan Broersen (VU), *Modal Action Logics for Reasoning About Reactive Systems*.
2003-01 Heiner Stuckenschmidt (VU), *Ontology-Based Information Sharing in Weakly Structured Environments*.
2002-17 Stefan Manegold (UVA), *Understanding, Modeling, and Improving Main-Memory Database Performance*.
2002-16 Pieter van Langen (VU), *The Anatomy of Design: Foundations, Models and Applications*.
2002-15 Rik Eshuis (UT), *Semantics and Verification of UML Activity Diagrams for Workflow Modelling*.
2002-14 Wieke de Vries (UU), *Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems*.
2002-13 Hongjing Wu (TUE), *A Reference Architecture for Adaptive Hypermedia Applications*.
2002-12 Albrecht Schmidt (UVA), *Processing XML in Database Systems*.
2002-11 Wouter C.A. Wijngaards (VU), *Agent Based Modelling of Dynamics: Biological and Organisational Applications*.
2002-10 Brian Sheppard (UM), *Towards Perfect Play of Scrabble*.
2002-09 Willem-Jan van den Heuvel (KUB), *Integrating Modern Business Applications with Objectified Legacy Systems*.
2002-08 Jaap Gordijn (VU), *Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas*.
2002-07 Peter Boncz (CWI), *Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications*.
2002-06 Laurens Mommers (UL), *Applied legal epistemology; Building a knowledge-based ontology of the legal domain*.
2002-05 Radu Serban (VU), *The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents*.
2002-04 Juan Roberto Castelo Valdueza (UU), *The Discrete Acyclic Digraph Markov Model in Data Mining*.
2002-03 Henk Ernst Blok (UT), *Database Optimization Aspects for Information Retrieval*.
2002-02 Roelof van Zwol (UT), *Modelling and searching web-based document collections*.
2002-01 Nico Lassing (VU), *Architecture-Level Modifiability Analysis*.
2001-11 Tom M. van Engers (VUA), *Knowledge Management: The Role of Mental Models in Business Systems Design*.

2001-10 Maarten Sierhuis (UvA), *Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and simulation language for work practice analysis and design*.
2001-09 Pieter Jan 't Hoen (RUL), *Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes*.
2001-08 Pascal van Eck (VU), *A Compositional Semantic Structure for Multi-Agent Systems Dynamics*.
2001-07 Bastiaan Schonhage (VU), *Diva: Architectural Perspectives on Information Visualization*.
2001-06 Martijn van Welie (VU), *Task-based User Interface Design*.
2001-05 Jacco van Ossenbruggen (VU), *Processing Structured Hypermedia: A Matter of Style*.
2001-04 Evgueni Smirnov (UM), *Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets*.
2001-03 Maarten van Someren (UvA), *Learning as problem solving*.
2001-02 Koen Hindriks (UU), *Agent Programming Languages: Programming with Mental Models*.
2001-01 Silja Renooij (UU), *Qualitative Approaches to Quantifying Probabilistic Networks*.
2000-11 Jonas Karlsson (CWI), *Scalable Distributed Data Structures for Database Management*.
2000-10 Niels Nes (CWI), *Image Database Management System Design Considerations, Algorithms and Architecture*.
2000-09 Florian Waas (CWI), *Principles of Probabilistic Query Optimization*.
2000-08 Veerle Coupé (EUR), *Sensitivity Analyis of Decision-Theoretic Networks*.
2000-07 Niels Peek (UU), *Decision-theoretic Planning of Clinical Patient Management*.
2000-06 Rogier van Eijk (UU), *Programming Languages for Agent Communication*.
2000-05 Ruud van der Pol (UM), *Knowledge-based Query Formulation in Information Retrieval*.
2000-04 Geert de Haan (VU), *ETAG, A Formal Model of Competence Knowledge for User Interface Design*.
2000-03 Carolien M.T. Metselaar (UVA), *Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectief*.
2000-02 Koen Holtman (TUE), *Prototyping of CMS Storage Management*.
2000-01 Frank Niessink (VU), *Perspectives on Improving Software Maintenance*.
1999-08 Jacques H.J. Lenting (UM), *Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation*.
1999-07 David Spelt (UT), *Verification support for object database design*.
1999-06 Niek J.E. Wijngaards (VU), *Re-design of compositional systems*.
1999-05 Aldo de Moor (KUB), *Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems*.
1999-04 Jacques Penders (UM), *The practical Art of Moving Physical Objects*.
1999-03 Don Beal (UM), *The Nature of Minimax Search*.
1999-02 Rob Potharst (EUR), *Classification using decision trees and neural nets*.
1999-01 Mark Sloof (VU), *Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products*.
1998-05 E.W. Oskamp (RUL), *Computerondersteuning bij Straftoemeting*.
1998-04 Dennis Breuker (UM), *Memory versus Search in Games*.
1998-03 Ans Steuten (TUD), *A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective*.
1998-02 Floris Wiesman (UM), *Information Retrieval by Graphically Browsing Meta-Information*.
1998-01 Johan van den Akker (CWI), *DEGAS – An Active, Temporal Database of Autonomous Objects*.