Université de Montréal

Concept Oriented Biomedical Information Retrieval

par
Wei SHEN

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en Informatique

August, 2015

RÉSUMÉ

Le domaine biomédical est probablement le domaine où il y a les ressources les plus riches. Dans ces ressources, on regroupe les différentes expressions exprimant un concept, et définit des relations entre les concepts. Ces ressources sont construites pour faciliter l'accès aux informations dans le domaine. On pense généralement que ces ressources sont utiles pour la recherche d'information biomédicale. Or, les résultats obtenus jusqu'à présent sont mitigés : dans certaines études, l'utilisation des concepts a pu augmenter la performance de recherche, mais dans d'autres études, on a plutôt observé des baisses de performance. Cependant, ces résultats restent difficilement comparables étant donné qu'ils ont été obtenus sur des collections différentes. Il reste encore une question ouverte si et comment ces ressources peuvent aider à améliorer la recherche d'information biomédicale. Dans ce mémoire, nous comparons les différentes approches basées sur des concepts dans un même cadre, notamment l'approche utilisant les identificateurs de concept comme unité de représentation, et l'approche utilisant des expressions synonymes pour étendre la requête initiale. En comparaison avec l'approche traditionnelle de "sac de mots", nos résultats d'expérimentation montrent que la première approche dégrade toujours la performance, mais la seconde approche peut améliorer la performance. En particulier, en appariant les expressions de concepts comme des syntagmes stricts ou flexibles, certaines méthodes peuvent apporter des améliorations significatives non seulement par rapport à la méthode de "sac de mots" de base, mais aussi par rapport à la méthode de Champ Aléatoire Markov (Markov Random Field) qui est une méthode de l'état de l'art dans le domaine. Ces résultats montrent que quand les concepts sont utilisés de façon appropriée, ils peuvent grandement contribuer à améliorer la performance de recherche d'information biomédicale. Nous avons participé au laboratoire d'évaluation ShARe/CLEF 2014 eHealth. Notre résultat était le meilleur parmi tous les systèmes participants.

Mots clés: UMLS, MetaMap, concept, recherche d'information biomédical, modèle de langue, expansion de requête, dépendance.

ABSTRACT

Health and biomedical area is probably the area where there are the richest domain resources. In these resources, different expressions are clustered into well defined concepts. They are designed to facilitate public access to the health information and are widely believed to be useful for biomedical information retrieval. However the results of previous works are highly mitigated: in some studies, concepts slightly improve the retrieval performance, while in some others degradations are observed. It is however difficult to compare the results directly due to the fact that they have been performed on different test collections. It is still unclear whether and how medical information retrieval can benefit from these knowledge resources. In this thesis we aim at comparing in the same framework two families of approaches to exploit concepts - using concept IDs as the representation units or using synonymous concept expressions to expand the original query. Compared to a traditional bag-of-words (BOW) baseline, our experiments on test collections show that concept IDs always degrades retrieval effectiveness, whereas the second approach can lead to some improvements. In particular, by matching the concept expressions as either strict or flexible phrases, some methods can lead to significant improvement over the BOW baseline and even over MRF model on most query sets. This study shows experimentally that when concepts are used in a suitable way, it can help improve the effectiveness of medical information retrieval. We participated at the ShARe/CLEF 2014 eHealth Evaluation Lab. Our result was the best among all the participating systems.

Keywords: UMLS, MetaMap, concept, medical information retrieval, language model, query expansion, dependency.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUI	Atom Unique Identifier
BOW	Bag Of Words
BOC	Bag Of Concepts
CLEF	Cross-Language Evaluation Forum
COLM	Concept Orient Information Retrieval
CUI	Concept Unique Identifier
EHR	Electronic Health Record
GO	Gene Ontology
HIPAA	Health Insurance Portability and Accountability Act
HMM	Hidden Markov Models
ICD	International Classification of Diseases
IR	Information Retrieval
LM	Language Model
LUI	Lexical Unique Identifier
MAP	Mean Average Precision
MeSH	Medical Subject Headings
MRF	Markov Random Field
NCBI	National Center for Biotechnology Information
NER	Named Entity Recognition
NLM	National Library of Medicine
POS	Part-of-Speech
RDF	Resource Description Framework

SCTID	SNOMED CT concept ID
SIGIR	Special Interest Group On Information Retrieval
SNOMED CT	Systematized Nomenclature of Medicine – Clinical Terms
SUI	String Unique Identifier
SVM	Support Vector Machine
SY	Asserted Synonyms
TREC	Text REtrieval Conference
UMLS	Unified Medical Language System

ACKNOWLEDGMENTS

I am extremely grateful to my supervisor, Jian-Yun NIE for his continual support and pertinent advice throughout this work.

This thesis is dedicated to my mother, ZhiPing WONG, and my precious Linqian FENG.

CHAPTER 1

INTRODUCTION

The volume of health and biomedical information is rapidly increasing. Some organizations such as *The National Library of Medicine (NLM)*, which is the world's largest biomedical library, has played an important role in providing access to biomedical and health information. Their public database, MEDLINE/PubMed containing over 24 million journal citations collected from 1946, is the most commonly used online scientific medical resource in the world. MedlinePlus [77] provides comprehensive and easy-to-read consumer health information to laypeople to search and understand their physical condition.

The Health and biomedical information retrieval (IR) is becoming more and more important not only for the health care professional, but also for lay people. Nearly 70% of search engine users in the U.S. have used web search to find information about a specific disease [36]. One of the most commonly used search engine PubMed has received 775,504,557 queries in 2008.

However, biomedical IR is a difficult task. Medical terminology always contains many synonyms, aliases or abbreviations. For example, to describe the symptom "*headache*", some professional physicians would use "*cephalodynia*", while laypeople may talk about "*head pain*" or even "*pain in head*". In addition, medical concepts are often described at different levels. For example, a patient wondering the effect of the "*antibiotics*" on his disease may miss a lot of useful information, because a large number of articles is talking about specific antibiotics such as "*Garamycin*", "*Kantrex*", "*Vancomycin*", but do not contain the word "antibiotics". Sometimes only the brand names are mentioned. This makes the expressions in documents hard to match with users' queries.

In order to standardize the expression in biomedical area, domain resources and thesauri, such as MeSH [60], Metathesaurus [12], SNOMED CT [89] are built, in which, the alias, acronyms or lexical variants of the biomedical terminologies are clustered into

controlled vocabularies and assigned a unique ID. Symptoms, diseases and their related therapies or drugs are connected to each other with semantic relations. To benefit from this kind of resources, previous works have explored different approaches, trying to take advantage of the unique concept ID or of different synonym and related concept expressions. However, the results were inconsistent. Some of them reached slight improvements while in some others degradations are observed.

In this thesis, we systematically explore the performances of different concept-based approaches on the same platform. Our goal is to test the true effectiveness of different methods and find out an appropriate way to benefit from the concepts in knowledge resource. We use UMLS Metathesaurus [12] and MetaMap [4] to identify concepts from documents and queries. The identified concepts are used in different ways - as the basic semantic representation units, as phrases or as independent words. In addition, we proposed a hybrid method to combine different concept expressions together.

Our general conclusion is that concepts IDs are too rigid to be representation units of document contants and queries. It is more reasonable to use concept expression as phrases. Our experiments were implemented on three different datasets: OHSUMED, CLEF 2013 and CLEF 2014 collections. Finally, the overall best result was achieved by the hybrid phrase combination method, which led to significant improvements over BOW baseline and traditional MRF model.

In addition, we participated in the CLEF/ShARe task on biomedical IR in 2014. Our result was ranked the best among 62 runs from 14 groups, and it outperforms the state-of-the-art in this domain.

The rest of the thesis is laid out as follows. In chapter 2, we first introduce what the biomedical information retrieval task is. After summarizing some well known medical thesauri and the concept mapping tools, we will review the previous works on concept-based medical IR. Chapter 3 mainly explains our approaches. Then in chapter 4 we present our experiments. The last chapter (Chapter 5) concludes the thesis and gives some possible directions of future work.

CHAPTER 2

BIOMEDICAL INFORMATION RETRIEVAL

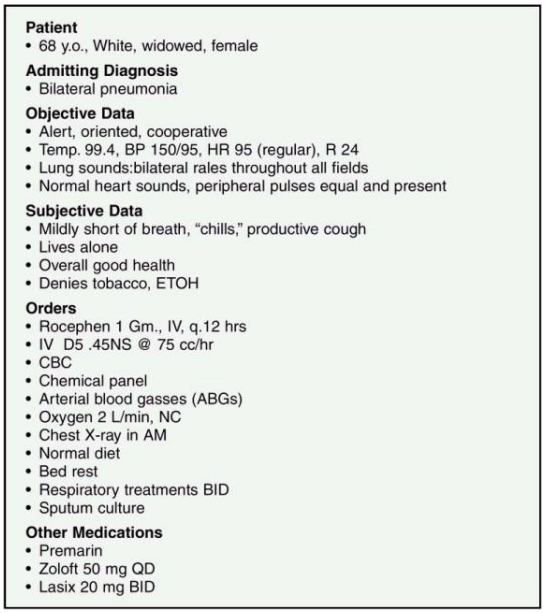
In this chapter, we explain in detail what the *Biomedical Information Retrieval* is, what the main difference is between this task and traditional IR, and what the recent progresses are. In the first section, we will show how medical data look like, and explain the main task of biomedical information retrieval. Section 2.2 is an introduction to the existing knowledge resources and their supporting knowledge extractors in medical area. And then in Section 2.3, we will review the previous works on concept-based biomedical IR.

2.1 Medical Data and Medical Information Retrieval Task

Medical Record (also called Health Record) is the first type of *Medical Data*. It occupies a large proportion of the Medical Data we have. A Medical Record contains the systematic documentation of a single patient's medical history and care across time. It can include a variety of types of "records": demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information created in healthcare organizations such as a hospital or physician's office [1]. Traditionally, medical records were written on paper and maintained in folders and usually housed at the governments, insurance companies and other large hospital or medical institutions. Figure 2.1 shows the typical data in medical record. Medical records are well structured, including some professional terms such as *Bilateral Pneumonia* or even some puzzling string/abbreviation like *Temp.99.4, BP 150/95, HR 95 (regular), R 24*.

In recent years, alone with the popularity of the *Electronic Medical Records* (EMR), the volume of medical record is exponentially increasing. Electronic health record

1. "Mobile Tech Contributions to Healthcare and Patient Experience". Top Mobile Trends. Retrieved 29 May 2014.



Patient
- 68 y.o., White, widowed, female

Admitting Diagnosis
- Bilateral pneumonia

Objective Data
- Alert, oriented, cooperative
- Temp. 99.4, BP 150/95, HR 95 (regular), R 24
- Lung sounds:bilateral rales throughout all fields
- Normal heart sounds, peripheral pulses equal and present

Subjective Data
- Mildly short of breath, "chills," productive cough
- Lives alone
- Overall good health
- Denies tobacco, ETOH

Orders
- Rocephen 1 Gm., IV, q.12 hrs
- IV D5 .45NS @ 75 cc/hr
- CBC
- Chemical panel
- Arterial blood gasses (ABGs)
- Oxygen 2 L/min, NC
- Chest X-ray in AM
- Normal diet
- Bed rest
- Respiratory treatments BID
- Sputum culture

Other Medications
- Premarin
- Zoloft 50 mg QD
- Lasix 20 mg BID

Figure 2.1: A typical patient record [53].

(EHR) system is designed to capture and represent the above information of the patient. Figure 2.2 shows an interface of electronic medical record, which is fully structured. The user is enforced to use a set of built-in controlled vocabularies to fill out the electronic forms.

However, because of the private nature of the individual medical records, their utilization is strictly limited by Health Insurance Portability and Accountability Act (HIPAA). Thus the medical records are normally only shared for enterprise-wide use. Only a very small part has been released to open access for research purpose. For example, the test document collection for TREC Medical Records track is a set of medical records made available for research use through the University of Pittsburgh BLULab NLP Repository. The repository contains one month of reports from multiple hospitals, and includes nine types of reports: Radiology Reports, History and Physicals, Consultation Reports, etc [99]. Before being released, all records are de-identified. After one year of TREC experiments, the data is no longer available. Therefore, while there are increasing needs to deal with patient's medical records, there is not publicly available data for researchers to perform experiments.

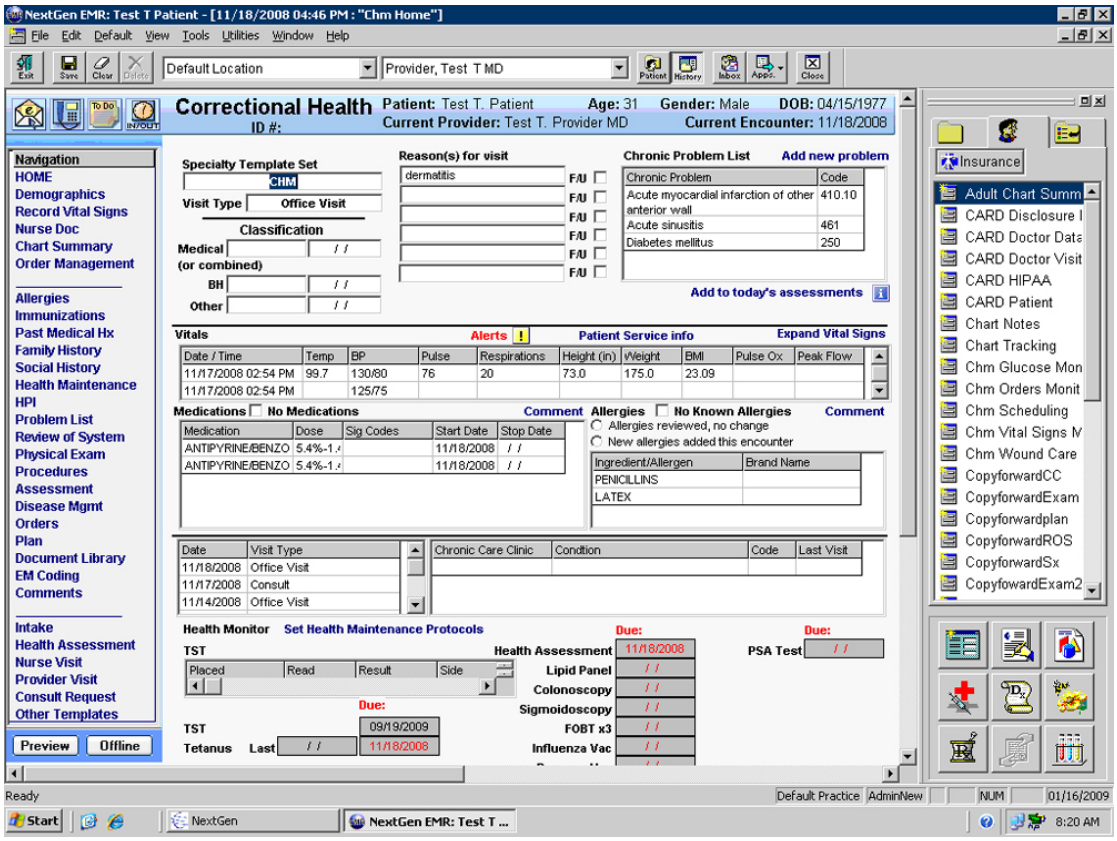


Figure 2.2: A practical Electronic Medical Record (Produced by NextGen EHR Software [63])

Another type of medical information refer to the biomedical literature in the fields of biomedicine and health, covering portions of the life sciences, behavioral sciences, chemical sciences, and bioengineering [32]. In 2009, Hersh [45] introduced a new broader term "*Biomedical and Health Informatics*" which extends the healthcare information to the public health and biomedical literature. Compared with the private medical record, this kind of data is much easier to get via some open accesses. MEDLINE is the U.S. National Library of Medicine (NLM) bibliographic database that contains over 24 million references to citations from over 5,600 worldwide life science journals and online books in about 40 languages [33]. PubMed [33] provides an online open access to MEDLINE database. In addition to the abstract, the citations come with links to its full-text web pages.

Figure 2.3 shows a top ranked citation for the query "*headache*". Title, author and abstract compose the main part of the citation. In the upper right corner, PUBMED provides the full text link to the JAMA [79] database. In this thesis, we will focus on the second type of data, the retrieval in biomedical literature.

2.2 Biomedical Knowledge Resource

Different from traditional ad hoc information retrieval task, medical IR is faced with an important challenge of vocabulary variation: although we have a set of well defined concepts, a concept can be described in many different ways with different terms and expressions. For example, the symptom *headache* or its acronym *HA* is commonly used in medical records. But another professional name: *Cephalgia* could be used in an Electronic Medical Records as the standard name. However, a layperson could describe it as *pain in head*. The large number of variations in concept expressions makes it hard to retrieve the right documents with a query.

In addition, some specific use of health data requires retrieval on the semantic level. A Clinical Decision Support System should generate case-specific advices on diagnosis according to patient's symptoms. And for patients, one would like to find out a rational treatment and care. The query may not contain the drug or operation that they are

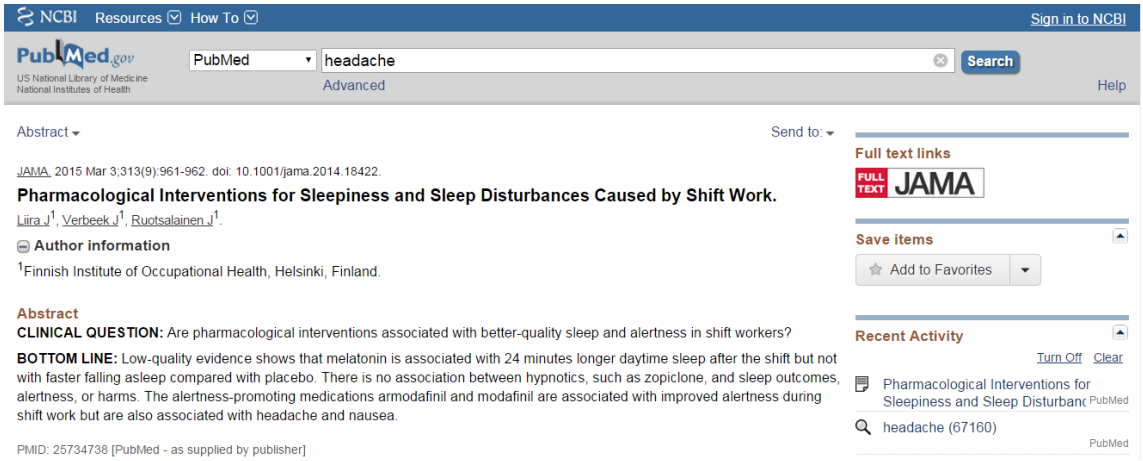


Figure 2.3: The top ranked document in PUBMED search engine with the query *headache* [73].

looking for. It requires inferring implicitly related drugs or treatments.

Under this condition, in order to unify the terminology expression and to construct a well-structured knowledge network, from the 1980s, some well-designed medical thesauri have been manually constructed. Nowadays, the U.S. National Library of Medicine[78] (NLM) maintains the largest source vocabulary Unified Medical Language System [12] (UMLS), including hundreds of independent thesaurus, such as the well known MeSH [60] (Medical Subject Headings), SNOMED [89] (Systematized Nomenclature of Medicine – Clinical Terms), ICD-9 [35] (International Classification of Diseases). In this section, we give a brief introduction to them. Especially, we focus on how the medical vocabularies are organized in different thesauri.

2.2.1 Thesauri in medical area

Regardless the large number of different thesauri and the different domains they focus on, a medical thesaurus is essentially a set of controlled vocabularies in a hierarchical structure. In this section, we show these fundamental structures of a typical thesaurus.

2.2.1.1 Concept

The two main goals of a medical thesaurus are to standardize the expressions of biomedical terminologies, and then to construct a structured vocabulary network. To do that, we firstly need to identify essential units which are unique and distinct from each other. Thesauri usually cluster terms by their meaning. A meaning is called a **Concept** and is assigned a unique ID. The terms which are strictly synonymous with each other, as well as their lexical variants are defined as the entries of that concept, and one of them will be denoted as the preferred name. For example, *Hyptertension* and *High blood pressure* refer to the same disease, which is defined as a *concept* identified by ID *C0020538* in UMLS Metathesaurus (2.4). *Hypertensive disease* is denoted as its standard expression (preferred name). Concepts and their corresponding expressions can be organized with some specific structure. For example, MeSH uses **Descriptor/Concept/Term Entry** structure, and UMLS Metathesaurus uses **Concept/Terms/String/Atom** structure for UMLS knowledge (They will be detailed in the next section).

2.2.1.2 Hierarchical relationships

Concepts are unique and distinct from each other, but don't exist in isolation. Similar to *"Class"* in Objet Oriented Programming Language, *"Concepts"* are always classified into a hierarchical structure. In a *Tree* structure, general concepts can be divided into more specific ones. For example, *hypertension* is a type of *Vascular Diseases*, and itself can be further divided into more specific types of hypertension, such as *malignant*

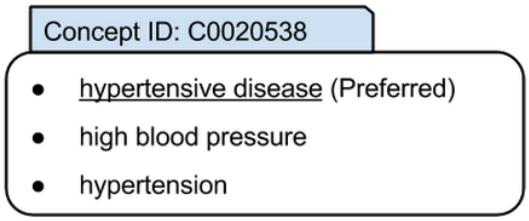


Figure 2.4: A concept can be expressed with different terms.

hypertension or *pregnancy-induced hypertension*. A broader concept is a **Hypernym**, denoted by *Parent* or *IS-A* relationship, while a narrower one is called **Hyponym** and comes with *Child* relation (2.5).

In terms of information retrieval, hierarchical relationship is useful to enhance the expression coverage. For example, suppose that we are looking for "*the treatment of hypertension*". A document talking about the treatment of "*malignant hypertension*" is probably relevant.

2.2.1.3 Semantic relationships

In addition to lexical and hierarchical relations of synonymy, hypernymy, hyponymy, many knowledge resources also connect concepts by semantic links to describe the relation or interaction between them. For example, the following assertion "*Influenza is caused by influenza viruses*" expresses the cause of the disease. As shown in Figure 2.6, it can be expressed by RDF (Resource Description Framework) [55] triples: {*influenza virus - cause - influenza*}. In which the two concepts, the subject "*influenza virus*" and the object "*influenza*" are connected by relation "*cause*". A semantic network could be helpful for some advanced usage. When medical decision support system needs to find out a proper treatment for a specific disease or symptom, a traditional retrieval model relies on terms is not sufficient. One may need to utilize the semantically related concepts.

2.2.2 Some Well Known Knowledge Resources

Obviously, not all thesauri are designed exactly in the same way. Each of them can have some particular features, which will lead to different performance when used in information retrieve. In this section, we will show different aspects of some commonly used thesauri.

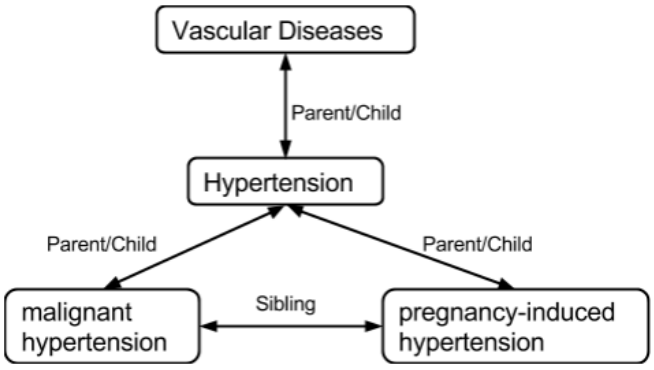


Figure 2.5: Concepts are classified into several of categories and are connected by hierarchical relationships.

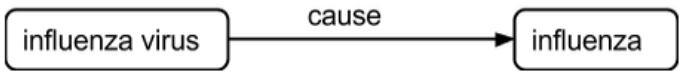


Figure 2.6: Semantic relationships represented by RDF triples.

2.2.2.1 MeSH - Medical Subject Headings

The **Medical Subject Headings (MeSH)** thesaurus is a controlled vocabulary produced by the **NLM** (National Library of Medicine). As the name suggests, it was designed to index citations in NLM's MEDLINE database, for cataloging of publications, and other databases. The important basic type of MeSH Records is the **Descriptor** (Also known as *Main Headings*). Different from the *Concept* introduced above, each MeSH Record is created in ***Descriptor/Concept/Term*** structure (as shown in Figure 2.7). The *Descriptor* is composed of some distinct but similar concepts referring to a same topic. For example, three different concepts *Headache*, *Bilateral Headache* and *Generalized Headache* are grouped under the same descriptor: "*Headache*".

All MeSH descriptors are organized into 16 root categories: *anatomic terms*, *organisms*, *diseases*, etc. Each of the root categories as well as their subcategories can be further divided into subcategories, which build up a *MeSH Tree* up to twelve levels. Each descriptor is allowed to appear in more than one tree. Figure 2.8 shows that the descriptor *Headache* is located in two trees. {*Pathological Conditions –> Signs and Symptoms –> Neurologic Manifestations –> Headache*} and {*Nervous System Dis-*

- **Headache** [Descriptor]
  - *Headache* [Concept, Preferred]
    - `Headache` [Term, Preferred]
    - `Cephalalgia` [Term]
    - `Cephalgia` [Term]
    - `Cephalodynia` [Term]
    - `Cranial Pain` [Term]
    - `Head Pain` [Term]
  - *Bilateral Headache* [Concept, Narrower]
    - `Bilateral Headache` [Term, Preferred]
  - *Generalized Headache* [Concept, Narrower]
    - `Generalized Headache` [Concept, Narrower]
  - *Ocular Headache* [Concept, Narrower]
    - `Ocular Headache` [Concept, Narrower]

Figure 2.7: The Descriptor/Concept/Term Structure of a MeSH Record "Headache".

eases –> Neurologic Manifestations –> Pain –> Headache}. This gives rise to the problem of ambiguity.

Currently, various online systems provide access to MeSH. The *MeSH Browser* [69], contains the complete contents of the vocabulary; the *MeSH Entrez* databases, which provides assistance using MeSH vocabulary in searching MEDLINE/PubMed.

A possible limitation of MeSH for the purpose of medical IR is that it only contains hierarchical relations between concepts, other types of semantic relation, such as causal relation, are not encoded. This may limit the coverage of the retrieval process: using MeSH, one can retrieve documents about more specific concepts, but not about related concepts.

2.2.2.2 SNOMED CT

SNOMED CT [49] (Systematized Nomenclature of Medicine – Clinical Terms) is another collection of medical controlled vocabulary, created by the College of American Pathologists. Different from MeSH, which mainly relies on hierarchy relationships, SNOMED contains extended lexical and semantic relations. Apart from *"IS-A"* subtype relation, concepts are inter-connected with each other by a set of object-attribute-value triples which express their defining characteristics. As showed in Figure 2.9, concept

```
- Pathological Conditions, Signs and Symptoms [C23]
  - Signs and Symptoms [C23.888]
    - Neurologic Manifestations [C23.888.592]
      - Pain [C23.888.592.612]
      - Back Pain [C23.888.592.612.107]
      - Chronic Pain [C23.888.592.612.274]
      - Facial Pain [C23.888.592.612.330]
      - Headache [C23.888.592.612.441]
      - Slit Ventricle Syndrome [C23.888.592.612.441.500]
      - Labor Pain [C23.888.592.612.451]
      - ...  ...
- Nervous System Diseases [C10]
  - Neurologic Manifestations [C10.597]
    - Pain [C10.597.617]
      - Acute Pain [C10.597.617.088]
      - Breakthrough Pain [C10.597.617.178]
      - Mastodynia [C10.597.617.205]
      - Musculoskeletal Pain [C10.597.617.231]
      - Back Pain [C10.597.617.232]
      - Chronic Pain [C10.597.617.258]
      - Facial Pain [C10.597.617.364]
      - Arrow pointing to current tree node
      - Headache [C10.597.617.470]
      - ...  ...
```

Figure 2.8: The descriptor "*Headache*" is located in more than one place in MeSH Trees.

"*viral pneumonia*", has two hypernyms "*Infective pneumonia*" and "*Viral lower respiratory infection*", which are linked by *IS-A* relation. In addition, the "*Causitive agent*" relationship denotes that the *infective pneumonia* is caused by *virus*, and "*Finding site*" indicates the affected organ is *Lung*. Relationships themselves are defined as concepts in SNOMED CT thesaurus and are unidirectional. Figure 2.10 lists the relationship triples in which concept "*Viral lower respiratory infection*" is involved.

Like most of the thesauri, the *Concept* is the representational unit identified by **SCTID** (SNOMED CT concept ID). Concept expressions are declared and classified in description table by *DescriptionType* indicator, respectively corresponding to "preferred" description (term), fully specified name and synonym (alternate). As shown in Figure 2.11, the concept "*Viral lower respiratory infection*" has two different expressions, one for fully specified name and another is assigned as synonym.

Among the existing thesauri, SNOMED contains the richest semantic relations. However, there is no reliable tool to identify SNOMED concepts from raw text. A common practice is to identify first UMLS concepts, which are then mapped to SNOMED CT concepts. Koopman et al. [56] reported that mapping between terminologies may result in a loss in meaning, because certain UMLS concepts have no equivalent in SNOMED CT. Currently, some online and offline browsers, such as *IHTSDO SNOMED CT Browser*[43], are available. But their accuracy has not been systematically evaluated.

Figure 2.9: Relationship assigned as attribute of concept "*Viral lower respiratory infection*"(Produced by IHTSDO SNOMED CT Browser [43]). Apart from the subtype hierarchy, *IS-A* link, SNOMED CT consist also semantic relationship.

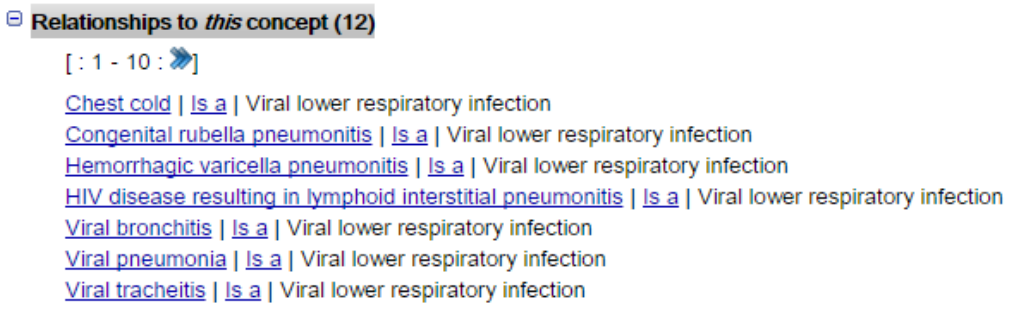


Figure 2.10: The SNOMED CT relationships in which concept "*Viral lower respiratory infection*" is involved (Produced by IHTSDO SNOMED CT Browser [43]).

Concept: `Viral lower respiratory infection`
SCTID: `312134000`

Descriptions (2)

Id	Description	Type	Status
708770018	Viral lower respiratory infection (disorder)	Fully specified name	Active
455799012	Viral lower respiratory infection	Synonym	Active

Figure 2.11: Two parallel expressions of concept "*Viral lower respiratory infection*"in SNOMED CT thesauri (Produced by IHTSDO SNOMED CT Browser [43]).

2.2.2.3 Other specific thesauri

In addition to the large and complete thesauri created for the usage on the entire medicine area, there are some smaller and domain-specific thesauri which only provide knowledge in a specific area. For example, the ***ICD*** *(International Classification of Diseases)* is currently the most authoritative and widely used classification system for diseases in the world. As shown in Figure 2.12, diseases are hierarchically organized. For example, *Viral Pneumonia* is further divided into sub-categories. Compared with the same entry in MeSH and SNOMED CT, the vocabulary *viral pneumonia* in ICD is more comprehensive. The Ninth Revision included an optional alternative method of classifying diagnostic statements, including information about both an underlying general disease and a manifestation in a particular organ or site. Another well known example is the **Entrez Gene** [68], which is a portal to gene-specific content. MeSH and Entrez Gene both act as a part of **NCBI** (The National Center for Biotechnology

Viral pneumonia 480- > Inflammation of the lung parenchyma that is caused by a viral infection. Pneumonia (inflammation of the lungs) caused by a virus.

- 480 **Viral pneumonia**
  - 480.0 Pneumonia due to adenovirus
  - 480.1 Pneumonia due to respiratory syncytial virus
  - 480.2 Pneumonia due to parainfluenza virus
  - 480.3 Pneumonia due to SARS-associated coronavirus
  - 480.8 Pneumonia due to other virus not elsewhere classified
  - 480.9 Viral pneumonia, unspecified

Figure 2.12: The disease *Viral Pneumonia* is further divided by its different causes in ICD-9, which is the most authority thesaurus of the classification of diseases.

Information) database. Different from general biomedical terminologies, gene and protein names can be highly variable, especially their acronyms and abbreviations are commonly shared between some totally different concepts. Their recognition is far from trivial and could result in some inappropriate expansion due to lexical ambiguities [7]. Obviously, a general thesaurus can hardly tackle this non trivial task. For example, the gene name abbreviation "*stx2*" can represent 104 different genomes of different species (partially listed in Figure 2.13). Within **GO** (Gene Ontology) [23], each gene is located in a large gene-centered semantic network. The related function, process, and component are connected to each gene (Figure 2.14). Stokes [92] summarizes some other similar gene symbol resources such as *ADAM* [108], *HUGO* [30], *OMIM* [13], *UniProt* [8], and so on. Another useful information could be the **GeneRIFs**, which allows scientists to add some sentences and published papers to describe a related function of a gene to enrich the semantic context of gene.

2.2.2.4 UMLS

The **UMLS** (Unified Medical Language System) was built to help computer systems to "understand" the language of biomedicine. It is composed of three main knowledge sources: *Metathesaurus*, *Semantic Network*, and *SPECIALIST Lexicon*.

Among them, the *Metathesaurus* is the main part, which is known as the largest

Gene Name	Gene ID	Description
Stx2	ID: 13852	syntax in 2 [Mus musculus (house mouse)] Chromosome 5
STX2	ID: 2054	syntax in 2 [Homo sapiens (human)] Chromosome 12
Stx2	ID: 25130	syntax in 2 [Rattus norvegicus (Norway rat)] Chromosome 12
Stx1b	ID: 24923	syntax in 1B [Rattus norvegicus (Norway rat)] Chromosome 1
stx2	ID: 780101	syntax in 2 [Xenopus (Silurana) tropicalis (western clawed frog)]
Syn16	ID: 811466	SNARE protein, putative [Plasmodium falciparum 3D7] Chromosome 12
...	...	...

Figure 2.13: In *Entrez Gene*, 104 genomes of different species shared a same abbreviation "*stx2*".

Function	Evidence Code	Pubs
SNAP receptor activity	IEA	
SNARE binding	ISO	
calcium-dependent protein binding	IPI	PubMed
protein binding	IPI	PubMed

Process	Evidence Code	Pubs
acrosome reaction	IDA	PubMed
cell differentiation	IDA	PubMed
digestive tract morphogenesis	ISO	
embryo implantation	IC	PubMed
epithelial cell differentiation	ISO	
intracellular protein transport	IEA	
microvillus assembly	ISO	
regulation of blood coagulation	ISO	
regulation of gene expression	ISO	
vesicle-mediated transport	IEA	

Component	Evidence Code	Pubs
basolateral plasma membrane	ISO	
cell surface	IDA	PubMed
cell-cell junction	IDA	PubMed
cytoplasmic vesicle	IDA	PubMed
extracellular vesicular exosome	ISO	
integral component of membrane	IEA	
membrane	IEA	
membrane raft	IDA	PubMed

Figure 2.14: For example, one of the well known Gene Ontology knowledge database: UniProt Knowledgebase (UniProtKB) [29] can provide related function, process, and component of a given protein (protein *syntaxin 2* [34] as example).

multi-lingual biomedical vocabulary database. It merges hundreds of different knowledge resources including the thesauri introduced above, MeSH, SNOMED CT, ICD, etc. The Semantic Network provides a consistent categorization of all concepts represented in the Metathesaurus and provides a set of useful relationships between these concepts. The main type of record in Metathesaurus is *Concept*, which is placed in the ***Concept/Lexical/String/Atom*** structure (See Figure 2.15).

Concept Unique Identifiers (CUI) A key goal of Metathesaurus construction is to link all names from hundreds of source vocabularies that have the same meaning. Each of them is assigned a unique ID: *CUI* starting with a letter C and followed by seven numbers. As shown in the example (Figure 2.15), the concept "*Headache*" is identified as *C0018681*.

Lexical Unique Identifiers (LUI) Sometimes a concept can be expressed by more than one term which are synonyms to each other. As example (Figure 2.15) concept "*Headache*" contains three synonyms: "*headache*", "*cephalalgia*" and an abbreviation "*ha*". Each of them is identified by *LUI* starting with letter *L*.

String Unique Identifiers (SUI) In natural language, a term can have different variations in character set, upper-lower case, or punctuation difference. Each of them is called a *String*, identified by *String Unique Identifiers* (SUI). In Figure 2.15, "*Headache*" and "*headaches*" are different variants of the same term "*headache*". On the other hand, a single string can correspond to more than one concept.

Atom Unique Identifiers (AUI) Considering that Metathesaurus is composed of hundreds of source vocabularies, a string may be included in more than one thesaurus, and its occurrence in each source vocabulary is assigned a unique atom identifier (AUI). As shown in Figure 2.15, the string *"Headache"* is included in both SNOMED CT and MeSH vocabulary, and is assigned two different AUI "*A2882187*" and "*A0066000*".

Metathesaurus also identifies useful relationships between concepts, but different from SNOMED CT, which mainly focuses on constructing a biomedical semantic network, Metathesaurus contains only hypernyms (PAR), hyponyms (CHD), synonyms (SY) and co-occurrence terms (stored in MRCOC.RRF data file). As shown in Figure 2.16, "*headache*" is a hyponym of "*Pain*" and "*Pain of head and neck region*" and

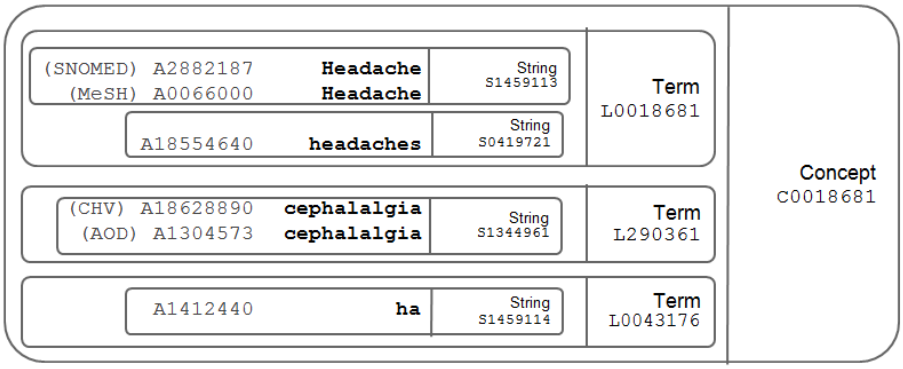


Figure 2.15: In Metathesaurus an expression is defined in 4 level: `Concept`, `Term`, `String` and `Atom` which is respectively identified by `AUI`, `SUI`, `LUI` and `CUI`.

a hypernym of "*Sinus headache*" and so on. All these concepts are annotated as the "*Sign or Symptom*" semantic type, which is described in more detail in UMLS Semantic Network [74].

2.3 Knowledge Extraction

Medical thesauri can be seen as a static dictionary of concepts. Before being used in a retrieve model, one has worked on extracting them from raw text, which is called *Biomedical Text Mining (BTM)* Task [114][21][58][88]. Depending on different targets, it can be roughly divided into three sub-tasks: *Concepts Extraction*, *Relations Extraction* [41][47][98] and *Event Extraction* [51]. For the purpose of medical IR, the most important task is concept identification, which is also the basis for the other two tasks [24].

Different from the traditional Named Entity Recognition (NER) task, which can rely on rule-based approach (for example, morphological analysis [38][2], lexical pattern [37]) and statistique/machine learning techniques (HMM [107][22], SVM [50]), concept mapping in medical texts usually uses straightforward dictionary look-up strategy combined with a series of NLP techniques such as parsing (chunking) [96], lexical variation [65] and disambiguation [86], etc. Aronson [5] concludes that the overall perfor-

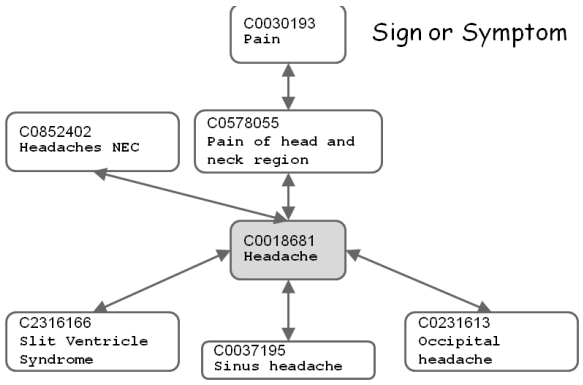


Figure 2.16: In UMLS, *headache* is defined as the hyponym of *Pain* and *Pain of head and neck region* and hypernym of *Sinus headache* and so on. All these concepts are annotated as the *Sign or Symptom* semantic type.

mance of mapping tools depends on how well these NLP problems has been solved. As introduced above, the structure of each thesauri is unique, a concept mapping tool is normally designed for a specific thesaurus, and released as a supporting module of it. The National Library of Medicine maintains the official online browsers of three above-mentioned thesaurus *MeSH* [70], *SNOMED CT* [75], and *Metathesaurus*[71]. There are also some well-performing offline programs available. One of the state-of-the-art systems, **MetaMap** [4], is embedded in the UMLS Metathesaurus, which achieved 84% in precision and 70% in recall [82]. Recently there have been some tools such as University of Michigan's **MGREP** [91] that also performs concept recognition. Stewart [91] reported that the main weakness of MetaMap is that chunking and variants generation are both time consuming tasks. MGREP works faster by generating directly the variants of single words, which will finally be combined with word order permutation. After a comprehensive comparison between these two tools, they found that MGREP has extremely fast execution speed, but fewer concepts are recognized.

There are several tools for specific uses. For example, RxNorm provides normalized names for clinical drugs by using the terminology National Drug File [15] - Reference Terminology (NDF-RT) from the Veterans Health Administration. NCBI PubMed Translator is another tool in charge of converting the original user queries into the corresponding MeSH heading for PubMed search engine.

Considering that MetaMap is the most widely used UMLS Metathesaurus mapping tool and results in the best mapping accuracy, in order to make our experiment comparable with most of the previous work, in our work, we will use MetaMap to recognize concepts from texts and queries. In the next section, we will briefly explain how concept are mapped by MetaMap.

2.3.1 MetaMap

MetaMap is a highly configurable program developed by Aronson [4] at the National Library of Medicine (NLM). It uses a knowledge-intensive approach together with some shallow morphological rules to map biomedical text to the UMLS Metathesaurus concept. It is widely regarded as the most advanced tool for this task.

The main precesses of MetaMap are as follows. A document will be cut into a list of sentences. Each of them will be further parsed by the SPECIALIST minimal commitment parser [66] which chunks the sentence into some short noun phrases and assigns a shallow syntactic tag for each. This step is called *Chunking* [96]. For example (Figure 2.17), given an input "*cerebral edema secondary infection diagnosis treatment*", two noun phrases "*cerebral edema secondary infection*" and "*diagnosis treatment*" can be detected, and each of them will be annotated by the *POS* (Part-of-Speech) tags.

Then metamap will conduct a lexical lookup within the phrases to find the longest spanning terms from the SPECIALIST Lexicon. For example, the noun phrase

Input:
```
cerebral edema secondary infection diagnosis
treatment
```
Output:
```
[mod(cerebral edema), head(secondary infection)],
[mod(diagnosis), head(treatment)].
```

Figure 2.17: SPECIALIST minimal commitment parser [66] chunks the sentence into two short noun phrases and assigns shallow syntactic tags for each. "*head*" refers to the core word of the noun phrase and "*mod*" means the modifier.
.

"cerebral edema secondary infection" is composed of two terms "cerebral edema" and "secondary infection". And the "diagnosis treatment" will be divided into "diagnosis" and "treatment".

The next step is to generate the variants for each extracted SPECIALIST term. A variant consists of a phrase of one or more words, together with all its acronyms, abbreviations, synonyms, derivational variants, inflectional and spelling variants [3]. The generation of abbreviations and synonyms is based on the knowledge in SPECIALIST Lexicon while the derivation variants and inflections are generated by handcrafted rules. During the variant generation process, each variant is assigned a variant distance score according to its history which records how it was created. Different type of variants correspond to different distance: 0 for p (spelling variants), 1 for i (inflections), 2 for s (synonyms) and for a (acronyms), 3 for d (derivational variants). For example, *Inflection*=1, *Synonym*=2, *Derivational variants*=3. The smaller the score is, the more semantically related the variant is. For example, in Figure 2.18, the synonym of "*treatment*" "*therapy*" is assigned a distance score of 2, and "*Tx*" is 4 and the history code "*sa*" means that it is a abbreviation of a synonym "*therapy*" of the original term "*treatment*".

Then as shown in Figure 2.19, these inferred variants are used to match the concept string in Metathesaurus. Once one of the concept string in Metathesaurus is matched with the variant, the CUI of this concept will be added to the *Candidate set*. In our example, term "*treatment*" corresponds to five different concepts "*C0039798 [therapeutic aspects]*", "*C1522326 [treating]*", "*C0087111 [Therapeutic procedure]*", "*C1533734 [Administration procedure]*" and "*C1705169 [Biomaterial Treatment]*" in UMLS Metathesaurus.

The concepts in the candidate list will be further evaluated and ranked according to a weighted average of the following four features [4],

- centrality (involvement of the head)
- variation (an average of inverse distance score)
- coverage
- cohesiveness

Finally the top ranked concept will be returned as the mapping result. Figure 2.20

```
- treatment [noun], 0= “”
  - therapy [noun], 2=“s”
    - Tx [noun], 4=“sa”
    - TH [noun], 4=“sa”
    - therapeutic [adj], 5=“sd”
      - Therapeutic [adj], 6=“sdi”
  - treat [verb], 3=“d”
    - treating [verb], 4=“di”
- ...
```

Figure 2.18: The variant generation process. Each variant is assigned a variant distance score according to its history which records how it was created. Different type of variants correspond to different distance: 0 for p (spelling variants), 1 for i (inflections), 2 for s (synonyms) and for a (acronyms), 3 for d (derivational variants).
.

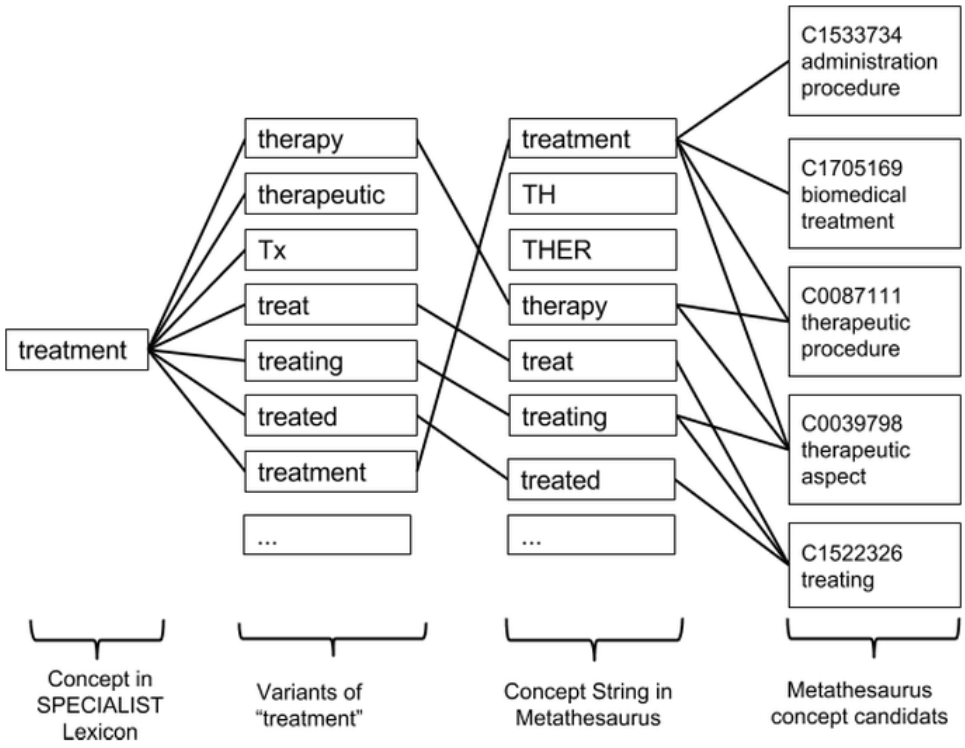


Figure 2.19: The variants of term “treatment” are used to match the concept strings in UMSL Metathesaurus. The matched strings belong to five different concepts.

gives an example of output of MetaMap on "treatment".

Considering that only the surface form similarity is used in MetaMap's candidate evaluation process, one may expect that more complex variations (e.g. those involving a large change in syntax) cannot be recognized. In addition, the simple dictionary look-up is unable to solve the problem of ambiguity - a term or expression corresponds to more than one concept. For the term "*treatment*" in our example (Figure 2.18 and Figure 2.19), in most of time MetaMap mapped it to concept "*C0039798 [therapeutic aspects]*" but we do observe that in some documents, the other four were selected. The selection of a concept for a term depends on the context in which the term is used. However, not all the selections are correct. This can lead to inconsistent concept mapping, i.e. a term is mapped to several concepts, even though the intended meaning is the same. Although MetaMap remains the current state-of-the-art for UMLS concept mapping, its mapping accuracy is far from perfect. In the rest of this thesis, our experiments will show that the concept mapping accuracy can largely affect the retrieve performance, which is consistent with the previous observations.

2.4 Concept-based Medical Information Retrieval

In the last two decades, researchers have constantly tried to find some proper ways to do medical IR to benefit from the knowledge resources. Recently, in 2011 and 2012, the *Text REtrieval Conference* (TREC) launched their Medical Record Track which aimed to find cohorts for the research on certain topics. The document set used in the track was based on a set of de-identified clinical reports made available by the University of Pittsburgh's BLULab NLP Repository. From 2013, ShARe/CLEF eHealth Evaluation Lab organized the User-Centered Health Information Retrieval evaluation Lab. The collections come from several online well-known medical sites and databases (e.g. Genetics Home Reference, ClinicalTrial.gov, Diagnosia). In 2014, TREC held the Clinical Decision Support track, aiming to help physicians to find out information about how to care their patients. We would have thought that with the help of rich domain resources, the retrieval performance can be largely increased. Actually in many cases, the result didn't

```
Phrase:
  text:  treatment
Candidates:
  Candidate:
    Score:  -1000
    Concept Id:  C0039798
    Preferred Name:  therapeutic aspects
  Candidate:
    Score:  -1000
    Concept Id:  C0087111
    Preferred Name:  Therapeutic procedure
  Candidate:
    Score:  -1000
    Concept Id:  C1522326
    Preferred Name:  Treating
  Candidate:
    Score:  -1000
    Concept Id:  C1533734
    Preferred Name:  Administration procedure
  Candidate:
    Score:  -1000
    Concept Id:  C1705169
    Preferred Name:  Biomaterial Treatment
Mappings:
  Map Score:  -1000
    Score:  -1000
    Concept Id:  C1705169
    Preferred Name:  Biomaterial Treatment
```

Figure 2.20: MetaMap's human-readable output for the input text "*treatment*". It corresponds to 5 possible candidates in UMLS Metathesaurus. MetaMap will return the top ranked concept.

show significant improvement compared to a traditional IR method. On the contrary, in some experiments, large degradation has been observed. When knowledge is incorporated into retrieval model in different ways, the result widely varied. In this section, we will give an overview of the previous works of concept-based medical information retrieval.

2.4.1 Indexing by controlled vocabulary

A straightforward method is to identify all concepts in queries and documents. The medical terms in both queries and documents are normalized and represented by unique concept IDs. As shown in Figure 2.21, different from the traditional Bag-of-words (BOW) approach, when "*diabetic gastroparesis*" is mapped to concept *C0267176 [Diabetic gastroparesis]*, it is represented by its ID, which can be treated as a single term and the retrieval process can be based on any of the ranking model, such as the traditional OKAPI BM25 term-weighting scheme [84] or Language Model [81] with Dirichlet or Two-stage smoothing [105].

The central question people try to address is how concept IDs compare to original query terms in the task of medical IR. The results obtained using concept IDs are disappointing: they usually underperform traditional BOW approaches. In an early work of Hersh [46], UMLS concepts are extracted from documents and represented by their concept ID (CUI) (shown as Figure 2.21), and so are for the queries. The result suggested that information retrieval using controlled vocabulary provides no apparent advantages over word-based methods. However, in recent years, some similar experiences such as Qi et al. [83] and Wang et al. [17] observed different results, in which the performance of vector space model [85] using UMLS concepts was largely higher than that using BOW. The contradictory results do not allow us to draw a clear conclusion. It is worth noting that Hersh [46] and Qi et al. [83] used only the MetaMap top ranked UMLS concept, but Wang [17] didn't use MetaMap's disambiguity module. All candidates were used to represent the concepts.

Instead of using the whole UMLS Metathesaurus, Koopman et al.[56][57] used SNOMED ID as the representation of the concept. They found that retrieval results are heavily dependent on the quality of concept extraction provided by MetaMap. So they also included all the candidate concepts rather than the top-ranked ones. With all the candidate SNOMED concepts, the average document length was 6066 terms per document, much more than using the top ranked concept - 1391 terms per document. The candidate concepts were included by query expansion technique. Finally the retrieval performance was improved, but the improvement was not statistically significant over the method using the whole UMLS.

2.4.2 Query expansion with controlled vocabulary

Query expansion [101][100] is commonly used in biomedical information retrieval. Considering that retrieving using concept IDs doesn't show clear advantage and their performances largely rely on concept mapping accuracy, researchers prefer to keep the original query. Thus concepts are only used to do expansion. When concepts are assigned small weights, it means that concepts just play a limited role to adjust the final

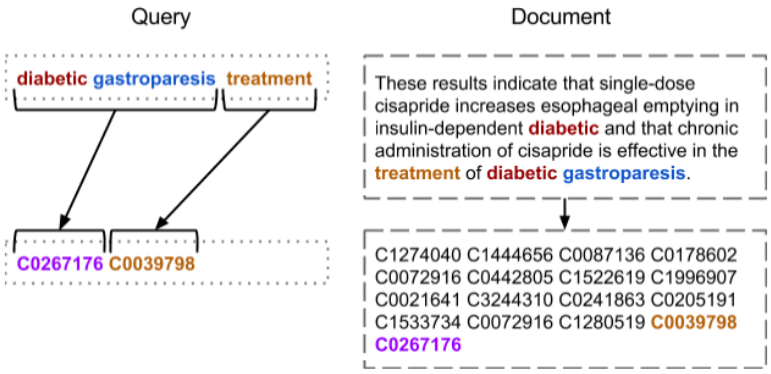


Figure 2.21: Concepts are identified in both query and document. Then document is indexed by their concept IDs (UMLS CUIs as example).

ranking result. For example, also in Qi et al.'s experiment [83], they have tried to mix the original query and their corresponding UMLS concept IDs, and so as for the documents. However, their experiment didn't gain large improvement. In some other experiments such as Zhu [110], concept IDs were not explicitly expanded into queries, but used to do result re-ranking. Their experiments were conducted in a two-step ranking strategy. In step 1, a normal retrieval was performed in raw text space. In step 2, they re-rank the output from step 1 in the concept space. However, this kind of attempt has shown to be ineffective.

Note that retrieving in concept IDs space could be a time consuming task. It requires terms to be normalized on both queries and documents. In some experiments, concepts are represented by their preferred names, thus documents can be kept as raw text. For example, Malagon [61] expanded query by recognized MeSH descriptors, which is similar to the matching strategy of the PubMed (the online open access to the MEDLINE biomedical literature). As shown in Figure 2.22, "*Scheuermann Disease*" and "*Therapeutics*" are two mapped standard MeSH heading. They are wrapped with the `#1()` operator (match the terms as an exact phrase) in Indri [93] and expanded with a small weight. The `#combine()` operator just considers the terms to be independent, i.e. as a bag of words. Similarly, Choi [19] expanded their query by UMLS preferred name. But expanded terms were not wrapped by `#1()` operator but retrieved as independent terms in language model. In addition, the terms not occurring in the discharge

summary (Each query in CLEF eHealth 2013 dataset comes with a discharge summary, which describes the history and context in which the patient has been diagnosed with a given disorder.) were filtered out. However, no improvement is observed in this kind of preferred name method. In another work, Aronson [6] also used the *preferred name* of concepts in UMLS Metathesaurus. Their original queries were expanded by two different parts:

- *phrases* determined by MetaMap processing
- concept *preferred name*

For the query *scheumann disease treatment* in Figure 2.24, their expanded query can be expressed in the query language of *Indri* Retrieval System as the Figure 2.23. In this example, *scheumann disease* and *treatment* are the identified phrases in the original query, and *vertebral epithysitis* and *therapy* are respectively their preferred name in Metathesaurus. In the same way, the phrases are wrapped as an exact phrase match with the `#1()` operator in Indri. In addition, the words, phrases and concepts are respectively assigned the weights of 2, 1 and 5 (this was the best weighting scheme obtained after a series of experiments). Finally, this method gained 4.4% improvement over the baseline (but not statistically significant). Compared with a similar previous work by Srinivasan [90], their conclusion is the same: the improvement brought by UMLS preferred name is not significant.

A large improvement in controlled vocabulary expansion approach is observed in King et al.'s work [52]. A big difference between their experiments and those mentioned above is that instead of mapping only the longest concepts, they used all "nested" concepts in a phrase. For example, for the query "*diabetic gastroparesis treatment*", MetaMap will find concept *C0267176 [Diabetic gastroparesis]*. But in their

#weight(λ 1 #combine(scheumann disease treatment)
λ 2 #combine (#1(Scheuermann Disease) #1(Therapeutics))

Figure 2.22: The concept-based preferred name expansion in Malagon's experiment. [61]

Prefered Name Expansion of Aronson [6] =
2 #combine(scheumann disease treatment)
1 #combine(#1(scheumann disease) #1(treatment))
5 #combine(vertebral epiphysitis therapy)

Figure 2.23: Aronson [6] also used the *preferred name* of concepts in UMLS Metathesaurus. Their original queries were expanded by two different parts: (1) *phrases* determined by MetaMap processing. (2) concept *preferred name*. The three conponents are respectively assigned the weights of 2, 1 and 5, which was the best weighting scheme obtained after a series of experiments.

experiment, *C0241863 [Diabetic]* and *C0152020 [Gastroparesis]* will also be expanded to the query. This method led to slight degradation (MAP decreased from 0.275 to 0.269) in pure concept IDs space. However, when concept IDs were mixed with original query, the results were largely boosted (increased from 0.275 to 0.325 in MAP).

2.4.3 Concept-based synonyms expansion

Instead of normalizing concepts by their standard form (IDs or preferred name), another alternative approach is to expand the synonyms of the concepts. As shown in Figure 2.24, when concept *C0036310 [Scheuermann's Disease]* and *C0039798 [therapeutic aspects]* are recognized in the original query, their synonyms enumerated in the thesauri can be used to enrich the query expression.

For example, in Claveau's work [20], the queries were expanded with synonyms

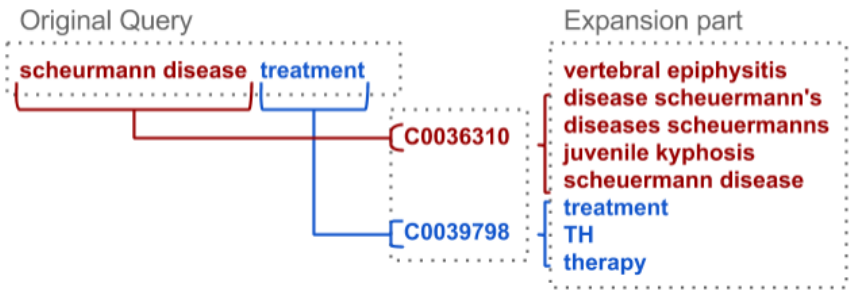


Figure 2.24: When concepts are recognized in query, their synonyms will be expanded into the original query.

found in the UMLS, and then retrieved by independent uni-gram language model [81]. As shown in Figure 2.25, original query terms were wrapped first by `#combine()` operator (combine the score of inner terms with the same weight), and were assigned a more important weight, while the weight of synonyms (wrapped by a second `#combine()` operator) did not exceed 0.1. Their experiments led to a slight improvement. Thakkar [95] expand MeSH entries in the same way, however the result was negative.

Similarly, Bedrick [10][9] also used MeSH entries. The difference is that the synonyms were further linked by `OR`, and different MeSH headings are connected with `AND` operator. In addition all synonyms were encapsulated as *exact phrases*, grouped by `#1()` operator. As an example, the query *scheumann disease treatment* will be transformed to the query in Figure 2.26. However this method is less effective than the traditional language model, because the Boolean operator `AND` is too strict.

In addition to the UMLS concepts or the MeSH headings, experiment based on Wikipedia also showed slight improvement. In Zhong's work [106], each title of Wikipedia article which describes a specific topic is denoted as a concept. As Wikipedia uses redirect hyperlink to redirect the user to different articles corresponding to the same title, and this kind of redirect pages are used as synonyms, which are used to be expanded into the query.

In some other experiences, the synonym expansion technique targeted some specific terms in query. For example, four groups [27][62][25][26] in TREC medical track filtered concepts by their Semantic Type (ST). Only concepts that belong to "*Pharmacologic Substance*", and "*Therapeutic or Preventive Procedure*" were added to the query. Three groups [10][27][62] implemented drug name or disease/symptom

```
#weight (0.9 #combine(scheumann disease treatment)
0.1 #combine(vertebral epiphysitis disease
scheuermann diseases scheuermanns juvenile
kyphosis scheuermann disease TH therapy) )
```

Figure 2.25: Synonyms expansion in Claveau's work [20].

```
("scheumann disease" OR "vertebral epiphysitis"
OR "disease scheuermann's" OR "diseases
scheuermanns" OR "juvenile kyphosis" OR
"scheuermann disease")
AND
("treatment" OR "TH" OR "therapy")
```

Figure 2.26: Synonyms expansion and boolean research in Bedrick's work [10][9].

expansion. In other cases, not all synonyms are used to expand the query. For example, Oh [80] and Zuccon [112] expanded only the full representation of acronyms or abbreviation into the queries. These conservative strategies always lead to quite slight improvements.

2.4.4 More than synonyms: Concept-based hybrid expansion

In some other works, not only synonyms are expanded in the query. For example, in Zhu's work [111][109][110], in addition to the detected MeSH concepts and their entry terms, all the descendant nodes down to *3* levels in the MeSH tree were also included. In addition, they model the term proximity by *Indri* operators. For example: for the MeSH terms such as "*Hearing Loss, High Frequency*", they can specify that the phrase "hearing loss" must occur as it is, while the latter phrase "high frequency" can occur within a text window of 16 words. This translates into the following Indri query: *#uw16(#1(hearing loss), high frequency)*, where `#uw16()` means that the three elements - #1(hearing loss), high and frequency, should appear within a window of 16 words. This method shows different performances on different datasets. For the official test collection for the TREC 2011 Medical Records Track, MAP (Mean Average Precision) was significantly boosted by up to 11%. However, the performance dropped on CLEF eHealth 2013 Evaluation Lab collection.

Drame [28] expanded their queries using concept synonymous terms, descendants in MeSH and UMLS thesaurus, and related MeSH terms (with the "*See Also*" relations in MeSH). This method can improve their baseline (*Precision at 10 (P10)* increased from

0.51 to 0.55). Unfortunately, since their baseline didn't perform well (much lower than the best run in their evaluation lab), their result was not competitive against other groups. In addition to the descendants (denoted by "*IS-A*" relationship), Koopman [56][57] extended the scope to all types of SNOMED CT relationships, but the improvement was also very limited.

2.4.5 Some conclusions

So far we can see that we cannot draw a clear conclusion on the performance of concept-based biomedical information retrieval approach. The results largely varied when different thesauri were used and when concept information was integrated into the query in different ways. In addition, the previous experiments have been carried out on different test collections, making it difficult to compare the results from different studies. This inconsistency can be partly explained by several factors:

- The experiments have been carried out on different test data. Although a large number of experiments used standard collections such as OHSUMED, others used their own test data, making the experimental results hardly comparable. In addition even on the same dataset,some experiments used only short queries and others used discharge summaries. These differences may have a significant impact on the results.
- The retrieval methods range from vector space model, language model, to other more heuristic methods, with or without pseudo relevance feedback.
- The resources used along with their mapping tools are often different. The resource can be: MeSH, SNOMED, ICD-9 or UMLS Metathesaurus. The mapping tools range from MetaMap, MedTex, SAPHIRE, etc.
- The resources are used in different ways: to normalize concept expressions or to expand queries.

So the question whether the rich resources in medical area can help to improve IR effectiveness is still open, and is what we will examine in this thesis.

The following Table 2.I, Table 2.III and Table 2.II provide a summary of the methods used in previous works. The platform and IR model column indicate the search

engine and the retrieve model used as the baseline. Knowledge resource and Concept mapping columns show the thesaurus used and the way how concepts are extracted from raw texts. The method section is a brief description of the method used, and the result section gives the performance obtained. Among them, Table 2.I lists some early works. Table 2.III illustrates the concept-based approaches implemented in TREC 2011/2012. Table 2.II summarizes the related works in CLEF eHealth 2013/2014 Ad hoc Biomedical Retrieval task. Medical Record track while

	Platform	IR Model	Knowledge Resource	Concept Mapping	Methods	Result
Hersh 92 [46]	SMART	tf.idf	UMLS	SAPHIRE	Bag of concept IDs	Concept space is less effective.
Srinivasan 96 [90]	SMART	tf.idf	Statistically produced	Statistical correlation [90]	Concept preferred name expansion	2.2% over baseline
Aronson 97 [6]	INQUERY [16]	Inference Network Model [97]	UMLS	MetaMap	Concept preferred name expansion	4.4% over baseline

Table 2.I: Early works of concept-based approach in biomedical information retrieval

	Platform	IR Model	Knowledge Resource	Concept Mapping	Methods	Result
snumed @CLEF13 @CLEF14 [19][18]	Indri	Language Model	UMLS	MetaMap	Concept preferred name expansion (filtered by discharge summary)	Only boosted MAP score by 1.35%
THCIB @CLEF13 [106]	Lucene	Okapi BM25 + PageRank [14] + HITS [54]	UMLS	Their propre extractor	Abbreviation/Acronym expansion	Outperforms the baseline slightly.
KISTI @CLEF14 [80]	Lucene	Language Model	UMLS	Rule-based abbreviation recognizing method [44]	Abbreviations/Acronyms expansion	Abbreviation expansion only got slight improvement.
DAI @CLEF14 [95]	Indri	Okapi BM25	MeSH	MetaMap	MeSH entries expansion	Performance dropped.
ERIAS @CLEF14 [28]	Lucene	Vectorial Space Model (VSM)	MeSH	MetaMap	MeSH entries, descendants and related terms (*see also* relation) expansion	Earned some improvement but the baseline didn't perform well.
RePALI @CLEF14 [20]	Indri	Language Model	MeSH	MetaMap	MeSH descriptors expansion	No improvement observed.
UHU @CLEF14 [61]	Lucene	Language Model	UMLS	MetaMap	UMLS asserted synonyms expansion	Only slight improvement.

Table 2.II: Concept-based approaches in CLEF eHealth 2013/2014 Evaluation Lab Task 3

	Platform	IR Model	Knowledge Resource	Concept Mapping	Methods	Result
AEHRC @TREC11 @TREC12 @CLEF13 [56][57][112]	Indri	unknown	SNOMED CT	MetaMap	SNOMED CT related concepts expansion in Concept IDs space	slight improvement
Cengage @TREC11 [52]	Lucene	Language Model	UMLS	Their own mapping tool	Concept IDs and raw text mixture space	Up to 18% significant improvement
NICTA @TREC12 [62]	Lucene	tf.idf	UMLS	MetaMap	Preferred name expansion	Dropped 2.68%
Delaware @TREC12 [17]	Indri	Language Model	UMLS	MetaMap	Bag of concept IDs (with candidate)	7% no significant improvement
OHSU @TREC12 @CLEF13 [10][9]	Lucene	Boolean retrieve	MeSH & ICD-9	MetaMap	MeSH entries expansion with *#1* exact match and Boolean retrieval.	Dramatically decrease
NEC @TREC12 [83]	Unknown	Language Model	UMLS	MetaMap	Concept IDs and raw text mixture space	Both performed better than raw text space.
ZHU @TREC12 @IEEE @CLEF13 [109][111][110]	Unknown	Language Model + MRF + MRM	MeSH	Unknown	Concept based synonyms expansion + hyponyms expansion	Got 11.9% improvement in [111] but caused slight degradation in [109] and [110].
LSIS @TREC12 [42]	Terrier	Divergence from Randomness (DFR)[1]	UMLS	MetaMap	Concept IDs and raw text mixture space	MAP decreased 11.6%.
York @TREC12 [48]	Lemur	Okapi BM25	MeSH	MetaMap	Bag of concept IDs and Boolean retrieval	Dropped 4.4% at P@10.
nlm @TREC11 @TREC12 [25][26]	Lucene [64]	Lucene "off-the-shelf"method	UMLS	MetaMap	Concept preferred name expansion (only for drugs and treatments)	Slight improvement.

Table 2.III: Concept-based approach in TREC 2011/2012 Medical Record Track Evaluation Lab

CHAPTER 3

OUR METHODS

As described in the last chapter, the previous research results of concept-based biomedical information leave the following questions wide open: can we really benefit from the rich knowledge resources in biomedical IR? Is it beneficial to perform concept-based retrieval? And little effort has been made to facilitate a comprehensive analysis and comparison of different methods under the same framework, which is the goal of our experiments. The methods we test are in line with those used in the previous studies, but some modifications are made.

In this chapter we will describe the methods implemented in our experiments. The experimental results will be described in Chapter 4.

3.1 Language Model - BOW baseline

Medical IR can be done using a traditional Bag-of-words (BOW) IR approach. In our experiments we use a **Language Model** [81][103][104] with Dirichlet smoothing [105] technique, which has been proven to be a strong baseline in biomedical IR [94]. A language model M for a document provides the probability of occurrence of terms in it. Documents are ranked according to the probability that its language model can generate the query. The score of a document D given a query $Q = q_1 q_2 ... q_n$ is determined as follows:

$$S(Q,D) = \frac{1}{n} \sum_{i=1}^{n} \ln P(q_i \mid D) \tag{3.1}$$

Where the sum of the generating log-probability will be normalized by the length of query n. The probability $P(q_i|D)$ can be estimated by Maximum Likelihood. For example (as shown in Figure 3.1), the probability that the model M generates the sentence "*patient with headache*" can be seen as a Markov chain, with the emission probabilities corresponding to those of the unigrams. Then the probability of the whole sentence T is shown as Equation 3.2.

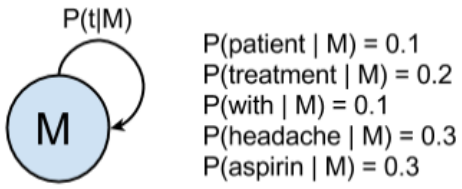


Figure 3.1: The probability distribution of a virtual *Language Model*.

$$Score(T \mid M) = P(patient \mid M) * P(with \mid M) * P(headache \mid M) = 0.1 * 0.1 * 0.3 = 0.003 \tag{3.2}$$

However, a core problem in language model estimation is the data sparseness. If one of the query terms doesn't occur in the document, the final score will be zero. This appears to be too rigid: many relevant documents do not contain all the query words. To solve this problem, we use smoothing. The goal of smoothing is to give a small probability to "unseen" terms. The Bayesian smoothing using Dirichlet priors is a commonly used method, which can be formulated as follows,

$$P(q_i \mid D) = \frac{tf_{q_i,D} + \mu \frac{tf_{q_i,C}}{|C|}}{|D| + \mu} \tag{3.3}$$

Here $tf_{q_i,D}$ means the frequency of query term q_i in document D. C represents the whole collection and $|C|$ is its size. Except solving data sparseness problem, smoothing was found to also add an IDF factor, a factor that assign more importance to less frequent terms, which is known very useful in IR. The dirichlet prior smoothing is one of the best smoothing methods for IR.

3.2 Bag-of-concepts (BOC) Approach

As we illustrated in Figure 2.21, a concept mapping tool recognizes the concepts in both queries and documents. The recognized concepts can then be represented by their concept IDs (e.g. CUI in Metathesaurus). The traditional retrieval model can then be used to match documents and queries in the concept IDs space. This approach can

be implemented easily: one can index both documents and queries by the concept IDs, and a traditional retrieval model can be conducted in concept IDs space. The possible benefits of such an approach are as follows:

- The concept normalization effect created with the unique concept ID allows us to match a document containing a concept (eg. "*hypertension*") with a query containing its synonym (eg. "high blood pressure").
- Retrieving using IDs forces terms to appear in the right way. A traditional BOW approach could find some document about *low blood pressure* for the query *high blood pressure* because of the high frequency of independent term *blood* and *pressure* in the document, or the word "high" appears at some places in it. This will not happen in the BOC approach because the concept expressions should be recognized in the document.

However, the previous result often show that the BOC approach underperforms a traditional BOW approach. One possible reason could be that the result is affected by the large number of concept mapping errors. As we mentioned earlier, MetaMap achieved 84% in precision and 70% in recall [82]. These numbers may not be high enough for a CUI-based approach. In our experiment, we reproduce a BOC run. Our goal is not to get the best overall result with this approach, but to confirm the performance of BOC approach and to observe how largely the retrieval result can be affected by these mapping errors. We will provide more details in Chapter 4.

3.3 Concept Hyponyms expansion

A problem often observed is that concepts in the query and the relevant documents can be at different hierarchical/granularity levels. As shown in Figure 3.2, a query can request for "*C0016658 [fracture]*", "*C0016662 [open fracture]*", "*C0149531 [fracture of pelvic]*", "*C0272577 [open fracture of pelvic]*" while a relevant document describes a more specific concept "*C0435785 [Open fracture pelvic, multiple public rami - unstable]*", which cannot be matched by CUI. However in BOW approach, they can be

partially matched using independent terms. This has been reported by Zuccon [113], and was called **Granularity Mismatch** problem. A straightforward solution is to find all the *nested concepts* of a long phrase. For example, from "open fracture of pelvis", we can extract four different concepts: *C0272577 [open fracture of pelvis]*, *C0016662 [open fracture]*, *C0149531 [fracture of pelvis]*, and *C0016658 [fracture]*. Finkel and Manning [31] provide a recognition method. However this task has to be conducted on both queries and documents, which is time consuming. An alternative way is to expand the query in concept ID space by their hyponym concepts to form an expansion query Q_E.

In UMLS Metathesaurus, the hyponyms are connected by the "*CHD (Child)*" relation. We develop a small program to extract all related "*Child*" concepts from the relation file `MRREL.RRF` in Metathesaurus. Expansion is conducted at two different levels:

- *BOC_Exp1*: Query is expanded only with the one level hyponyms of each concept.
- *BOC_Exp2*: The two level hyponyms are included.

In practice, Figure 3.3 shows a "one level" hyponym expansion. The concept *C1963154 [Renal Failure Adverse Event]* has three hyponyms: *C1558058 [CTCAE Grade 3 Renal Failure]*, *C1558059 [CTCAE Grade 4 Renal Failure]* and *C1558060 [CTCAE Grade 5 Renal Failure]*. They are expanded as the synonyms by *#syn()* operator in Indri.

3.4 BOW and BOC mixture space

One may believe that the BOC approach is capable of finding highly relevant documents (high precision) while the BOW approach can ensure a wide coverage (recall). A simple method is to combine the two scores as follows (Equation 3.4):

$$Score_{BOW+BOC}(Q,D) = \beta Score_{BOW}(Q,D) + (1-\beta) Score_{BOC}(Q,D) \tag{3.4}$$

Where β is a parameter to be tuned. In practice, documents are indexed in two different fields: <*original*> and <*cui*>, respectively corresponding to `BOW` and `BOC` space. In

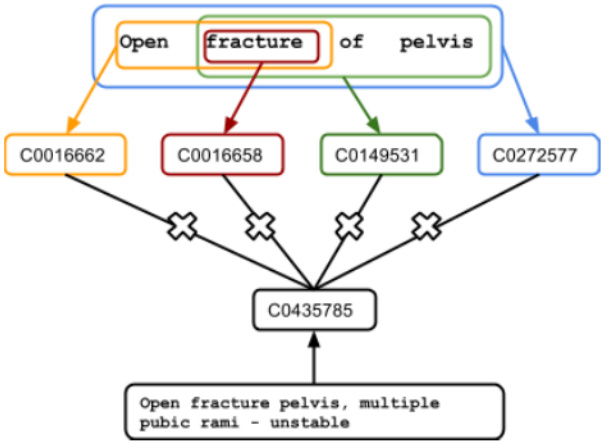


Figure 3.2: Granularity mismatch problem. The concept *C0435785* in document cannot be matched with any concepts in query "*open fracture of pelvis*".

BOC Query:

```
<query>
<number>001</number>
<text>
#combine(C1963154)
</text>
</query>
```

Query after Hyponyms Expansion:

```
<query>
<number>001</number>
<text>
#combine(#syn( C1963154 C1558058 C1558059 C1558060
))
</text>
</query>
```

Figure 3.3: The concept *C1963154 [Renal Failure Adverse Event]* was expanded by its three hyponyms: *C1558058 [CTCAE Grade 3 Renal Failure]*, *C1558059 [CTCAE Grade 4 Renal Failure]* and *C1558060 [CTCAE Grade 5 Renal Failure]* They are further wrapped by *#syn()* operator in Indri as synonyms.

Indri structured query language the "*term.fieldname*" operator specifies a term to be matched only with the content in a certain field. The example in Figure 3.4 shows how a BOW+BOC query looks like in Indri.

3.5 Concepts as phrases

As mentioned earlier, MetaMap does not recognize all the concepts correctly. The BOC method relying solely on the recognized concepts is thus prone to the recognition errors of MetaMap. In chapter 2, we have mentioned that a concept can be expressed in two different ways: concept IDs or a set of concept variants. By expanding the original queries with concept variants (shown in equation 3.5), queries and documents can be matched in word space.

$$\begin{aligned} &Score_{BOWPhrase}(Q,D) = \gamma Score_{BOW}(Q,D) + (1-\gamma)\, Score_{cptPhrase}(Q,D) \\ &where, Score_{cptPhrase}(Q,D) = \frac{1}{n} \sum_{cpt_i \in Q} \ln P(cpt_i \mid D) \end{aligned} \tag{3.5}$$

In equation 3.5, γ is a parameter to be tuned. The remaining question is: how should these concept variants be matched in documents? In other words, how to estimate the probability $P(cpt_i|D)$ of concept given a document. In spite of some previous works, we tested three different ways in our experiments.

Query after Hyponyms Expansion:

```
<query>
<number>001</number>
<text>
#weight( 0.9 #combine(Coronary.original artery.original
disease.original) 0.1 #combine(C0010054.cui) ) </text>
</query>
```

Figure 3.4: Documents are indexed in two fields: raw text and concept IDs. The original query "*Coronary artery disease*" is retrieved in original field while the concept ID "*C0010054*" focus in the CUI field. The original query is given a more important weight.

3.5.1 Exact Phrase Match (*Phrase_Exact*)

We consider that the occurrence of contiguous query terms shows strong evidence, that the document contains the right concept. By using Indri query language, the concept variants are encapsulated in the "`#1()`" operator, which matches the terms as an exact phrase. They are then grouped by "`#syn()`" operator which sums up the frequencies of different variants of a concept as shown in equation 3.6.

$$P_{\#1}(cpt_i, D) = \frac{\sum_{string_j \in cpt_i} tf_{\#1(string_j),D} + \mu \frac{\sum_{string_j \in cpt_i} tf_{\#1(string_j),C}}{|C|}}{|D| + \mu} \tag{3.6}$$

where, $string_{1,2,...,j}$ represent the different expressions of the concept cpt_i and $tf_{\#1(string_j),D}$ represents the frequency of the exact phrase of $string_j$ in document D. The final score is also smoothed by Dirichlet prior.

In Indri, a query will be translated to the following form:

Original Query :

```
<text>Anoxic brain injury</text>
```

Concept string expansion :

```
<text>#combine( #syn( #1(anoxic brain damage) #1(anoxic
brain injury) #1(anoxic disorder dup encephalopathy)
#1(anoxic disorder encephalopathy) #1(anoxic
encephalopathy) #1(anoxic enceph) #1(encephalopathy
hypoxic ischemic) #1(encephalopathy hypoxic) #1(hie) )
)</text>
```

In this example, `#combine` is used to combine different concepts (here only one is recognized in the query) and `#syn` is used to combine different concept expressions, treated as exact phrases (by #1).

3.5.2 Proximity full phrase match (*Phrase_Prox*)

Some previous works which match concept string as an exact phrase didn't perform well. It's probably because natural language, concept expressions could be much more varied than expected. In our experiment, we tried to give more flexibilities to "*phrase match*" method by modeling terms proximity.

For example, for concept "*C0746131 [lung lesion cavitary]*", the following expres-

sion "*pneumonia with cavitary lesions*" is not stored in Metathesaurus (Fig. 3.5) and cannot be matched. The phrase "*kidney of mouse*" is not stored as a string of the concept "*C1517673 [Mouse Kidney]*" in Metathesaurus. Using MetaMap, it will be divided into two parts: "*C0022646 [Kidney]*" "*C0025929 [Laboratory mice]*"and "*C0032285 [Pneumonia]*" "*C0221198 [lesion]*".

In order to cover such missed variants, it would be better to implement more flexible phrase matching strategies such as `#uwN()` operator in Indri, which refers to an unordered window – all terms must appear within a window of length N in any order. Thus the formula 3.6 will become as follows (Equation 3.7):

$$P_{\#uwN}(cpt_i, D) = \frac{\sum_{string_j \in cpt_i} tf_{\#uwN(string_j),D} + \mu \frac{\sum_{string_j \in cpt_i} tf_{\#uwN(string_j),C}}{|C|}}{|D| + \mu} \tag{3.7}$$

where $tf_{\#uwN(string_j),D}$ represents the frequency of unordered terms in $string_j$ within N-word windows in D (N is the length of each concept string). The following is an expanded query wrapped by #uwN operator in Indri. To give more flexibility, we can choose a larger size of window N. If we don't limit the window size, operator `#uw()` will allow all terms to appear anywhere in any order. In that way, for the concept "*Type 1 diabetes mellitus*", phrase "*Type 1*" and "*diabetes mellitus*" can appear in two different paragraphs. In the same way, two separate terms "*brain injuries*" and "*hypoxic*" can be matched to the query "*Anoxic brain injury*".

Query : `Type 1 diabetes mellitus`
Document : **`Diabetes mellitus`** `is a group of metabolic...There are three main types: - ` **`Type1`** `-Type2 -Gestational diabetes`

Query : `lung lesion cavitary`
Document : `a 44-year-old man ..... such as pneumonia with cavitary lesions on chest.`
Query : `mouse kidney`
Document : `.....proteins in the kidney of mouse treated with ASB identified by TOF/TOF MS`

Figure 3.5: Concepts can be represented by the expressions not included in the Metathesaurus.

Original Query :

```
<text>Anoxic brain injury</text>
```

Concept string expansion :

```
<text>#combine( #syn( #uw3(anoxic brain damage)
#uw3(anoxic brain injury) #uw4(anoxic disorder dup
encephalopathy) #uw3(anoxic disorder encephalopathy)
#uw2(anoxic encephalopathy) #uw2(anoxic enceph)
#uw3(encephalopathy hypoxic ischemic) #uw2(encephalopathy
hypoxic) #uw1(hie) ) )</text>
```

Figure 3.6: For the concept "*Type 1 diabetes mellitus*", phrase "*Type 1*" and "*diabetes mellitus*" can appear in two different paragraphs. In the same way, two separate terms "*brain injuries*" and "*hypoxic*" can be matched to the query "*Anoxic brain injury*".

Query : `Anoxic brain injury`

Document : `.....you remember that` **`brain injuries`** `in adults both` **`hypoxic`** `and traumatic are increasingly treated with therapeutic hypothermia......`

3.5.3 phrase partial match (*Phrase_Bow*)

In the above two methods, all terms in a concept string are constrained to appear in the document, ordered or unordered. In some cases, that is still too strict. Assume a query "*occult blood screening*", corresponding to the concept "*C0028792 [Occult blood screen]*" or "*C0201811 [Fecal occult blood test]*". In many documents, they are expressed as "*occult blood test*" which is not stored as a string of these two concepts in UMLS Metathesaurus. In order to allow a partial match between them, we break all concept string into words, and use a BOW query to match them. This corresponds to the following Indri query:

Original Query :

```
<text>Anoxic brain injury</text>
```

Concept string in Bag-of-words :

```
<text>#combine( anoxic brain damage anoxic brain injury
anoxic disorder dup encephalopathy anoxic disorder
encephalopathy anoxic encephalopathy anoxic enceph
encephalopathy hypoxic ischemic encephalopathy hypoxic
hie )</text>
```

3.5.4 Collect the Synonyms of concept

So far we have described how a recognized concept expression from a query can be processed. A remaining problem is how to extract the synonyms of a concept. In UMLS Metathesaurus, there are two different ways to extract the set of synonyms of a concept. Here we first introduce both of them, and explain our choice.

3.5.4.1 "RQ": Possible Synonyms and "SY": Asserted Synonyms in UMLS Metathesaurus

A straightforward way is to use the synonyms relationship explicitly defined in Metathesaurus. "**RQ**" (Related, possible synonym) and "**SY**" (Asserted Synonyms) are two different types of synonym, respectively referring to the possible synonyms and synonyms for sure. A big difference between them is that *RQ* connects different concepts, but *SY* only connects a concept with its different expressions (terms, strings and atoms). As shown in Figure 3.I, the majority of the possible synonyms of "*headache*" extracted by *RQ* relationship are narrower hyponyms. In contrast, the *SY* relation better meets our requirement. For our purpose of finding different expressions of a concept, the *SY* relation appears to be more appropriate. However none of these two relations were used in our experiment. What we choose is the third one, *Concept Strings*, which will be described in the next section.

3.5.4.2 UMLS concept Strings

As described in chapter 2, in Metathesaurus, the synonymous relationship is defined in "Concept/Term/String/Atom" framework. Another way to find all possible expressions of a concept is to extract related **Concept Strings**, which contain not only the synonyms but also the lexical variants as well as the acronyms/abbreviations. As shown in Figure 3.7, the asserted synonyms is just a subset of the group concept strings, which contains only the synonyms confirmed by physicians. In order to increase the coverage of concept expression, we use "Concept Strings" as the synonyms in our test.

Relation	Content
RQ	*Headache* || *Chronic Headache* || *Occipital headache* || *Head pressure* || *Headache dull* || *Frontal headache* || *Headache recurrent* || *incomplete anencephaly, hemicrania* || *Throbbing Headache* || *Temporal headache* || *Parietal headache* || *Primary Stabbing Headache* || *Headache fullness* || *Headache occurring* || *Nocturnal headache* || *Pounding in head* || *Headache (except migraine) aggravated* || *Cephalalgia or cephalgia* || *Head throbbing* || *Headache discomfort* || *Headache aggravated* || *Hemicrania* || *Frequent headaches* || *Drug-induced headache* || *Intermittent headache* || *Retroauricular pain* || *Hemicephalalgia* || *control of headache by self regulation* || *Facial Pain* || *Cluster Headache* || *Tension Headache* || *Vascular Headaches* || *Headache Disorders* || *Other specified headache syndromes* || *Headache; including migraine* || *Other headache* || *C/O - a headache*
SY	*Headache* || *pain in head* || *Pain in head* || *Cephalodynia* || *head pain* || *head pains* || *Head pain* || *HEAD PAIN CEPHALGIA* || *cephalgia* || *Cephalgia* || *Cephalalgia* || *cephalalgia* || *Cranial Pain* || *Head Pain* || *head ache* || *ache head* || *HEAD ACHE* || *HA-Headache* || *Cranial pain* || *cranial pain*

Table 3.I: The "Possible synonyms (RQ)"and "Asserted synonyms (SY)"of concept "headache".

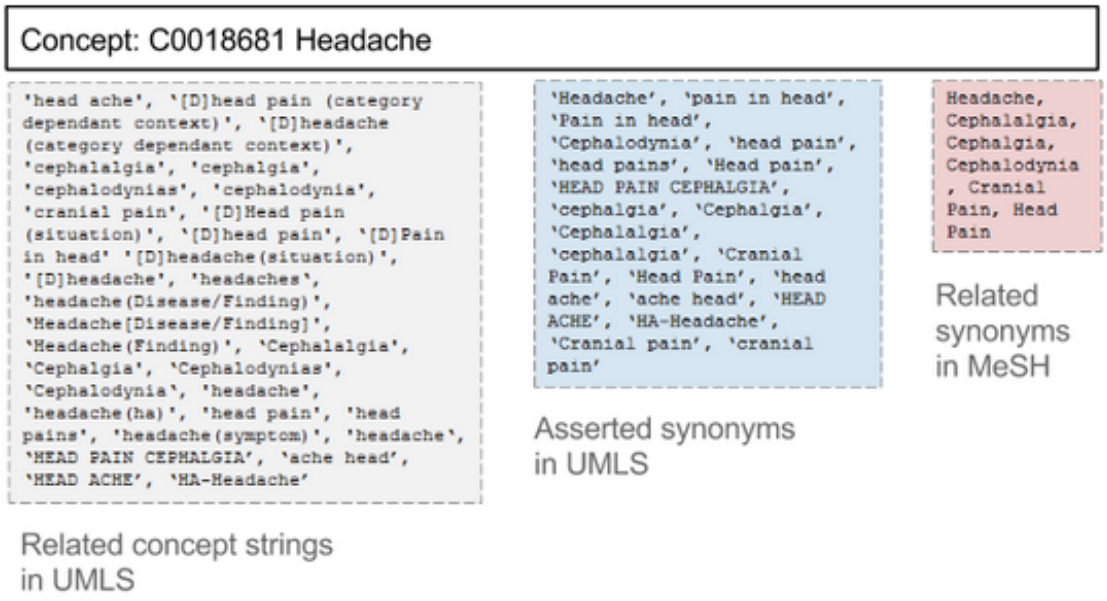


Figure 3.7: Asserted Synonyms with relation "SY", or MeSH Synonyms are only subsets of UMLS concept strings. In our experiments, we use the full Concept Strings to expand the original queries.

3.5.5 Concept Strings extraction

Technically, three methods can be used to extract concept strings from a given CUI. First, UMLS database has an official online access: UTS (UMLS Terminology Services) Java API 2.0 [76] provided by the web service. We can use the `getConcept` function to get the properties of concepts by their Concept IDs. But in terms of speed of execution, UTS API requires much processing time. The second way is to load Metathesaurus resource to local database such as MySQL, Oracle. NLM provide with the load scripts [72]. The third method is to extract them directly from the Metathesaurus data file: `MRCONSO.RRF`. There is exactly one row in this file for each atom in the Metathesaurus. The first column is the Concept ID, and the 15th one represents the String of that Atom. Figure 3.8 gives an example of one row in file `MRCONSO.RRF`. In our experiments we use the second method and loaded Metathesaurus to MySQL.

3.6 Markov Random Field and our hybrid approach (*Phrase_Comb*)

Some previous works modeling terms dependence, such as *Markov Random Field Model* [67], found that the combination of different dependence models could be more effective than each individual components. In [67], Metzler proposed three variants: *full independence* (FI), *sequential dependence* (SD), and *full dependence* (FD). Each of them are under different assumptions. The full independence model is a unigram bag-of-words language model. As the name suggests, the sequential dependence model considers dependence between adjacent query terms. These sets of terms are required to appear contiguously with the same order, which is similar to the *Phrase_Exact* matching. The third one, full dependence model, relies on the occurrence of non-contiguous terms within a text window. This strategy is similar to the *Phrase_Prox* method. There

C0001175|ENG|P|L0001175|VO|S0010340|Y|A0019182||M0000245|D000163|MSH|PM|D000163|**Acquired Immunodeficiency Syndromes**|0|N|1792|

Figure 3.8: One row in file MRCONSO.RRF in Metathesaurus. The first column is Concept ID and the 15th one refers to one of the concept string.

are however two differences between MRF model and our concept-based phrase matching: (1) The phrases we use correspond to concept expressions, while those in [67] are free sequences words in the query. For example, in query *bleeding after hip operation*, any two adjacent terms will be considered as a pseudo-phrase: *bleeding after*, *after hip*, and *hip operation*. We thus expect less noisy phrases in our case. (2) Our phrases also include the synonyms of the concepts; thus a query is naturally expanded by synonym concepts.

Metzler's experimental results [67] revealed that the sequential dependence model is more effective for some homogeneous collections (for example newswire collections) with verbose queries. In contrast, the full dependence model works better on less homogeneous collections (such as web pages) with shorter queries. The linear combination of three models (FI, SD, and FD) outperformed all three single models and achieved the best overall results. The combination is conducted by assigning a different weight for each dependence model, λ_{FI} for full independence model, λ_{SD} for sequential model, and λ_{FD} for full dependence model. Three weights are enforced to sum to one:

$$\begin{aligned} &S(Q,D) = \lambda_{FI} S_{FI}(Q,D) + \lambda_{SD} S_{SD}(Q,D) + \lambda_{FD} S_{FD}(Q,D) \\ &where, \quad \lambda_{FI} + \lambda_{SD} + \lambda_{FD} = 1 \end{aligned} \tag{3.8}$$

Inspired from the MRF model, in our experiments we combined the above three phrase matching methods, S_{Phrase_Exact}, S_{Phrase_Prox} and S_{Phrase_Bow}, with the traditional BOW scores using linear combination (shown in Equation 3.9),

$$\begin{aligned} S(Q,D) = & \\ & (1-\lambda_1-\lambda_2-\lambda_3) S_{BOW}(Q,D) \\ & + \lambda_1 S_{Phrase\_Exact}(Q,D) \\ & + \lambda_2 S_{Phrase\_Prox}(Q,D) + \lambda_3 S_{Phrase\_Bow}(Q,D) \end{aligned} \tag{3.9}$$

Notice that we will set the interpolation parameters manually. We test a range of values for the parameters. Each parameter is allowed to vary within {0, 0.025, 0.05, 0.075,..., 0.975, 1}. Our goal is not to see how we can automatically set the parameters to

their best, but to see how concepts can potentially help improving retrieval effectiveness when the parameters are set at some reasonable values.

CHAPTER 4

EXPERIMENT

In this chapter, we will first explain how our experiments are carried out. In the second section, we will describe the results. The section three will contain some analysis on our results obtained.

4.1 Experimental setup

Our experiments were performed on the following three standard data sets for MIR: OHSUMED collection and ShARe/CLEF eHealth 2013 and 2014 collections (task 3a). The OHSUMED collection contains a set of 1987-1991 MedLine Database Abstract. The document collections of ShARe/CLEF eHealth 2013 and 2014 are the same, collected from EU-FP7 Khresmoi project's 2012 medical documents web crawl. In 2014 collection, only a small number of documents from the 2013 dataset were removed. In our experiment, queries of CLEF 2013 and 2014 are all retrieved on 2014 documents collection. Table 4.I provides some statistics about the collections.

Before being used to build the index, the collections were cleaned up and were transformed to TREC Style file. Figure 4.1 shows a reference from MEDLINE (via OHSUMED dataset, 1988) in XML formatted text. Each citation is labeled by a set of meta data such as the `PMID` number, author, title, abstract, etc. CLEF collection is crawled from a large number of web site (e.g. Figure 4.2). All *HTML* tags, *CSS* and *JavaScript* or *JQuery* scripts have been removed in the clean-up. Figure 4.3 shows a

Collection	Num Doc	Query	Num qrel
OHSUMED	336133	1-106	2252
CLEF 2013	1101228	qtest1-qtest50	6218
CLEF 2014	1101228	qtest2014.1-qtest2014.50	6800

Table 4.I: Experiment data sets statistics

cleaned TREC style file. Only the highlighted sentences in Figure 4.2 are extracted to be the <*text*> field.

The queries used in three collections are developed by physicians. Figure 4.4 shows two original topics in OHSUMED and CLEF datasets and the corresponding TREC format queries converted from them. Queries in OHSUMED come with two parts: "*.W*" field is the request information while "*.B*" field describes the patient's information. In our experiments, we use only the "*.W*" field. Queries in CLEF dataset consist of a topic <*title*> field (text of the query), <*description*> field (longer description of what the query means), and a <*narrative*> field (expected content of the relevant documents). The query can be generated at different levels of detail according to the user's needs. In our experiments, we build two different type of queries. The <*title*> field is used to build a short query. And in long query <*desc*> field is added.

Notice that CLEF topics also include a discharge summary, which describes the patient health problems and history, and a patient profile. OHSUMED topics provide patient information as well. Although some previous experiments found this information useful [110], we will not use it in order to focus on the problem of MIR in its traditional setting.

We use Indri [93], a state-of-the-art language model based search engine as our retrieval platform. As shown in Figure 4.5, words are stemmed using Porter stemmer, and terms in PubMed stop words list are removed. Rather than the built-in Lemur stopwords, we choose PubMed stopwords, which include 133 terms selected by NCBI physicians for the purpose of medical information retrieval. Some units of measurement, such as *km, kg, mg* and some specific symbol like *PMID* are filtered.

The performance are evaluated by Mean average precision (MAP) at top 1000 documents retrieved and the precision at 10 top ranked documents (P@10). OHSUMED dataset provides two versions of relevance judgment file (Definitely relevant or probably relevant). We use the first one. For CLEF dataset, we have Graded and binary relevance judgment. We choose the graded version. The test of significance is carried out by one-tailed Student's t-test, with $p<0.05$ as statistically significant, and $p<0.01$ as extremely statistically significant.

```
<document>
<PMID>88000001</PMID>
<resource>Alcohol Alcohol 8801; 22(2):103-12</resource>
<mesh>
<meshTerm>Acetaldehyde</meshTerm>
<meshTerm>Buffers</meshTerm>
<meshTerm>Catalysis</meshTerm>
<meshTerm>HEPES</meshTerm>
<meshTerm>Nuclear Magnetic Resonance</meshTerm>
<meshTerm>Phosphates</meshTerm>
... ...
</mesh>
<title>
The binding of acetaldehyde to the active site of
ribonuclease:  alterations in catalytic activity and effects
of phosphate.
</title>
<publicationType>JOURNAL ARTICLE</publicationType>
<abstract>
Ribonuclease A was reacted with acetaldehyde and sodium
cyanoborohydride in the presence or absence of 0.2 M
phosphate.  ...  ...  and that modification of this lysine
by acetaldehyde adduct formation resulted in inhibition of
catalytic activity.
</abstract>
<authors>
<author>Mauch TJ</author><author>Tuma
DJ</author><author>Sorrell MF</author>
</authors>
</document>
```

Figure 4.1: A reference in MEDLINE.

Our retrieval experiment are set up as follows.

- **BOW:** A traditional language modeling approach with Dirichlet smoothing ($\mu =$ 3000) produces a baseline.
- **MRF:** A standard Markov Random Field combination, with the classic weighting schema: λ_{FI}=0.8, λ_{SD}=0.1, and λ_{FD}=0.1 [67].
- **BOC:** We use MetaMap to extract concept IDs (CUIs) from documents and queries.
- **BOC_Exp1 & BOC_Exp2:** Mapped concepts in queries are then further expanded by adding hyponym concepts, which corresponding to the use of one level and two levels of synonyms.
- **BOW+BOC:** It combines the original queries in BOW space and the mapped concept IDs with different weights. We test a range of values for the interpolation parameter β, varying within {0.7,0.8,0.9}.
- **Phrase_Exact (PE):** We further extract the UMLS "*concept strings*" of the iden-

```
#UID:river4274_12_000200
#DATE:201204-06
#URL:http://www.riverbendds.org/palatal.html
#CONTENT:
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 3.2 Final//EN">
<html>
<head>
<meta name=description content=Palatal plates and oral motor function-
children with Down syndrome/>
<meta name=keywords content="Palatal plates, oral motor, Down syndrome,
Down's syndrome, Trisomy 21 />
<meta name=author content=Irene Johansson />
<title> Palatal Plates and Oral Motor Function - Children with Down
Syndrome</title>
<link rel=StyleSheet href=global.css type=text/css media=screen/>
<link rel=stylesheet type=text/css media=handheld href=pocketpc.css />
<script language=JavaScript1.2 src=menusync.js
type=text/javascript></script>
</head>
<body bgcolor=#FFFFFF text=#000000 link=#0000FF vlink=#800080 alink=#EE0000
onload=self.focus();>
<h1> Palatal Plates and Oral Motor Function - Children with Down
Syndrome</h1>
<p></p>
<table width=100% border=0 cellspacing=0 cellpadding=0>
<tbody>
<tr valign=top>
<td width=50%><a href=mailto:ireneswipnet.se>
Irne Johansson</a><br/>
Department of Education, University of Karlstad<br/>
Birgitta Bckman<br/>
Department of Pedodontics, University of Ume;</td>
<td>Reprinted with permission of the publisher<br/>
Phonum 4, 9-12, Department of Phonetics, University of Ume, Sweden</td>
</tr>
</tbody>
...
...
#EOR
```

Figure 4.2: Web crawled document *river4274_12_000200* in CLEF eHealth collection. All *HTML*, *CSS* and *JavaScript* or *JQuery* scripts should be cleaned up from the document. The highlighted contents are those will be extracted as the *text* of the document.

```
<DOC>
<DOCNO>river4274_12_000200</DOCNO>
<TEXT>Palatal Plates and Oral Motor Function - Children with Down Syndrome
Palatal Plates and Oral Motor Function - Children with Down Syndrome Irne
Johansson Department of Education University of Karlstad Birgitta Bckman
Department of Pedodontics University of Ume Reprinted with permission
of the publisher Phonum 4 9-12 Department of Phonetics University of
Ume Sweden Abstract An evaluation of the effects of early oral motor and
sensory intervention including palatal plate therapy in children with
Down syndrome indicates positive effects at the age of 18 months The final
evaluation will be carried out when the children are 8 years old.
...
...</TEXT>
</DOC>
```

Figure 4.3: A clean TREC Style document extracted from OHSUMED web page *river4274_12_000200*.

Original OHSUMED topic file:

```
.I 1
.B
60 year old menopausal woman without hormone replacement
therapy.
.W
Are there adverse effects on lipids when progesterone is
given with estrogen replacement therapy.
```

OHSUMED input query for Indri search engine:

```
<query>
<number>1</number>
<text>Are there adverse effects on lipids when progesterone
is given with estrogen replacement therapy</text>
</query>
```

Original CLEF topic:

```
<topic>
<id>qtest2014.47</id>
<discharge summary>
22821âĹŠ026994âĹŠDISCHARGESUMMARY.txt
</discharge summary>
<title>
treatment for subarachnoid hemorrage
</title>
<desc>
What are the treatments for subarachnoid hemorrage?
</desc>
<narr>
Relevant documents should contain information on the
treatment for subarachnoid hemorrage.
</narr>
<profile>
This 36 year old male patient does not remember how he was
treated in the hospital.  Now he wants to know about the
care for subarachnoid hemorrage patients.
</profile>
</topic>
```

CLEEF short query for Indri search engine:

```
<query>
<number>short_1</number>
<text>treatment for subarachnoid hemorrage</text>
</query>
```

CLEEF long query for Indri search engine:

```
<query>
<number>long_1</number>
<text>treatment for subarachnoid hemorrage What are the
treatments for subarachnoid hemorrage</text>
</query>
```

Figure 4.4: How the original topics look like. And how they are transform to TREC Style queries. (Note: the spelling error in “hemorrage”is in the original topic.)

```
<DOC>
<DOCNO>001</DOCNO>
<TEXT>
Gabexate as a therapy for disseminated intravascular
coagulation.  [...]
</TEXT>
</DOC> <DOC>
<DOCNO>002</DOCNO>
<TEXT>
Activation and complexation of protein C and cleavage
and decrease of protein S in plasma of patients with
intravascular coagulation.  [...]
</TEXT>
</DOC>
```

Figure 4.5: TREC style document after collection pre-processing. Before being indexing, the meaningless stop words have been already removed, some words are reverted to their base form by removing the suffix (called *stemming*)

tified concepts in queries. The `Phrase_Exact` is an exact phrase matching with `#1()` operator.

- **BOW+Phrase_Exact (BOW+PE):** Original query is kept and is expanded by UMLS concept strings with `#1()` exact matching operator. We test three different weights $\gamma = 0.7, 0.8, 0.9$.
- **BOW+Phrase_Prox (BOW+PP):** It is the linear combination of original query and flexible concept strings with a series of different Indri Query Language operators (`#uwN()`, `#uwN+1()`, `#uwN+2()`, `#uw()`, `#uw5`, `#uw6()`...). We test three different weights $\gamma = 0.7, 0.8, 0.9$.
- **BOW+Phrase_Bow (BOW+PB):** The original query is expanded by independent concept strings in BOW space. We test three different weights $\gamma = 0.7, 0.8, 0.9$.
- **BOW+Phrase_Combine (BOW+PC):** Finally, Run `BOW+Phrase_Combine` is an interpolated combination of BOW, Phrase_Exact, Phrase_Prox and Phrase_Bow. We test a range of values for the interpolation parameter λ_1, λ_2, and λ_3, varying within {0, 0.025, 0.05, 0.075, ..., 0.975, 1}.

Table 4.II summarizes the methods tested in our experiments. (Note: We have participated in the 2014 CLEFeHealth Lab. Our `baseline` and `BOW+phrase_prox` were submitted as GRIUM_EN_Run1 and GRIUM_EN_Run5.)

Run name	Description
BOW (Baseline)	LM+dirichlet smoothing using bag-of-words space of original query.
MRF	Reproduce standard Markov Random Field model.
BOC	Using concept IDs (CUI,Concept Unified Identifier) as index in a language model.
BOC_Exp1	CUI Space + direct hyponym CUIs expansion.
BOC_Exp2	CUI Space + two level hyponym CUIs expansion.
BOW+BOC	Linear combination of BOW and BOC.
Phrase_Exact (PE)	Retrieved by exact ordered matching operator `#1()`.
BOW+Phrase_Exact (BOW+PE)	Linear combination of BOW and Phrase_Exact.
BOW+Phrase_Prox (BOW+PP)	Combination of original query and unordered concept strings matching operators.
BOW+Phrase_Bow (BOW+PB)	Combination of original query and independent words in concept string.
BOW+Phrase_Combine (BOW+PC)	Linear combination of BOW and Phrase_Exact and Phrase_Prox and phrase_Bow.

Table 4.II: Experiment methods

4.2 Result and Analysis

The overall best results in all five query sets achieved significant improvement against the language model baseline. Specific to each method, retrieval in concept IDs space is less effective than original BOW approach. And we could hardly benefit from the further concept hyponyms expansion. The results of BOW+BOC mixture space are inconsistent, which led to significant improvement over baseline in OHSUMED collection, however the score decreased in CLEF collections. On the other hand, concept phrases expansion approaches perform much better, which can significantly boost the baseline performance for most query sets. The overall best results in three datasets were achieved by our hybrid concept phrases combination approach. The improvements are statistically significant over both BOW baseline and MRF model. These results confirm the usefulness of concepts when used as phrases. The Table 4.III shows the results of our experiments [2]. In this section, we will give a brief analysis to reveal both their advantages and disadvantage.

4.2.1 BOC (Bag-of-concepts) space V.s. BOW (Bag-of-words) space

As shown in Table 4.III, the BOC approach is less effective than the traditional BOW approach. This confirms our earlier intuition that such an approach is too rigid. On CLEF dataset the BOC approach always underperforms the BOW approach by large margins. Figure 4.6 show the difference of BOW and BOC on the CLEF collection. There are only a small number of queries that can benefit from the concepts in Metathesaurus. However on OHSUMED collection, BOC run obtains 0.1474 at MAP, only 8.3% (relative change) less than BOW baseline. As shown in Figure 4.7, the results appears to be more balanced: 42 queries got better MAP than baseline while 56 decreased.

However we cannot say that retrieving in concept IDs space is completely useless. In chapter 3 we have described the two main advantages of BOC space. First, retrieving compound words and phrases as a lexical unit can eliminate the ambiguous words in

2. In OHSUMED collection, five queries (8, 28, 49, 86, and 93) have no definitely relevant documents. Indri system automatically drop them in our experiment.

Method	OHSUMED		CLEF13 short		CLEF13 long		CLEF14 short		CLEF14 long	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
BOW	0.1607	0,2099	0.2844	0.4940	0.2709	0.4680	0.3945	0.7180	0.4026	0.6680
BOC	0.1474	0.1823	0.1677	0.3220	0.1745	0.3160	0.2276	0.4920	0.2494	0.5260
BOC_Exp1	0.1595	0.1960	0.1700	0.3120	0.1711	0.3060	0.2377	0.5040	0.2590	0.5360
BOC_Exp2	0.1596	0.1931	0.1749	0.3447	0.1618	0.2940	0.2303	0.4760	0.2542	0.5320
Method	OHSUMED (7:3)		CLEF13 short (9:1)		CLEF13 long (9:1)		CLEF14 short (9:1)		CLEF14 long (9:1)	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
BOW+BOC	0.1838 **	**0.2495 ** **	0.2512	0.4360	0.2205	0.3960	0.3325	0.6280	0.2913	0.5320
PE	0.1787	0.2228	0.2044	0.3600	0.2089	0.3400	0.2580	0.5400	0.2897	0.5620
BOW+PE	0.1858 **	0.2337 **	0.2895 M	0.4880 M	0.2812 * M	0.4740	0.4060 * M	0.7408	0.4202 * M	0.6940 * M
BOW+PP (#uwN+1)	**0.1888 ** M**	0.2277 **	**0.2908 * M**	0.4960 M	0.2817 * M	0.4680	0.4075 * M	0.7420	0.4221 ** MM	0.6940 * M
BOW+PB	0.1707 **	0.2267 **	0.2835	**0.5020 M**	0.2718	**0.4780**	0.3990	0.7260	0.4118 *	0.6820 * M
Method	OHSUMED		CLEF13 short		CLEF13 long		CLEF14 short		CLEF14 long	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
MRF	$\lambda_{FI} = 0.8$, $\lambda_{SD} = 0.1$, $\lambda_{FD} = 0.1$									
	0.1729 *	0,2297 **	0.2750	0.4680	0.2735	0.4560	0.3904	**0.7620**	0.4099	0.6760
BOW+PC	$weight_{original} = 0.8$, $\lambda_1 = 0$, $\lambda_2 = 0.1$, $\lambda_3 = 0.1$									
	0.1815 **	0.2267 **	0.2892 M	0.4940 M	**0.2838 ***	0.4700	**0.4137 * MM**	0.7580 *	**0.4316 ** MM**	**0.7180 * M**

Table 4.III: Results of our experiments. The *(7:3)* and *(9:1)* indicate the weight of the original query and of the expanded query. The overall best result for each dataset are highlighted. ***** **(**)** and M (MM) respectively indicate the statistically significant improvement over BOW baseline and over MRF in student one-tailed test with p<0.05 (p<0.01).

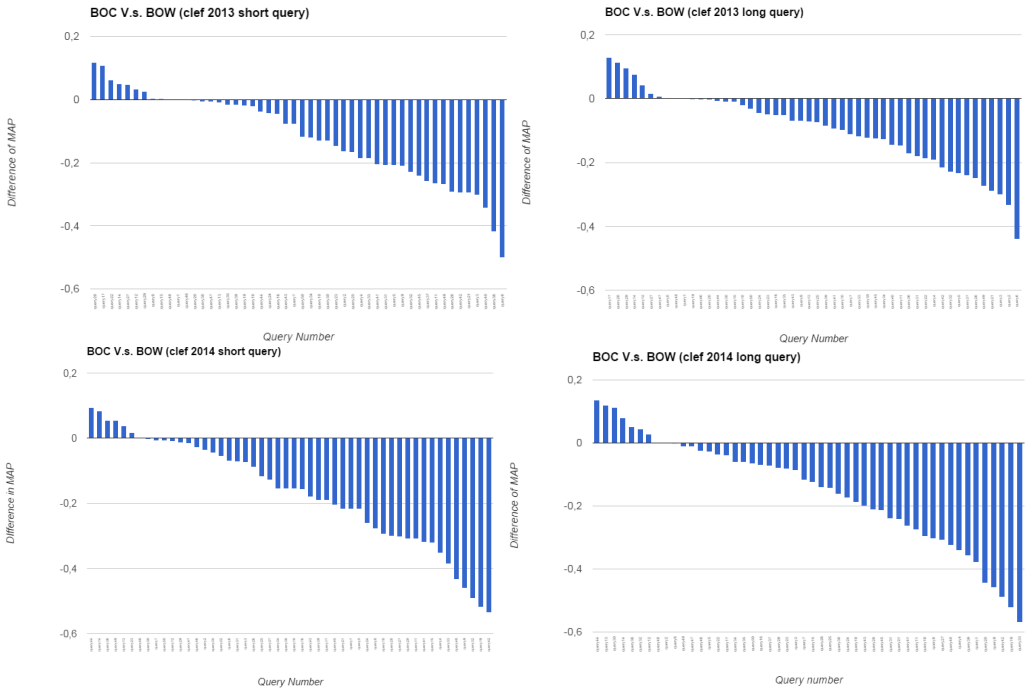


Figure 4.6: BOC V.s. BOW on CLEF 2013 and 2014 datasets.

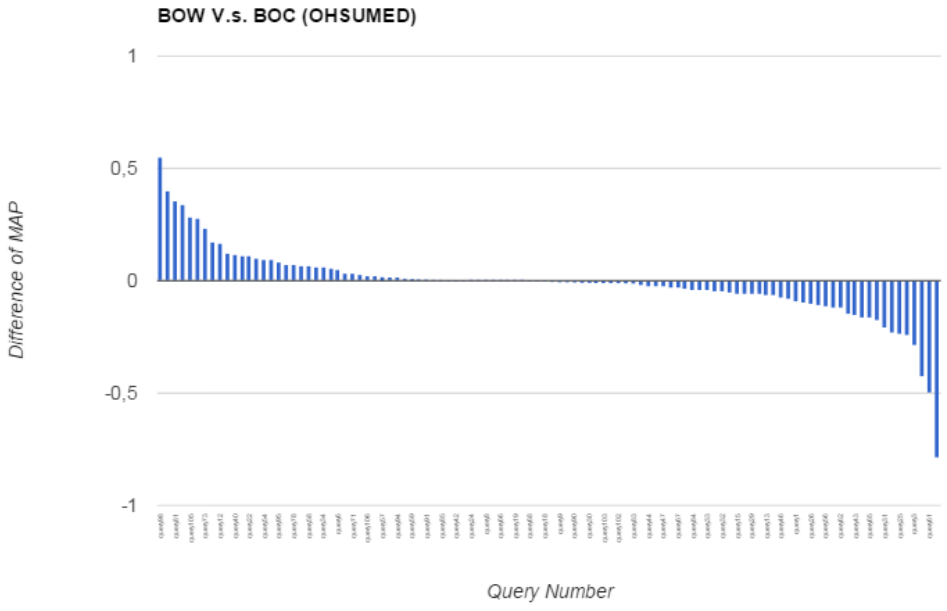


Figure 4.7: BOC V.s. BOW OHSUMED

BOW space and can improve the retrieval accuracy. Second, the normalization effect of synonyms enriches the terminology expression. We do observe both of them in our experiment. They can explain why 42 OHSUMED queries got better MAP in BOC approach, as well as for those queries that obtained lower MAP but higher precision. The following are some example of them.

- For query 21 of OHSUMED dataset "*secondary hypertension recent strategy workup*", the word "*secondary*" appears 7 times in the 10 top ranked documents of BOW approach, for example, "*secondary unexplained infertility*", "*secondary to type I Chiari malformation*", "*secondary tenth cranial nerve deficits*", and so on. None of them occur together with the word "*hypertension*". In BOC approach, these kind of noises will be eliminated. Finally in BOC approach, the MAP increases to `0.3447` from `0.0034`.

- For query 52 "*indications success pericardial windows pericardectomies*", word "*windows*" appears 37 times in top 10 documents of BOW approach, but only 14 of them refer to "*pericardial windows*". The others are used in irrelevant documents such as "*Nasal antral windows in children: a retrospective study.*". In BOC approach, when concept "*C0031041 [Pericardiostomy]*" was identified, these irrelevant documents are eliminated, and the MAP increase to `0.3215` from `0.0458`.

- The word "*scheurmann*" in query 98 has never appeared in relevant document. Instead, one of its aliases "*scheuermann*"is the most commonly used expression in OHSUMED collection. In BOC approach these two different spellings are both mapped to concept "*C0036310 [Scheuermann's Disease]*". In BOC approach, the MAP increased to `0.5515` from `0`.

- MetaMap mapped the word "*tips*" to concept "*C0339897 [Transjugular intrahepatic portosystemic shunt procedure]*". But in BOW, it is always matched to the ambiguous word "*tips*" (a useful hint or idea), which largely affected the retrieval performance.

All the above benefits are based on the basic premise that the concepts in both queries and documents are correctly recognized. The low effectiveness of BOC obtained may be explained by the relatively low concept mapping accuracy of MetaMap. When important concepts are not identified or when they are mapped to inappropriate concepts, the retrieval effectiveness will be greatly affected. It would be interesting to analyze how the mapping errors affect retrieval effectiveness.

We analyze the MetaMap mapping result of 50 CLEF 2014 short queries. In total 182 non-stopwords, MetaMap identified 85 concepts which cover 161 words. The raw coverage is 88.5%. Among them, only 61 concepts are correct (we examined them manually), which correspond to 125 words. Therefore the coverage of correct concepts is only 68.7%. This ratio is not high enough to support the CUI-based approach. A large number of concepts (14) have been mapped to an incorrect CUI or to an inconsistent CUI with respect to that in the relevant documents. 10 concept expressions are not recognized, or only partially recognized. There are four general types of error observed. Table 4.IV shows some typical examples observed in CLEF 2014 short queries.

Type 1. First, if a meaning has no corresponding concept in Metathesaurus, MetaMap will fail to identify it. MetaMap will rather choose one that looks similar. For the query "*foramen ovale*", the two most similar concepts are "*C1110599 [cranial foramen ovale]*" and "*C0016521 [diac foramen ovale]*". On the other hand, for the concept "*gallstone*", MetaMap has to choose one from the two candidates "*C0242216 [Biliary calculi]*" and "*C0947622 [Cholecystolithiasis]*" according to their context. None of them perfectly matches the query expression.

Type 2. An expression stored in Metathesaurus may often correspond to more than one CUI (e.g. "*Myocardial infarction*" may correspond to "*C2926063*" or "*C0027051*" depending on context). The large number of ambiguities in Metathesaurus is well known. Bodenreider [11] reported that 22% of the concepts related to "*heart*" in Metathesaurus are ambiguous. MetaMap uses heuristic rules, taking into account the context, to make the selection. As the context in a document may be different from that in a query, the selections in them can be inconsistent.

Type 3. The corresponding concepts exist in Metathesaurus, but the variant expressions encountered are not stored in Metathesaurus and not recognized by MetaMap. For example the phrase "*pneumonia with cavitary lesion*" is not stored as one of the concept names of the concept "*C0746131 [lung lesion cavitary]*" in Metathesaurus. Thus it is segmented into two

short concepts "*C0032285 [Pneumonia]*" and "*C0221198 [Lesion]*". In practice, especially in web search, such cases can be frequent. More powerful concept identification tools are required.

Type 4. Some errors are due to misspelling or unrecognized variations of words (We observed 5 misspelling in CLEF 2014 queries: "*repiratory (respiratory)*", "*gynecolocigal (gynecology)*", "*hemorrage (hemorrhage)*", "*urinanalysis (urinalysis)*", "*hereditarity (hereditary)*). Spelling error can frequently occur in practical uses, especially in web search. To cope with this problem, previous studies (e.g. Zuccon [112]) used spelling correction and query suggestion of search engines, which have been found useful.

In addition to concept mapping errors, another big problem is that using concept as the lexical unit in retrieve model is too rigid. Before conducting our experiments, we estimated (in chapter 3) that concepts in the query and the relevant documents can be at different hierarchical/granularity levels (Figure 3.2). Although the concepts are correctly recognized, the concepts used in a query and in the corresponding relevant documents are not the same. According to our observation, this problem is more complicated than expected. Three different cases have been observed.

Case 1. In the following Table 4.V, the first column gives the original query. The second and third column show respectively the concepts identified in query and relevant documents. We can see that users are more likely to use some general terms to describe the information they need, but medical literatures usually address a more specific issue. In this case, BOW approach can still partially match with a part of terms in documents. In BOC space however, they will be represented by two totally different concept IDs. For example, a query in OHSUMED asked for the effectiveness of the chemical elements "*C0016980 [Gallium]*" in treatment of disorder "*C0020437 [Hypercalcemia]*". However the chemical elements usually appears in the form of its chemical compound such as "*C0061008 [gal-*

Query	***CUI identified in query***	***CUI identified in relevant documents***
Coronary artery disease	*C0010054 [Coronary arteriosclerosis]*	*C0010068 [Coronary heart disease]*
Dropped **gallstone** *abscess right flank*	*C0242216 [Biliary calculi]*	*C0242216 [Biliary calculi] and C0947622 [Cholecystolithiasis]*
foramen ovale	*C1110599 [cranial foramen ovale]*	*C1110599 [cranial foramen ovale] and C0016521 [diac foramen ovale]*
Myocardial infarction	*C2926063 [Myocardial infarction:Finding:Point in time:Patient:Ordinal]*	*C0027051 [Myocardial Infarction]*
renal failure	*C1963154 [Renal Failure Adverse Event]*	*C0035078 [Kidney Failure]*
Right upper lobe **pneumonia with cavitary lesion**	*C0032285 [Pneumonia] and C0221198 [Lesion]*	*C0746131 [lung lesion cavitary]*
White blood cells *with moderate bacteria in urinanalysis*	*C0023508 [white blood cell count procedure]*	*C0023516 [Leukocytes]*
aspiration pneumonia due to misplacement of **dobhoff tube**	*C0175730 [biomedical tube device]*	*C3204189 [Dobhoff Tube]*
Right upper lobe **pneumonia with cavitary lesion**	*C0032285 [Pneumonia] and C0221198 [Lesion]*	*C0746131 [lung lesion cavitary]*
advices for patient with **acute infarctus myocardi**	*C0205178 [Acute]*	*C0155626 [Acute myocardial infarction]*
Bilateral pulmonary contusions *and safety belt*	*C0238767 [Bilateral] and C0347625 [Contusion of lung]*	*C2836276 [Contusion of lung, bilateral]*
Repiratory failure *and CHF*	*C0231174 [failure, biologic function]*	*C1145670 [respiratory failure]*
causes for **gynecolocigal bleeding**	*None*	*C0018417 [Gynecology]*
treatment for **subarachnoid hemorrage**	*C1515008 [subarachnoid route of drug administration]*	*C0038525 [subarachnoid hemorrhage]*
White blood cells with moderate bacteria in **urinanalysis**	*None*	*C0042014 [urinalysis]*
Chronic lymphocytic leukemia and **hereditarity**	*None*	*C0439660 [hereditary]*

Table 4.IV: The mapping error in CLEF 2014 queries.

lium nitrate]", which fails to match the concept ID in the query.

Case 2. MetaMap performs a lexical lookup within the sentence to find the longest spanning terms from the SPECIALIST Lexicon. This is correct in terms of concept mapping but not always reasonable for the purpose of information retrieval. Some long specific concepts in query may be too rigid and can be expressed by some short concepts in relevant documents. For example, given a query "*carcinoid tumors liver pancreas research treatments*", MetaMap recognized the concept "*C0345933 [Carcinoid tumor of pancreas]*". However this concept is usually expressed by two nested short concepts "*C0007095 [Carcinoid Tumor]*" and "*C0030274 [Pancreas]*". Some similar examples are shown in the following Table 4.VI.

Case 3. More often, the concept is neither too broader, nor too narrow. But language is so varied that we observed many concepts expressed in a totally different way in some relevant documents. For example, in a document talking about the disease "*Hypoaldosteronism*" the concept "*C0020595 [Hypoaldosteronism]*" is discribed as "*Such as urinary aldosterone levels and plasma renin activity, showed lower individual test performance characteristics*", and MetaMap is not able to extract the concept "*C0020595 [Hypoaldosteronism]*" from it. So this relevant document can not be matched in BOC approach when only "*Hypoaldosteronism*" is explicitly mentioned in the query. Table 4.VII shows some examples observed in OHSUMED dataset.

In addition, we observe that not all terms in the query can contribute to the retrieval performance. In the following examples, two queries come with their mapped concepts and their frequencies in judged definitely relevant documents. The occurrence of terms "*evaluation complications management*" don't provide strong evidence.

Query: *Evaluation complications management bulimia.*
- `C2945623 [evaluation and management]` [0 times]
- `C0009566 [Complication]` [6 times]
- `C0006370 [Bulimia]` [161 times]

Query: *Prevention risk factors pathophysiology hypothermia.*

Original Query	**Concepts in Query**	**Concepts commonly used in relevant documents**
effectiveness **gallium** *therapy hypercalcemia*	*C0016980 [Gallium]*	*C0061008 [gallium nitrate]*
back pain *information diagnosis treatment*	*C0004604 [Back Pain]*	*C0024031 [Low Back Pain]*
complications management **anorexia bulimia**	*C0003123 [Anorexia]* and *C0006370 [Bulimia]*	*C0003125 [Anorexia Nervosa]* and *C2267227 [Bulimia Nervosa]*
diverticulitis *differential diagnosis management*	*C0012813 [Diverticulitis]*	*C0518989 [Acute diverticulitis]* and *C0581275 [Colonic diverticular abscess]*
adverse effects lipids progesterone given **estrogen** *replacement therapy*	*C0014939 [Estrogens]*	*C0014938 [Estrogens, Conjugated (USP)]* and *C1136013 [Conjugated Equine Estrogens]*

Table 4.V: Example of topics represented by broad concepts in queries but described by more specific concepts in documents.

Original Query	Concepts in Query	Concepts commonly used in relevant documents
guillain barre syndrome sensitivity specificity **nerve conduction velocity tests**	*C0429381 [Nerve Conduction velocity]*	*C0027788 [Nerve conduction function]*, *C0501384 [Motor nerve]*, *C0501385 [Sensory nerve]*, *C0429379 [Motor nerve conduction block]*
chronic pain management review article tricyclic antidepressants	*C0747141 [chronic pain management]*	*C0150055 [Chronic pain]* and *C0030193 [Pain]*
carcinoid tumors liver pancreas *research treatments*	*C0345933 [Carcinoid tumor of pancreas]*	*C0007095 [Carcinoid Tumor]* and *C0030274 [Pancreas]*
adverse effects lipids progesterone given **estrogen replacement therapy**	*C0014935 [Estrogen Replacement Therapy]*	*C0014939 [Estrogens]*
rh isoimmunization *review topics*	*C0035404 [Rh Isoimmunization]*	*C2699077 [Rh Negative Blood Group]* and *C0302020 [Isoimmunization]*

Table 4.VI: Long and specific concepts are used in query, but in relevant documents the same topics are represented by a group of nested short concepts.

Concepts in Query	**Expression in Relevant Documents**
C0020595 [Hypoaldosteronism]	*Such as urinary aldosterone levels and plasma renin activity, showed lower individual test performance characteristics.*
C0020621 [Hypokalemia]	*low serum potassium levels*
C0030312 [Pancytopenia]	*leukopenia, anaemia and thrombocytopenia*
C0005586 [Bipolar Disorder]	*the occurrence of four or more mood episodes during the previous 12 months*
C0001849 [AIDS Dementia Complex]	*Twenty-nine patients at risk of developing acquired immunodeficiency syndrome (AIDS) presented with cognitive, motor, and behavioral dysfunctions characteristic.*
C3665358 [galactorrhea]	*Nipple discharge in women.*
C0220655 [Malignant pericardial effusion]	*The involvement of the pericardium by metastatic tumors is not uncommon, particularly in patients with lung cancer, breast cancer, lymphomas, leukemias, and melanomas.*

Table 4.VII: Some concepts can be described by the context of the documents, and MetaMap fails to identify the same concept as in the query.

- C1706420 [Prevention Study] [0 times]
- C0035648 [risk factors] [3 times]
- C0031847 [physiopathological] [3 times]
- C0020672 [Hypothermia, natural] [100 times]

Here we list more of this type of concepts. They should be treated in a different way, rather than to match them directly as for the other concepts. A way to address this issue is to use hyponym concepts. For example, if the concept "treatment" in a query is expanded to all the treatments, then the relevant documents containing any such treatment could be found.

- C1705169 [Biomaterial Treatment]
- C0521116 [Current (present time)]
- C0011906 [Differential Diagnosis]
- C0750430 [Work-up]
- C0029235 [Organism]
- C1704338 [diagnosis aspect]
- C0039798 [therapeutic aspects]
- C0332185 [Recent]
- C0679199 [Strategies]
- C0282443 [Review [Publication Type]]
- C0597535 [Success]
- C0376636 [Disease Management]
- C0039798 [therapeutic aspects]
- C0678257 [Description]
- C0332281 [Associated with]
- C0441655 [Activities]
- C0087111 [Therapeutic procedure]
- C0205179 [Advanced phase]
- C1706852 [Article]
- C1533716 [Information]
- C0011900 [Diagnosis]
- C0376636 [Disease Management]
- C0582205 [Utilities]
- C0015127 [Etiology aspects]
- C1280500 [Effect]
- C0332138 [Secondary diagnosis]
- C1518601 [Options]
- C1522427 [best (quality)]

4.2.2 Bag-of-concepts + Sub-concept Expansion

In the following table (Table 4.VIII), we report the results of the BOC approach and the hyponym concepts expansion approach. When we expand the BOC queries by sub-concepts, the overall performance of BOC approach is slightly improved (as shown in Figure 4.8). Such an expansion is really effective on several queries but leads to large degradations for some others. This method targets the queries that use broad concepts, such as those listed in Table 4.V. For these queries, their performance can be largely improved if some important sub-concepts used in relevant documents are expanded. For a query mentioned in the past section "*effectiveness galium therapy hypercalcemia.*", the expansion of the hyponym concept "*C0061008 [gallium nitrate]*" is particularly useful for this query.

However, on the other hand, we do observe large degradation on some queries. Some general concepts such as "*C0008679 [Chronic disease]*" is not adapted to the concept hyponyms expansion. They usually possess a large number of commonly used hyponym concepts, which can largely dilute the importance of the original concepts in query. The immediate hyponyms of concept "*C0008679 [Chronic disease]*" are listed in Figure 4.9.

Overall, concept expansion helped improving the results for 34 of the 101 queries only. Level 1 hyponyms expansion (BOC_Exp1) performs slightly better than a deeper one (BOC_Exp2). That's maybe due to Globally, even with hyponyms expansion, the BOC approach does not outperform the traditional BOW method.

Method	OHSUMED		CLEF13 short		CLEF13 long		CLEF14 short		CLEF14 long	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
BOC	0.1474	0.1823	0.1677	0.3220	0.1745	0.3160	0.2276	0.4920	0.2494	0.5260
BOC_Exp1	0.1595	0.1960	0.1700	0.3120	0.1711	0.3060	0.2377	0.5040	0.2590	0.5360
BOC_Exp2	0.1596	0.1931	0.1749	0.3447	0.1618	0.2940	0.2303	0.4760	0.2542	0.5320

Table 4.VIII: Result of BOC approach and hyponyms expansion approach. The overall best result for each dataset are highlighted. * and ** respectively indicate the statistically significant and extremely significant over BOW baseline in student one-tailed test with $p<0.05$ and $p<0.01$.

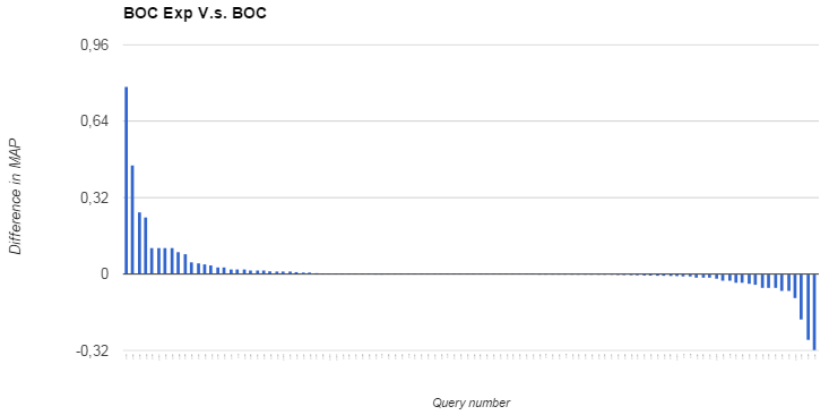


Figure 4.8: BOC_Exp1 V.s. BOC in OHSUMED dataset.

C0018621 [Hay fever], C0151317 [Chronic infectious disease] C0152210 [Intermittent tropia], C0232598 [Chronic vomiting] C0264220 [Chronic disease of respiratory system], C0264741 [Recurrent rheumatic fever], C0264742 [Chronic rheumatic fever], C0267763 [Chronic peritonitis] C0269041 [Chronic salpingitis], C0277556 [Recurrent disease] C0277583 [Chronic drug overdose], C0401066 [Unilateral recurrent inguinal hernia], C0423062 [Intermittent divergent squint], C0520801 [Chronic radiation sickness], C0558364 [Acute recurrent cystitis], C0581384 [Chronic anemia], C0585398 [Acute-on-chronic renal impairment], C0730314 [Chronic central serous chorioretinopathy], C0743244 [Chronic drug abuse], C0870281 [Chronic mental disorder], C1263722 [Chronic metabolic disorder], C1263743 [Chronic disease of genitourinary system], C1263765 [Chronic disease of breast], C1263865 [Chronic disease of ocular adnexa], C1263876 [Chronic disease of ear], C1264046 [Chronic disease of lymphatic vessels], C1264527 [Chronic poisoning], C1275398 [Chronic disease of hematopoietic system], C1290009 [Chronic disease of skin], C1290136 [Chronic disease of musculoskeletal system] C1290380 [Chronic disease of cardiovascular system], C1290611 [Chronic digestive system disorder], C1290882 [Chronic nervous system disorder], C1290886 [Chronic inflammatory disorder], C1290894 [Chronic disease of immune system], C1531663 [Chronic disease of immune function], C1531664 [Chronic disease of immune structure], C2316225 [Chronic headache disorder], C0001973 [Alcoholic Intoxication, Chronic], C0015674 [Chronic Fatigue Syndrome], C0150055 [Chronic pain] ...

Figure 4.9: The level one hyponyms of concept "*C0008679 [Chronic disease]*".

4.2.3 Combining BOW and BOC

The following table (Table 4.IX) reports the result of BOC approach and BOW+BOC approach.

This combination indeed produces significantly better results on OHSUMED than BOW. We observe that 73 queries (denoted in Figure 4.10) out of 101 benefit more or less from the combination while the rest get a lower score (shown in Figure 4.11). The bars in the figure denote the difference between BOW+BOC approach and BOW baseline in MAP, while the circles represent the score of BOW run and the crosses refer to the result of BOC approach. We have described in Section 4.2.1 that the BOC approach can give a higher precision, while the BOW approach result is good at recall. When the BOW part is given a more important weight, it can ensure a wide coverage of possible relevant documents. And when BOC approach is integrated, some highly relevant documents rank higher.

On the other hand, in Figure 4.11, we show the 28 queries that didn't perform well in BOW+BOC mixture space. We observed that the majority of them got a lower MAP in BOC approach than in BOW baseline because of the MetaMap mapping errors.

However, as the result of the poor performance of BOC approach, the BOW+BOC mixture space didn't show effectiveness on CLEF 2013 and 2014 collection. Different datasets require different interpolation weights β. As shown in Table 4.X, `7:3` is more suitable for OHSUMED collection, while the best weight for CLEF datasets is `9:1`.

Method	OHSUMED (7:3)		CLEF13 short (9:1)		CLEF13 long (9:1)		CLEF14 short (9:1)		CLEF14 long (9:1)	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
BOW	0.1607	0,2099	0.2844	0.4940	0.2709	0.4680	0.3945	0.7180	0.4026	0.6680
BOC	0.1474	0.1823	0.1677	0.3220	0.1745	0.3160	0.2276	0.4920	0.2494	0.5260
BOW+BOC	0.1838 **	**0.2495 **	0.2512	0.4360	0.2205	0.3960	0.3325	0.6280	0.2913	0.5320

Table 4.IX: Result of BOW, BOC and their combination: BOW+BOC approach. The *(7:3)* and *(9:1)* indicate the weight of the original query and of the expanded query. The overall best result for each dataset are highlighted. * and ** respectively indicate the statistically significant and extremely significant over BOW baseline in student one-tailed test with $p<0.05$ and $p<0.01$.

4.2.4 BOW+BOC V.s. BOW+Phrase_Exact

The Table 4.XI compares BOW+BOC and BOW+Phrase_Exact (PE) method. Both the BOC approach and the exact phrase matching approach rely strictly on the identified concepts. One may consider exact phrase matching as a naive concept identification process. However, this process simply compares the phrase in queries and in documents, but does not try to solve the possible ambiguities of an expression, which MetaMap tries to do, nor the possible variants. The results show that the exact phrase matching method works much better than BOC except for OHSUMED. For example, the phrase "*lupus anticoagulants*" appears 213 times in relevant documents. 140 of them are mapped to concept "*C0085240 [Lupus Coagulation Inhibitor]*" while the remaining 73 ones are identified as "*C0311370 [Lupus anticoagulant disorder]*" (these two concepts share the same string name). The inconsistent concept mapping largely affects the performance of BOC approach. But with exact phrase matching, however, we don't care which concept a phrase refers to. Once it occurs in a document, it will be matched. For this query, the 209 phrase "*lupus anticoagulants*" are all matched with the query using exact phrase match.

Alternatively, if an expression in document is allowed to map all the concept candidates, a higher recall can be expected. A suitable concept mapping process for MIR may be the one that identifies all concept candidates. Koopman [57] used all the SNOMED concept candidates instead of the top ranked concepts suggested by MetaMap. MetaMap can be configured to do this, but we did not test this option in our experiments.

Same as BOW+BOC approach, the results are largely varied with different interpola-

BOW+BOC	OHSUMED		CLEF13 short		CLEF13 long		CLEF14 short		CLEF14 long	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
7:3	**0.1838**	**0.2495**	**0.2522**	0.4320	**0.2305**	0.4060	0.3189	0.5900	0.2901	0.5460
8:2	0.1792	0.2406	0.2516	**0.4380**	0.2281	**0.4160**	0.3268	0.6220	0.2965	0.5500
9:1	0.1716	0.2347	0.2512	0.4360	0.2205	0.3960	**0.3325**	**0.6280**	**0.2913**	**0.5320**

Table 4.X: The results of BOW+BOC method vary with interpolation weight β. The best results for each dataset in our experiments are highlighted.

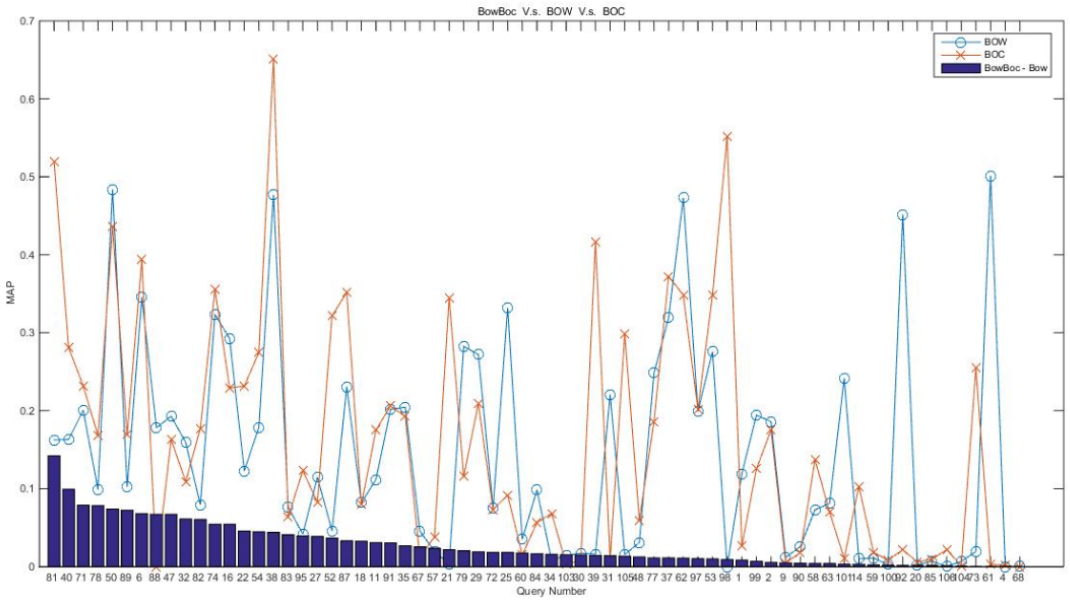


Figure 4.10: The performance of the OHSUMED queries that benefit from the BOW+BOC method. The bars denote their benefit (@MAP) from BOW+BOC run over BOW baseline, while the circles represented the scores of BOW and the crosses refer to the results of BOC approach.

Method	OHSUMED (7:3)		CLEF13 short (9:1)		CLEF13 long (9:1)		CLEF14 short (9:1)		CLEF14 long (9:1)	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
BOW	0.1607	0,2099	0.2844	0.4940	0.2709	0.4680	0.3945	0.7180	0.4026	0.6680
BOW+BOC	0.1838 **	**0.2495 **	0.2512	0.4360	0.2205	0.3960	0.3325	0.6280	0.2913	0.5320
PE	0.1787	0.2228	0.2044	0.3600	0.2089	0.3400	0.2580	0.5400	0.2897	0.5620
BOW+PE	0.1858 **	0.2337 **	0.2895	0.4880	0.2812 *	0.4740	0.4060 *	0.7408	0.4202 *	0.6940 *

Table 4.XI: Result of BOW+BOC approach and BOW+Phrase_Exact (PE) method. The *(7:3)* and *(9:1)* indicate the weight of the original query and of the expanded query. The overall best result in our experiments for each dataset are highlighted. * and ** respectively indicate the statistically significant and extremely significant over BOW baseline in student one-tailed test with $p<0.05$ and $p<0.01$.

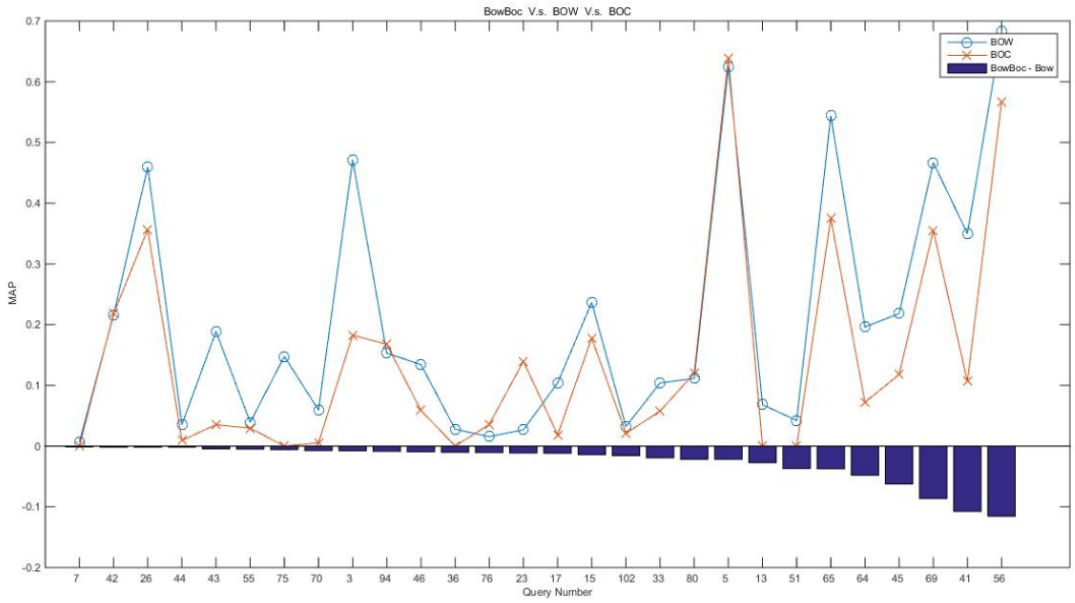


Figure 4.11: The performance of the OHSUMED queries that become worse in BOW+BOC approach. The bars denote their benefit (@MAP) from BOW+BOC run over BOW baseline, while the circles represented the scores of BOW and the crosses refer to the results of BOC approach.

tion weights. OHSUMED collection requires a more important weight of concept phrase expansion: `7:3`, while `9:1` is enough for CLEF 2013 and 2014 collections (As shown in Table 4.XII).

4.2.5 Giving more flexibilities to phrase matching

Table 4.XIII reports the results of different concept phrase matching strategies. We can see that the flexible concept phrases expansion method (*BOW+PP*) can lead to sig-

BOW + Phrase_Exact	OHSUMED		CLEF13 short		CLEF13 long		CLEF14 short		CLEF14 long	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
7:3	**0.1858**	0.2337	0.2828	0.4780	0.2805	0.4660	0.3783	0.7380	0.4037	0.7020
8:2	0.1853	**0.2366**	0.2857	0.4820	**0.2845**	**0.4780**	0.3936	0.7420	0.4188	**0.7100**
9:1	0.1740	0.2228	**0.2895**	**0.4880**	0.2812	0.4740	**0.4060**	**0.7460**	**0.4202**	0.6940

Table 4.XII: The results of BOW+Phrase_Exact (PE) method vary with interpolation weight γ. The optimal result for each dataset are highlighted.

nificant improvement over baseline on all five datasets. And the hybrid phrases combination method performs even better and achieves the overall best results in most of collections. In practice, as usual, the results of concept phrase expansion methods are varied with different interpolation weights. Our results show that (Table 4.XIV and 4.XV) `7:3` is suitable for OHSUMED collection and `9:1` for CLEF datasets.

And then, Table 4.XVI shows how the results vary with different size of windows. We have tried a dynamic unordered window with size N, N+1, and N+2 (N is the length of each concept phrase), as well as a fixed size in the range of 6 and 10. The optimal size is not the same for different datasets. But overall, `#uwN+1` or `#uwN+2` can be a good choice. And we do observe that when we allow more flexibility to concept phrase using proximity matching, word order is not fixed, some additional word can also be inserted, we observe that the expressions in queries and documents can be more easily matched. The BOW+Phrase_Prox method always produce significant improvement over baseline. For example, in BOW+Phrase_Prox method, when the two terms in concept "*C0008679 [Chronic disease]*" are allowed to appear in a window of size 4, a commonly used expressions in relevant documents "*chronic and inapparent disease*" can be matched. For another query "*differential diagnosis elevated alkaline phosphatase ldh levels*", when an important concept "*C0428332 [Alkaline phosphatase level - finding]*" is allowed to appear in a large windows, the expressions such as "*increased intestinal alkaline phosphatase levels*", and "*level of serum alkaline phosphatase is almost invariably elevated*" in the relevant documents can be matched.

We also combine three different phrase matches: Phrase_Exact (PE), Phrase_Prox (PP), and Phrase_Bow (PB). Finally, the *BOW+Phrase_Comb (PC)* method got the overall best result. On all query sets, the results are significantly better than those of BOW, and usually outperform a unique method. In Figure 4.12 and 4.13, we show more details comparing Phrase_Comb against BOW and BOC. The bars show the benefits of Phrase_Comb (PC) method over BOW baseline with descending ranks. The circles represent the MAP in BOW baseline, while crosses refer to those in BOC approach. The combination of three methods with different flexibility could be more effective than each individual components, which is consistent with results of the MRF.

Method	OHSUMED		CLEF13 short		CLEF13 long		CLEF14 short		CLEF14 long	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
BOW	0.1607	0,2099	0.2844	0.4940	0.2709	0.4680	0.3945	0.7180	0.4026	0.6680
MRF	$\lambda_{FI} = 0.8, \lambda_{SD} = 0.1, \lambda_{FD} = 0.1$									
	0.1729 *	0,2297 **	0.2750	0.4680	0.2735	0.4560	0.3904	**0.7620**	0.4099	0.6760

Method	OHSUMED (7:3)		CLEF13 short (9:1)		CLEF13 long (9:1)		CLEF14 short (9:1)		CLEF14 long (9:1)	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
BOW+PE	0.1858 **	0.2337 **	0.2895 M	0.4880 M	0.2812 * M	**0.4740**	0.4060 * M	0.7408	0.4202 * M	0.6940 * M
BOW+PP	**0.1888 ** M**	0.2277 **	**0.2908 * M**	0.4960 M	0.2817 * M	0.4680	0.4075 * M	0.7420	0.4221 ** MM	0.6940 * M
BOW+PB	0.1707 **	0.2267 **	0.2835	**0.5020 M**	0.2719	0.4660	0.3990	0.7260	0.4118 *	0.6820 * M
BOW+PC	$weight_{original} = 0.8, \lambda_1 = 0, \lambda_2 = 0.1, \lambda_3 = 0.1$									
	0.1815 **	0.2267 **	0.2892 M	0.4940 M	**0.2838 ***	0.4700	**0.4137 * MM**	0.7580 *	**0.4316 ** MM**	**0.7180 * M**

Table 4.XIII: Result of concept synonyms expansion as phrases. The *(7:3)* and *(9:1)* indicate the weight of the original query and of the expanded query. The overall best result among our experiments for each dataset are highlighted. *** (**)** and M (MM) respectively indicate the statistically significant improvement over BOW baseline and over MRF in student one-tailed test with p<0.05 (p<0.01).

BOW + Phrase_Prox (#uwN+1)	OHSUMED		CLEF13 short		CLEF13 long		CLEF14 short		CLEF14 long	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
7:3	**0.1888**	**0.2277**	0.2891	0.4880	0.2860	0.4680	0.3846	0.7660	0.4081	0.7080
8:2	0.1866	0.2267	0.2900	0.4860	**0.2870**	**0.4820**	0.3989	**0.7660**	0.4200	**0.7100**
9:1	0.1731	0.2218	**0.2908**	**0.4960**	0.2817	0.4680	**0.4075**	0.7420	**0.4221**	0.6940

Table 4.XIV: The results of BOW+Phrase_Prox (PP) method vary with interpolation weight γ. The optimal choice for each dataset are highlighted.

BOW + Phrase_Bow	OHSUMED		CLEF13 short		CLEF13 long		CLEF14 short		CLEF14 long	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
7:3	**0.1707**	0.2267	**0.2853**	**0.5020**	0.2681	**0.4780**	0.3989	0.7100	0.4026	**0.6900**
8:2	0.1677	0.2149	**0.2853**	**0.5020**	0.2718	**0.4780**	**0.4006**	0.7240	**0.4126**	0.6800
9:1	0.1652	0.2129	0.2835	**0.5020**	**0.2719**	0.4660	0.3990	**0.7260**	0.4118	0.6820

Table 4.XV: The results of BOW+Phrase_Bow (PB) method vary with interpolation weight γ. The highest score for each dataset are highlighted.

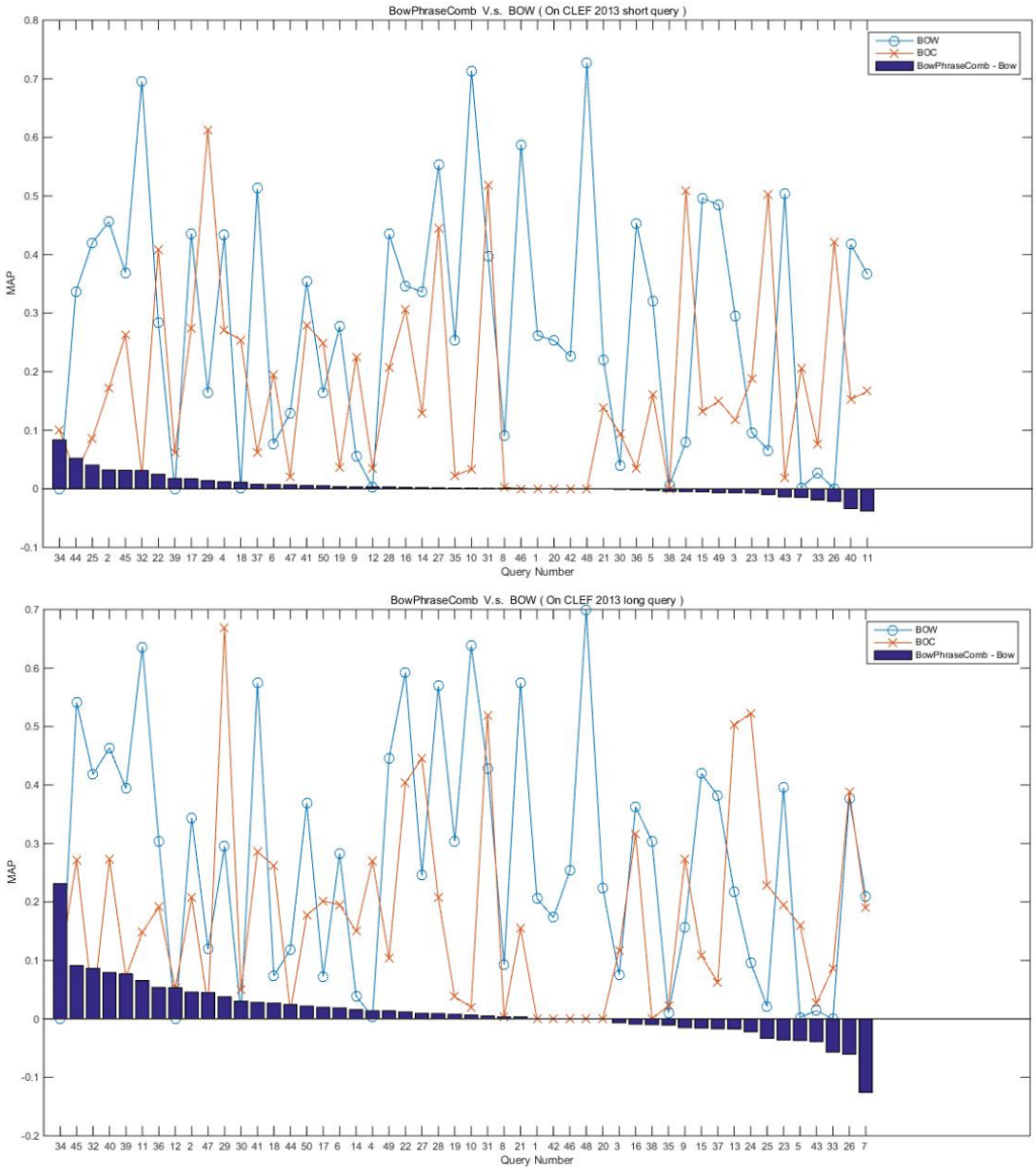


Figure 4.12: The difference between BOW+Phrase_Comb (PC) run and baseline on CLEE 2013 dataset (shown as the bars). The circles and the crosses respectively denote the performance of each query on the BOW baseline and the BOC approach.

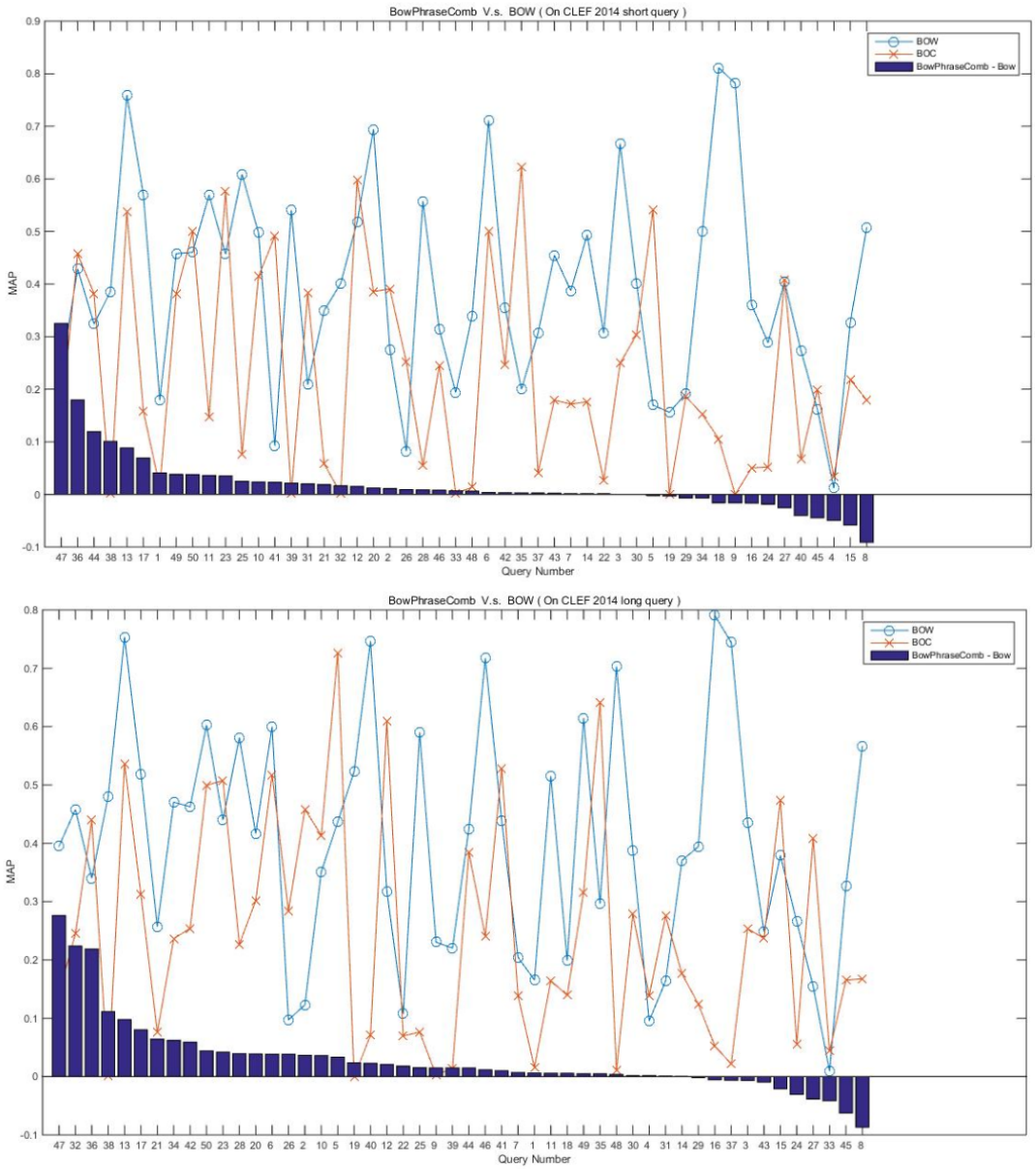


Figure 4.13: The difference between BOW+Phrase_Comb (PC) run and baseline on CLEE 2014 dataset (shown as the bars). The circles and the crosses respectively denote the performance of each query on the BOW baseline and the BOC approach.

BOW + Phrase_Bow	OHSUMED (7:3)		CLEF13 short (9:1)		CLEF13 long (9:1)		CLEF14 short (9:1)		CLEF14 long (9:1)	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
uwN	0.1862	0.2327	0.2899	0.4920	0.2824	0.4680	**0.4076**	**0.7480**	0.4215	0.6960
uwN+1	**0.1888**	0.2277	0.2908	0.4960	0.2817	0.4680	0.4075	0.7420	0.4221	0.6940
uwN+2	0.1884	0.2277	0.2914	0.4960	0.2820	0.4720	0.4045	0.7320	0.4219	0.6900
uw5	0.1883	0.2267	0.2910	0.4960	0.2812	0.4720	0.4055	0.7340	0.4228	0.6900
uw6	0.1883	0.2287	**0.2917**	**0.4980**	0.2824	0.4720	0.4055	0.7360	**0.4237**	0.6960
uw7	0.1878	0.2307	0.2914	0.4960	0.2827	0.4720	0.4066	0.7420	0.4236	**0.6980**
uw8	0.1874	0.2307	0.2914	**0.4980**	0.2829	0.4720	0.4057	0.7360	0.4231	0.6960
uw9	0.1875	0.2317	0.2913	0.4940	0.2846	0.4740	0.4055	0.7380	0.4228	0.6940
uw10	0.1877	0.2337	0.2914	0.4960	**0.2849**	**0.4760**	0.4064	0.7360	0.4231	0.6960
uw	0.1879	**0.2366**	0.2822	0.4800	0.2798	0.4660	0.4002	0.7340	0.4131	0.6940

Table 4.XVI: The BOW+Phrase_Prox (PP) method obtains different results when different windows are used. The highest score for each dataset are highlighted.

If we compare our concept phrases matching approach with MRF, which has been used in some previous studies that produced the best results in CLEF 2013 and TREC 2014, we find that in most of the case, retrieving concepts as phrases can produce significantly better results. This result demonstrates strongly the potential benefits of using existing knowledge resources, especially when concepts are retrieved as flexible phrases. It proves that instead of using arbitrary adjacent words in queries (like MRF), concept phrases can provide real dependences in the query, eliminate noisy phrases and enrich the query expressions by using synonyms.

The result of the Phrase_Comb method in Table 4.XIII and in Table 4.III is produced with a fixed weighting schema: (0.8, 0, 0.1, 0.1). We also tested a range of values for the parameters varying within {0, 0.025, 0.05, 0.075,...,1.0}. The BOW component always takes the largest part in the best combinations. The Figure 4.14 shows the improvements of Phrase_Comb method over baseline at MAP (Original queries in BOW space are assigned the weight of 0.8 and 0.9, which means $\lambda 1+\lambda 2+\lambda 3=0.2$ and $\lambda 1+\lambda 2+\lambda 3=0.1$). The location of the spots denotes three different weights $\lambda 1$, $\lambda 2$, $\lambda 3$ for *Phrase_Exact*, *Phrase_Prox* and *Phrase_Bow*. The depth of color reflect the degree of the improvement. According to the color bar, the deeper the color, the greater the improvement. Table 4.XVII lists some results in detail. We can see that, for OHSUMED and CLEF

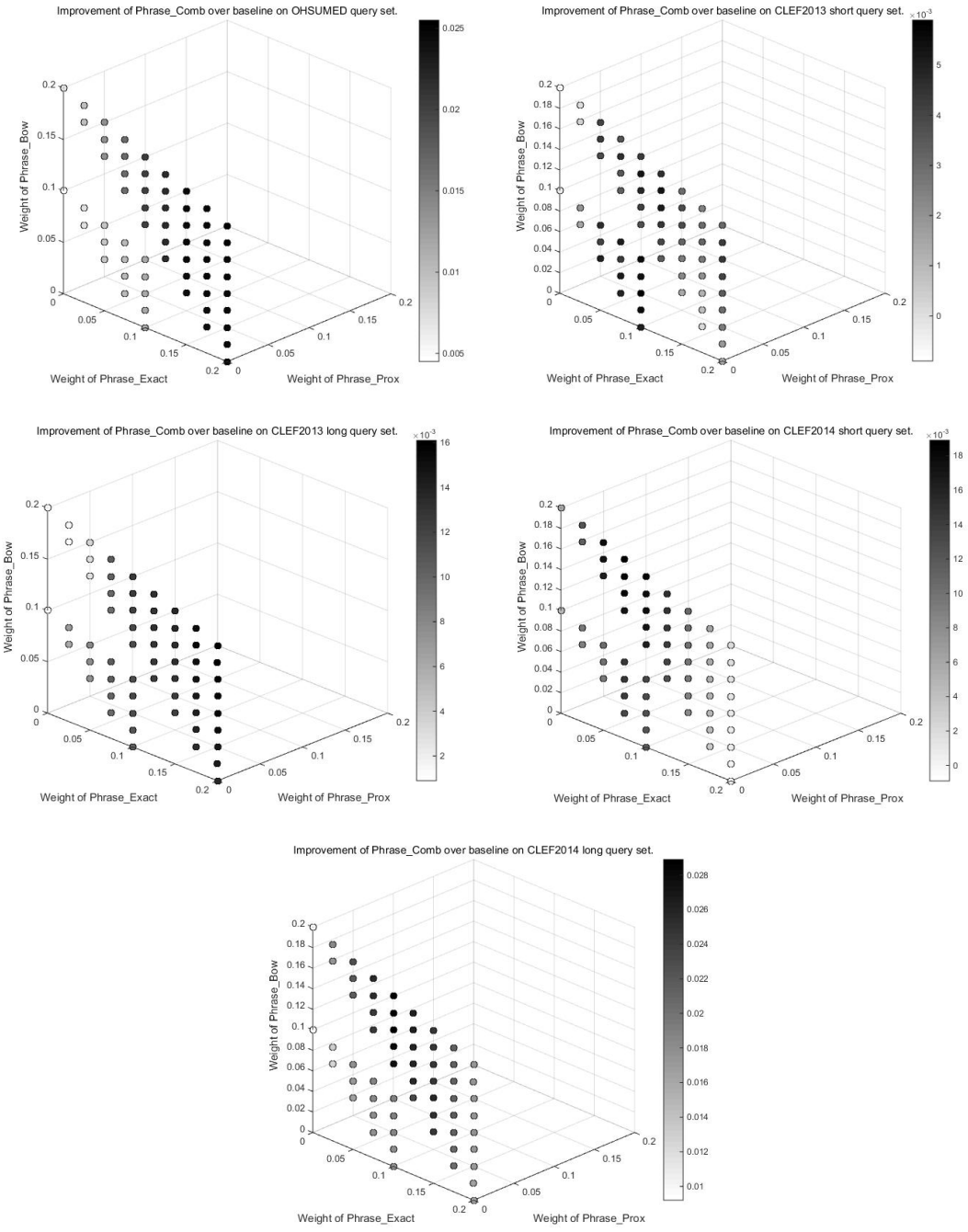


Figure 4.14: The spots represent the improvements of Phrase_Comb method over baseline at MAP. The location of the spots denotes three different weights $\lambda 1$, $\lambda 2$, $\lambda 3$ for *Phrase_Exact*, *Phrase_Prox* and *Phrase_Bow*. The depth of color reflect the degree of the improvement. The deeper the color, the greater the improvement is.

2013 long query sets, the best weights are (0.8, 0, 0.2, 0) for the four components in Formula 3.9, i.e. 0.8 is assigned to BOW, 0 to *Phrase_Exact* and 0.2 to *Phrase_Prox* and 0 to *Phrase_Bow* (when the other two components are assigned weight of zero, which is equal to the Phrase_Prox approach). For CLEF 2013 short queries, the optimal parameters are (0.9, 0, 0.1, 0). And for CLEF 2014 collections, the best weights are (0.8, 0, 0.1, 0.1). The small weights assigned to phrases mean that phrases just play a limited role to help re-rank the retrieval results of BOW in favor of the documents containing the phrases. As for the three phrase matching methods, there is no unique best weighting scheme for all different data set. But overall, the*Phrase_Prox* is more important than the *Phrase_Exact*, and should be given more weight. As for *Phrase_Bow*, it seems also useful to some extent. Our results suggest that (0.8, 0, 0.1, 0.1) could be a suitable weight on our three test collections.

In our combination method, the interpolation parameters are set globally regardless to the query at hand. A more reasonable approach to be investigated in the future is to determine the parameters according to the characteristics of the query. We expect such an approach to produce further improvements in MIR.

4.2.6 Our experience in the ShARe/CLEF eHealth Evaluation Lab 2014

We have participated in the ShARe/CLEF eHealth Evaluation Lab 2014. Our official submission to ShARe/CLEF eHealth Evaluation Lab 2014 - *GRIUM_EN_Run.5* [87] used the *BOW+Phrase_Prox (PP)* method with a weight of 0.8 for BOW part and 0.2 for the phrase expansion. These parameters have been tuned using CLEF2013 queries. Our run obtained the best result among all the submissions. It is interesting to notice that the best submissions in CLEF 2013 - Runs Team Mayo [110] used a Markov random field model and the SNUMEDINFO runs [18] runs also used concept expressions as phrases, similarly to our combination method. All these results confirm the usefulness of concepts when used as phrases. As a reference, we show the 5 best submissions in CLEF 2013 [94] and CLEF 2014 [40] in Table 4.XVIII and 4.XIX.

Notice also that the best runs in both TeamMayo and SNUMEDINFO used discharge summaries, which are not used in our experiments. In addition, some TeamMayo runs

BOW + Phrase_ Comb	OHSUMED		CLEF13 short		CLEF13 long		CLEF14 short		CLEF14 long	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
8:0:0:2	0.1667	0.2149	0.2835	**0.5020**	0.2718	0.4780	0.4006	0.7240	0.4126	0.6800
8:0:1:1	0.1815	0.2267	0.2892	0.4940	0.2838	0.4700	**0.4137**	0.7580	**0.4316**	**0.7180**
8:0:2:0	**0.1866**	0.2267	0.2900	0.4860	**0.2870**	0.4820	0.3989	**0.7660**	0.4200	0.7100
8:1:1:0	0.1857	0.2366	0.2892	0.4900	0.2863	**0.4820**	0.3983	0.7600	0.4215	0.7100
8:1:0:1	0.1808	0.2307	0.2881	0.4920	0.2824	0.4620	0.4109	0.7560	0.4302	0.7100
8:2:0:0	0.1853	**0.2366**	0.2857	0.4820	0.2845	0.4780	0.3936	0.7420	0.4188	0.7100
9:0:0:1	0.1652	0.2129	0.2840	0.4940	0.2719	0.4660	0.3990	0.7260	0.4118	0.6820
9:0:0.5:0.5	0.1702	0.2188	0.2902	0.4980	0.2806	0.4680	0.4048	0.7320	0.4202	0.6880
9:0:1:0	0.1731	0.2218	**0.2908**	0.4960	0.2817	0.4680	0.4075	0.7420	0.4221	0.6960
9:0.5:0.5:0	0.1732	0.2208	0.2904	0.4920	0.2821	0.4720	0.4084	0.7480	0.4215	0.6960
9:0.5:0:0.5	0.1702	0.2178	0.2894	0.4940	0.2784	0.4660	0.4040	0.7340	0.4202	0.6880
9:1:0:0	0.1740	0.2228	0.2895	0.4880	0.2812	0.4740	0.4060	0.7460	0.4202	0.6940

Table 4.XVII: Some detailed results of BOW+Phrase_Comb (PC) method with different interpolation parameters. The optimal result for each dataset are highlighted.

Run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	Rel Ret
Team Mayo 2.3	0,4960	0,5180	0,4391	0,4665	0,3108	1673
Team Mayo 5.3	0,5120	0,5040	0,4645	0,4618	0,3061	1689
Team Mayo 6.3	0,5160	0,4940	0,4639	0,4579	0,2953	1689
Team Mayo 3.3	0,5280	0,4880	0,4742	0,4584	0,2900	1689
Team Mayo 4.3	0,5240	0,4820	0,4837	0,4637	0,2967	1689

Table 4.XVIII: Top 5 submissions in CLEF 2013 task 3a.

Run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	Rel Ret
GRIUM EN Run.5	0,7680	0,7560	0,7423	0,7445	0,4016	2550
SNUMEDINFO CZ Run.5	0,7592	0,7551	0,6998	0,7011	0,3494	2147
SNUMEDINFO EN Run.2	0,7840	0,7540	0,7502	0,7406	0,3753	2307
SNUMEDINFO EN Run.5	0,8160	0,7520	0,7749	0,7426	0,3814	2305
SNUMEDINFO CZ Run.6	0,7388	0,7469	0,6834	0,6871	0,3395	2147

Table 4.XIX: Top 5 submissions in CLEF 2014 task 3a.

used pseudo relevance feedback (relevance model [59]), which we do not use. Despite the fact that less information is used, our combination results are only slightly lower than the best runs in CLEF 2013. In addition our latter phrase combination method further boost the overall best performance.

Compared with the previous works the concept-based query expansion (described in section 2.5.2), where significant improvement is rarely observed, our experiments mainly have two differences: (1) Different from a traditional query expansion method which directly add synonyms into BOW space, expanding concept synonyms in the form of phrase takes into account the dependency between concept terms, meanwhile keeps the full flexibility on phrase matching strategy. (2) Instead of using only the asserted synonyms or even the concept preferred name as the set of concept phrase, we extracted all corresponding strings of a concept which include not only the synonyms but also all variations.

CHAPTER 5

CONCLUSION AND FUTURE WORK

Medical IR is highly demanding and can possibly benefit from very rich knowledge resources. Previous studies have tested different ways to take advantage of the existing resources in medical area to improve the retrieval effectiveness of traditional BOW methods. However, in some studies, concepts have been able to slightly improve the retrieval effectiveness, while in some others, they degrade the effectiveness instead. In this study, we aim at comparing different methods to use concepts in the same framework. Our results showed that the traditional BOW methods could be significantly improved by incorporating the identified concepts as phrases. By matching the phrases in documents in either strict or flexible ways, concepts can help better determine relevant documents. This method led to significant improvements on 4 of the 5 queries sets. This confirms the usefulness of concepts in MIR.

Furthermore, within the same framework, different matching methods were systematically compared with each other for the first time. Our results showed that replacing the traditional bag of words by bag of concept IDs did not yield good results. This indicates that a rigid use of concept IDs may not be suitable for MIR. An important problem we observed in our experiments is related to the concept mapping process. In a number of cases, the concept mapping tool was unable to identify the concepts correctly and consistently. This problem greatly reduced the potential benefits of concepts in MIR. As described in this thesis natural language is highly varied, so queries and relevant document can be hardly matched through identified concepts. This observation is consistent with the observations in the previous studies. To improve concept-based MIR, we have tried to expand concepts by their hyponyms or to combine BOW and BOC space. But no significant improvement is observed until we finally used concepts as phrase. Actually, the flexible phrase matching strategy can be seen as a kind of compromise between BOW and BOC approaches, which keeps a balance between the recall and precision. It can benefit from the terms dependency meanwhile maximize the concept expressions

coverage in the relevant document.

Despite the good results we obtained in our experiments, our study has a number of limitations. First, we only aimed at showing the potential of a concept-based approach to MIR by examining a series of simple methods. The method can be improved in several ways in the future.

First, we can work on finding a way to determine the best parameters for a combination, and using a more sophisticated way to combine different uses of concepts. For example, concepts could be differently weighted according to their semantic type, etc.

Second, the queries used in our tests have been created by healthcare professionals, and are assumed to be formulated correctly (although some spelling errors appeared in them). In practice, laypeople often do not know what concepts to include in a query and what terms to use to express the concepts [102]. There may be a much larger vocabulary discrepancy between queries and documents than observed in our experiments. The methods described in this paper, which rely on concept mapping, will fail in many practical cases. A crucial aspect for practical MIR is to connect the vocabulary used by laypeople users to that of the authors. Notice that the new CLEF eHealth evaluation lab in 2015 will focus on user-centered Medical IR task, in their queries a long, ambiguous wording is used in place of the actual medical term to refer to a condition or disease. A possible solution is to exploit search logs, which record the user queries and the documents they clicked on. Such information has been found valuable in general web search [39] to link the terms in user's queries and the terms in documents. A similar approach could be used in MIR.

BIBLIOGRAPHY

[1] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

[2] Sophia Ananiadou. A methodology for automatic term recognition. In *Proceedings of the 15th conference on Computational linguistics-Volume 2*, pages 1034–1038. Association for Computational Linguistics, 1994.

[3] Alan R. Aronson. The effect of textual variation on concept based information retrieval. In *Proceedings of the AMIA Annual Fall Symposium*, page 373. American Medical Informatics Association, 1996.

[4] Alan R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[5] Alan R. Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[6] Alan R. Aronson and Thomas C Rindflesch. Query expansion using the umls metathesaurus. In *Proceedings of the AMIA Annual Fall Symposium*, page 485. American Medical Informatics Association, 1997.

[7] Alan R. Aronson, Dina Demner-Fushman, Susanne M Humphrey, Jimmy J Lin, Patrick Ruch, Miguel E Ruiz, Lawrence H Smith, Lorraine K Tanabe, W John Wilbur, and Hongfang Liu. Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. In *TREC*, 2005.

[8] Amos Bairoch, Rolf Apweiler, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo

Lopez, Michele Magrane, et al. The universal protein resource (uniprot). *Nucleic acids research*, 33(suppl 1):D154–D159, 2005.

[9] Steven Bedrick and G. Sheikhshabbafghi. Lucene, metamap, and language modeling: Ohsu at clef ehealth 2013. *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*, 2013.

[10] Steven Bedrick, Tracy Edinger, Aaron Cohen, and William Hersh. Identifying patients for clinical studies from electronic health records: Trec 2012 medical records track at ohsu. Technical report, DTIC Document, 2012.

[11] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[12] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[13] S.A. Boyadjiev and E.W. Jabs. Online mendelian inheritance in man (omim) as a knowledgebase for human developmental disorders. *Clinical genetics*, 57(4): 253–266, 2000.

[14] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[15] Steven H. Brown, Peter L. Elkin, S.T. Rosenbloom, C. Husser, B.A. Bauer, MJ Lincoln, J Carter, M Erlbaum, and MS Tuttle. Va national drug file reference terminology: a cross-institutional content coverage study. *Medinfo*, 11(Pt 1): 477–81, 2004.

[16] James P. Callan, W. Bruce Croft, and Stephen M. Harding. The inquery retrieval system. In *Database and expert systems applications*, pages 78–83. Springer, 1992.

[17] P. Callejas, A. Miguel, Yue Wang, and Hui Fang. Exploiting domain thesaurus for medical record retrieval. Technical report, DTIC Document, 2012.

[18] Sungbin Choi and Jinwook Choi. Exploring effective information retrieval technique for the medical web documents: Snumedinfo at clefehealth2014 task 3. *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*, 2014.

[19] Sungbin Choi, Jeongeun Lee, and Jinwook Choi. Snumedinfo at imageclef 2013: Medical retrieval task. In *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*, 2013.

[20] Vincent Claveau, Thierry Hamon, Natalia Grabar, and Sébastien Le Maguer. Repali participation to clef ehealth ir challenge 2014: leveraging term variation. In *Conference and Labs of the Evaluation Forum CLEF*, pages 13–p, 2014.

[21] Aaron M Cohen and William R Hersh. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71, 2005.

[22] Nigel Collier, Chikashi Nobata, and Jun-ichi Tsujii. Extracting the names of genes and gene products with a hidden markov model. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 201–207. Association for Computational Linguistics, 2000.

[23] Gene Ontology Consortium. Gene ontology, Last visit: July 2015. URL `http://geneontology.org/`.

[24] Berry De Bruijn and Joel Martin. Getting to the (c) ore of knowledge: mining biomedical literature. *International journal of medical informatics*, 67(1):7–18, 2002.

[25] Dina Demner-Fushman, Swapna Abhyankar, Antonio Jimeno-Yepes, Russell F Loane, Bastien Rance, François-Michel Lang, Nicholas C Ide, Emilia Apostolova, and Alan R. Aronson. A knowledge-based approach to medical records retrieval. In *TREC*, 2011.

[26] Dina Demner-Fushman, Swapna Abhyankar, Antonio Jimeno-Yepes, Russell Loane, Francois Lang, James G Mork, Nicholas Ide, and Alan R. Aronson. Nlm at trec 2012 medical records track. Technical report, DTIC Document, 2012.

[27] Alberto Diaz, Miguel Ballesteros, Jorge Carrillo-de Albornoz, and Laura Plaza. Ucm at trec-2012: Does negation influence the retrieval of medical reports? Technical report, DTIC Document, 2012.

[28] Khadim Dramé, Fleur Mougin, and Gayo Diallo. Query expansion using external resources for improving information retrieval in the biomedical domain. *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*, 2014.

[29] EMBL-EBI. Gene ontology annotation (uniprot-goa) database, Last visit: July 2015. URL `http://www.ebi.ac.uk/GOA`.

[30] Tina A Eyre, Fabrice Ducluzeau, Tam P Sneddon, Sue Povey, Elspeth A Bruford, and Michael J Lush. The hugo gene nomenclature database, 2006 updates. *Nucleic acids research*, 34(suppl 1):D319–D321, 2006.

[31] Jenny Rose Finkel and Christopher D Manning. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 141–150. Association for Computational Linguistics, 2009.

[32] The National Center for Biotechnology Information. Pubmed quick start, Last visit: July 2015. URL `http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.PubMed_Quick_Start`.

[33] The National Center for Biotechnology Information. Medline, Last visit: July 2015. URL `http://www.ncbi.nlm.nih.gov/pubmed`.

[34] The National Center for Biotechnology Information. Gene, Last visit: July 2015. URL `http://www.ncbi.nlm.nih.gov/gene/2054`.

[35] Centers for Disease Control and Prevention. Cdc home, Last visit: July 2015. URL `http://www.cdc.gov/nchs/icd/icd9cm.html`.

[36] Susannah Fox. *Health topics: 80% of internet users look for health information online*. Pew Internet & American Life Project, 2011.

[37] Ken-ichiro Fukuda, Tatsuhiko Tsunoda, Ayuchi Tamura, Toshihisa Takagi, et al. Toward information extraction: identifying protein names from biological papers. In *Pac Symp Biocomput*, volume 707, pages 707–718. Citeseer, 1998.

[38] Robert Gaizauskas, George Demetriou, and Kevin Humphreys. Term recognition and classification in biological science journal articles. In *In Proc. of the Computional Terminology for Medical and Biological Applications Workshop of the 2 nd International Conference on NLP*. Citeseer, 2000.

[39] Jianfeng Gao, Xiaodong He, and Jian-Yun Nie. Clickthrough-based translation models for web search: from word models to phrase models. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1139–1148. ACM, 2010.

[40] Lorraine Goeuriot, Liadh Kelly, Wei Li, Joao Palotti, Pavel Pecina, Guido Zuccon, Allan Hanbury, Gareth JF Jones, and Henning Mueller. Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval. 2014.

[41] Jörg Hakenberg, Conrad Plake, Ulf Leser, Harald Kirsch, and Dietrich Rebholz-Schuhmann. Lll'05 challenge: Genic interaction extraction-identification of language patterns based on alignment and finite state automata. In *Proceedings of the 4th Learning Language in Logic workshop (LLL05)*, pages 38–45, 2005.

[42] Hussam Hamdan, Shereen Albitar, Patrice Bellot, Bernard Espinasse, and Sébastien Fournier. Lsis at trec 2012 medical track-experiments with conceptualization, a dfr model and a semantic measure. Technical report, DTIC Document, 2012.

[43] Leading healthcare terminology. ihtsdo, Last visit: July 2015. URL `http://browser.ihtsdotools.org/`.

[44] MAAS Schwartz Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. 2003.

[45] William Hersh. A stimulus to define informatics and health information technology. *BMC Medical Informatics and Decision Making*, 9(1):24, 2009.

[46] William R Hersh, David H Hickam, and TJ Leone. Words, concepts, or both: optimal indexing units for automated information retrieval. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 644. American Medical Informatics Association, 1992.

[47] Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. Overview of biocreative: critical assessment of information extraction for biology. *BMC bioinformatics*, 6(Suppl 1):S1, 2005.

[48] Qinmin Hu, Zhi Xu, Xiangji Huang, and Zheng Ye. York university at trec 2012: Crowdsourcing track. Technical report, DTIC Document, 2012.

[49] ihtsdo (Leading healthcare terminology). Snomed ct e-learning server, Last visit: July 2015. URL `http://ihtsdo.org/fileadmin/user_upload/doc/elearning.html`.

[50] Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta, and Jun'ichi Tsujii. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, pages 1–8. Association for Computational Linguistics, 2002.

[51] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9. Association for Computational Linguistics, 2009.

[52] Benjamin King, Lijun Wang, Ivan Provalov, and Jerry Zhou. Cengage learning at trec 2011 medical track. In *TREC*, 2011.

[53] Marjorie King and Renee Shell. Teaching and evaluating critical thinking with concept maps. *Nurse Educator*, 27(5):214–216, 2002.

[54] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[55] Graham Klyne and Jeremy J Carroll. Resource description framework (rdf): Concepts and abstract syntax. 2006.

[56] Bevan Koopman, Peter Bruza, Laurianne Sitbon, and Michael Lawley. Aehrc & qut at trec 2011 medical track: a concept-based information retrieval approach. In *Proceedings of 20th Text REtrieval Conference (TREC 2011)*, pages 1–7. National Institute of Standards and Technology (NIST), 2011.

[57] Bevan Koopman, Guido Zuccon, Anthony Nguyen, Deanne Vickers, Luke Butt, and Peter Bruza. Exploiting snomed ct concepts & relationships for clinical information retrieval: Australian e-health research centre and queensland university of technology at the trec 2012 medical track. Technical report, DTIC Document, 2012.

[58] Michael Krauthammer and Goran Nenadic. Term identification in the biomedical literature. *Journal of biomedical informatics*, 37(6):512–526, 2004.

[59] Victor Lavrenko and W Bruce Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127. ACM, 2001.

[60] Carolyn E. Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[61] Juan Manuel Córdoba Malagón and Manuel Jesús Maña López. Laberinto at share/clef ehealth evaluation lab 2014.

[62] David Martinez, Arantxa Otegi, and Eneko Agirre. Nicta and ubc at the trec 2012 medical track. Technical report, DTIC Document, 2012.

[63] EMR Matrix. Emr reviews nextgen, Last visit: July 2015. URL http://emr-matrix.org/2010/11/nextgen-healthcare-ehr-frontiers-corrections/.

[64] Michael McCandless, Erik Hatcher, and Otis Gospodnetic. *Lucene in Action: Covers Apache Lucene 3.0*. Manning Publications Co., 2010.

[65] Alexa T McCray, Suresh Srinivasan, and Allen C Browne. Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 235. American Medical Informatics Association, 1994.

[66] Alexa T McCray, Suresh Srinivasan, and Allen C Browne. Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 235. American Medical Informatics Association, 1994.

[67] Donald Metzler and W Bruce Croft. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479. ACM, 2005.

[68] NLM (U.S.National Library of Medicine). Entrez gene, Last visit: July 2015. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene.

[69] NLM (U.S.National Library of Medicine). Mesh browser, Last visit: July 2015. URL http://www.nlm.nih.gov/mesh/2015/mesh_browser/MBrowser.html.

[70] NLM (U.S.National Library of Medicine). Medical subject headings, Last visit: July 2015. URL http://www.nlm.nih.gov/mesh/MBrowser.html.

[71] NLM (U.S.National Library of Medicine). Unified medical language system - metathesaurus, Last visit: July 2015. URL http://uts.nlm.nih.gov/metathesaurus.html.

[72] NLM (U.S.National Library of Medicine). Umls load script, Last visit: July 2015. URL http://www.nlm.nih.gov/research/umls/implementation_resources/scripts/index.html.

[73] NLM (U.S.National Library of Medicine). Pubmed database, Last visit: July 2015. URL http://www.ncbi.nlm.nih.gov/pubmed/25734738.

[74] NLM (U.S.National Library of Medicine). The umls semantic network, Last visit: July 2015. URL http://semanticnetwork.nlm.nih.gov/.

[75] NLM (U.S.National Library of Medicine). Snomed ct browser, Last visit: July 2015. URL http://uts.nlm.nih.gov/snomedctBrowser.html.

[76] NLM (U.S.National Library of Medicine). Umls terminology services api 2.0 documentation, Last visit: July 2015. URL http://uts.nlm.nih.gov/home.html#apidocumentation.

[77] U.S.National Library of Medicine. Medlineplus - trusted health information for you, Last visit: July 2015. URL http://www.nlm.nih.gov/medlineplus/.

[78] U.S.National Library of Medicine. Nlm, Last visit: July 2015. URL http://www.nlm.nih.gov.

[79] The Journal of the American Medical Association. Jama, Last visit: July 2015. URL http://jama.jamanetwork.com/journal.aspx.

[80] Heung-Seon Oh and Yuchul Jung. A multiple-stage approach to re-ranking clinical documents. *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*, 2014.

[81] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[82] Wanda Pratt and Meliha Yetisgen-Yildiz. A study of biomedical concept identification: Metamap vs. people. In *AMIA Annual Symposium Proceedings*, volume 2003, page 529. American Medical Informatics Association, 2003.

[83] Yanjun Qi and Pierre-François Laquerre. Retrieving medical records with sennamed: Nec labs america at trec 2012 medical records track. Technical report, DTIC Document, 2012.

[84] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109, 1995.

[85] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[86] Martijn J Schuemie, Jan A Kors, and Barend Mons. Word sense disambiguation in the biomedical domain: an overview. *Journal of Computational Biology*, 12 (5):554–565, 2005.

[87] Wei Shen, Jian-Yun Nie, Xiaohua Liu, and X Liui. An investigation of the effectiveness of concept-based approach in medical information retrieval grium@ clef2014ehealthtask 3. *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*, 2014.

[88] Matthew S Simpson and Dina Demner-Fushman. Biomedical text mining: A survey of recent progress. In *Mining text data*, pages 465–517. Springer, 2012.

[89] Kent A Spackman, Keith E Campbell, and Roger A Côté. Snomed rt: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association, 1997.

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

[91] Samuel Alan Stewart, Maia Elizabeth Von Maltzahn, and SS Raza Abidi. Comparing metamap to mgrep as a tool for mapping free text to formal medical lexicons. In *Proceedings of the 1st international workshop on knowledge extraction & consolidation from social-media in conjunction with the 11th international semantic web conference (ISWC 2012), Boston, USA*, pages 63–77. Citeseer, 2012.

[92] Nicola Stokes, Yi Li, Lawrence Cavedon, and Justin Zobel. Exploring criteria for successful query expansion in the genomic domain. *Information Retrieval*, 12(1):17–50, 2009.

[93] Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, volume 2, pages 2–6. Citeseer, 2005.

[94] Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. Overview of the share/clef ehealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231. Springer, 2013.

[95] Harsh Thakkar, Ganesh Iyer, Kesha Shah, and Prasenjit Majumder. Team irlabdaiict at share/-clef ehealth 2014 task 3: User-centered information retrieval system for clinical documents. *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*, 2014.

[96] Erik F Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language*

learning-Volume 7, pages 127–132. Association for Computational Linguistics, 2000.

[97] Howard Turtle and W Bruce Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems (TOIS)*, 9(3): 187–222, 1991.

[98] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[99] E Voorhees and R Tong. Overview of the trec 2011 medical records track. In *Proc. of TREC*, 2011.

[100] Ellen M Voorhees. Query expansion using lexical-semantic relations. In *SIGIRâĂŹ94*, pages 61–69. Springer, 1994.

[101] Jinxi Xu and W Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11. ACM, 1996.

[102] Q. Zeng, S. Kogan, N. Ash, R.A. Greenes, and A.A. Boxwala. Characteristics of consumer terminology for health information retrieval. *Methods of information in medicine*, 41(4):289–298, 2002.

[103] ChengXiang Zhai. Statistical language models for information retrieval (synthesis lectures series on human language technologies). *Morgan & Claypool Publishers*, (2):1, 2008.

[104] Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410. ACM, 2001.

[105] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM, 2001.

[106] Xiaoshi Zhong, Yunqing Xia, Zhongda Xie, Sen Na, Qinan Hu, and Yaohai Huang. Concept-based medical document retrieval: Thcib at clef ehealth lab 2013 task 3. *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*, 2013.

[107] Guodong Zhou, Jie Zhang, Jian Su, Dan Shen, and Chewlim Tan. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7): 1178–1190, 2004.

[108] Wei Zhou, Vetle I. Torvik, and Neil R. Smalheiser. Adam: another database of abbreviations in medline. *Bioinformatics*, 22(22):2813–2818, 2006.

[109] Dongqing Zhu and Ben Carterette. Exploring evidence aggregation methods and external expansion sources for medical record search. Technical report, DTIC Document, 2012.

[110] Dongqing Zhu, S Wu, Masanz James, Ben Carterette, and Hongfang Liu. Using discharge summaries to improve information retrieval in clinical domain. *Proceedings of the ShARe/-CLEF eHealth Evaluation Lab*, 2013.

[111] Donqqing Zhu and Ben Carterette. Improving health records search using multiple query expansion collections. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–7. IEEE, 2012.

[112] G Zuccon, B Koopman, and A Nguyen. Retrieval of health advice on the web: Aehrc at share/clef ehealth evaluation lab task 3. In *Proceedings of CLEF Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis*, 2013.

[113] Guido Zuccon, Bevan Koopman, Anthony Nguyen, Deanne Vickers, and Luke Butt. Exploiting medical hierarchies for concept-based information retrieval. In

Proceedings of the Seventeenth Australasian Document Computing Symposium, pages 111–114. ACM, 2012.

[114] Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B Cohen. Frontiers of biomedical text mining: current progress. *Briefings in bioinformatics*, 8(5):358–375, 2007.