Mining the Medical and Patent Literature to Support Healthcare and Pharmacovigilance

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Harsha Gurulingappa
aus
Davangere, Indien

Bonn 2012

Angefertigt mit Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen
Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. rer. nat. Martin Hofmann-Apitius

2. Gutachter: Prof. Dr. med. Alexander Pfeifer

Tag der Promotion: 29.06.12

Erscheinungsjahr: 2012

Abstract

Recent advancements in healthcare practices and the increasing use of information technology in the medical domain has lead to the rapid generation of free-text data in forms of scientific articles, e-health records, patents, and document inventories. Capturing the valuable and novel information from these document sources can have several benefits such as medical decision support, timely alerts, and reduced medication errors that in-turn can enhance the effectiveness of clinical care and reduce the treatment costs. These inherent benefits and the sheer amount of free-text data has urged the development of sophisticated information retrieval and information extraction technologies.

A fundamental requirement for the automatic processing of biomedical text is the identification of information carrying units such as the concepts or named entities. In this context, this work focuses on the identification of medical disorders (such as diseases and adverse effects) which denote an important category of concepts in the medical text. Two methodologies were investigated in this regard and they are dictionary-based and machine learning-based approaches. Abilities of standard medical terminologies and the Conditional Random Fields (CRF) to support the recognition of disorders were examined on a common platform. An outcome of this investigation showed that a hybrid approach that utilizes the strengths of dictionaries and CRF is highly suitable for the disorder recognition in free-text. Furthermore, the capabilities of this hybrid model were customized for the recognition of different categories of medical concepts (such as medical problems, treatments, and tests) in e-health records. Within the same framework, a Support Vector Machine (SVM)-based system was implemented for the classification of assertions made over medical problems (such as negations and uncertainties). Performances of the adapted systems for concept identification and assertion classification in e-health records were evaluated as a part of open assessment challenge (*i.e.* I2B2 2010) where both demonstrated highly competitive results in comparison to several state-of-the-art medical information extraction technologies. The developed strategies can be integrated into the semantic information retrieval and information extraction platforms for improved literature searches in the medical domain.

A precise semantic platform for searching and retrieval of concise information from voluminous biomedical document archives can inherently support researchers and medical professionals to fetch the exquisite knowledge quickly. In this context, capabilities of the concept recognition techniques were systematically exploited to build a semantic search platform for the retrieval of e-health records. The system facilitates conventional text search as well as semantic and ontological searches. Later on, capabilities of the retrieval platform were scaled for searching and retrieval from

biomedical and chemical patents. Performance of the adapted retrieval platform for e-health records and patents was evaluated within open assessment challenges (*i.e.* TRECMED and TRECCHEM respectively) wherein the system was best rated in comparison to several other competing information retrieval platforms.

Finally, from the medico-pharma perspective, adverse effects of medications is a challenging issue that confronts healthcare and pharmaceutical industries. Therefore, a strategy for the identification of adverse drug events from medical case reports was developed. Considering the extremely limited availability of the annotated textual data for training an information extraction system for drug safety research, a sufficiently large corpus containing double annotated medical case reports was generated. Later on, the corpus was applied for the development of a Maximum Entropy-based model for the identification of adverse event assertive sentences. Qualitative evaluation as well as an expert validation of the system's performance showed robust results. It allows the development of alerting systems capable of capturing the drug safety issues published in different literature sources.

In conclusion, this thesis presents approaches for efficient information retrieval and information extraction from various biomedical literature sources in the support of healthcare and pharmacovigilance. The applied strategies have potential to enhance the literature-searches performed by biomedical, healthcare, and patent professionals. This can promote the literature-based knowledge discovery, improve the safety and effectiveness of medical practices, and drive the research and development in medical and healthcare arena.

Publication Record

1. Harsha Gurulingappa, Bernd Müller, Roman Klinger, Heinz-Theodor Mevissen, Martin Hofmann-Apitius, Juliane Fluck, and Christoph M. Friedrich. Patent Retrieval in Chemistry based on semantically tagged Named Entities. In The Eighteenth Text RETrieval Conference (TREC 2009) Proceedings, Gaithersburg, Maryland, USA, 2009.

2. Harsha Gurulingappa, Roman Klinger, Martin Hofmann-Apitius, and Juliane Fluck. An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature. In 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining (7th edition of the Language Resources and Evaluation Conference), Valetta, Malta, 2010.

3. Harsha Gurulingappa, Bernd Müller, Roman Klinger, Heinz-Theodor Mevissen, Martin Hofmann-Apitius, Christoph M. Friedrich, and Juliane Fluck. Prior Art Search in Chemistry Patents Based on Semantic Concepts and Co-Citation Analysis. In The Nineteenth Text RETrieval Conference (TREC 2010) Proceedings, Gaithersburg, Maryland, USA, 2010.

4. Harsha Gurulingappa, Martin Hofmann-Apitius, and Juliane Fluck. Concept Identification and Assertion Classification in Patient Health Records. In Proceedings of the 2010 I2B2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA, 2010.

5. Bernd Müller, Roman Klinger, Harsha Gurulingappa, Heinz-Theodor Mevissen, Martin Hofmann-Apitius, Juliane Fluck, and Christoph M. Friedrich. Abstracts versus Full Texts and Patents: A Quantitative Analysis of Biomedical Entities. In Proceedings of the 1st IRF Conference, Lecture Notes in Computer Science. Springer, 2010.

6. Mihai Lupu, Harsha Gurulingappa, Igor Filippov, Zhao Jiashu, Juliane Fluck, Marc Zimmermann, Jimmy Huang, and John Tait. Overview of the TREC 2011 Chemical IR Track. In The Twentieth Text RETrieval Conference (TREC 2011) Proceedings, Gaithersburg, Maryland, USA, 2011.

7. Harsha Gurulingappa, Bernd Müller, Martin Hofmann-Apitius, and Juliane Fluck. Information Retrieval Framework for Technology Survey in Biomedical and Chemistry Literature. In The Twentieth Text RETrieval Conference (TREC 2011) Proceedings, Gaithersburg, Maryland, USA, 2011.

8. Harsha Gurulingappa, Bernd Müller, Martin Hofmann-Apitius, and Juliane Fluck. A Semantic Platform for Information Retrieval from E-Health Records. In The Twentieth Text RETrieval Conference (TREC 2011) Proceedings, Gaithersburg, Maryland, USA, 2011.

9. Harsha Gurulingappa, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Identification of Adverse Drug Event Assertive Sentences in Medical Case Reports. In First International Workshop on Knowledge Discovery and Health Care Management (KD-HCM), European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 2011.

10. Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a Benchmark Corpus to Support the Automatic Extraction of Drug-related Adverse Effects from Medical Case Reports. Journal of Biomedical Informatics. (Accepted) In Press, 2012.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview on Healthcare and Medicine

Medicine is a science and art of healing that encompass various healthcare practices applied to maintain the health of individuals or population by prevention and treatment of health disturbances. Contemporary medicine uses healthcare sciences, biomedical sciences, medical technology, pharmaceutical sciences, and information technology to diagnose and treat various forms of medical problems [Schlich (2007), Ellner and Joyner (2012)]. Healthcare is provided by medical practitioners, dentistry, nursing, pharmacy, allied health professionals as well as solitary care providers. Healthcare is not only associated with personalized care of patients or population but also accounts for country's economy [Arrow et al. (2009)]. In the year 2010, healthcare industry contributed to over 10% of the Gross Domestic Product (GDP)[1] across lot of developed countries[2]. Healthcare is therefore regarded as a major determinant in promoting good health and well being of people around the world. An example is malaria eradication in 1980s that has been declared by World Health Organization (WHO)[3] as a first disease in the human history to be eliminated by deliberate medical interventions [Arita (2011)].

The goal of any medical practice is to promote health tranquility to the public. Nevertheless, sometimes the medical interventions can result in failure to deliver the expected results or may cause unexpected deleterious effects [Vincent et al. (2001)]. Adverse effects are unexpected harmful effects resulting from any medical intervention to the patients [Marken and Pies (2006), Poppenga (2001)]. They are sometimes referred to as *iatrogenic* because they are generated by a physician or treatment[4]. Adverse effects pose major challenges to the healthcare industry. The problem of adverse effects is more severe in the pharmaceutical domain associated with pharmaceutical preparations [Vervloet and Durham (1998)], however, they are not confined to any particular type of medical treatment. Examples of adverse effects include abdominal pain resulting from a drug; infection, inflammation, or scarring resulting from a surgery; or perforation of the intestinal wall resulting from a diagnostic procedure such as colonoscopy.

Pharmacovigilance is a healthcare research that deals with detection, assessment,

[1] http://en.wikipedia.org/wiki/Gross_domestic_product
[2] http://www.healthleadersmedia.com/
[3] http://www.who.int/en/
[4] http://www.halexandria.org/dward048.htm

analysis, and prevention of adverse effects of medicinal drugs [Härmark and van Grootheest (2008)]. The complete adverse effect profile of a drug is not known at the time of approval because of the small test population size, short duration, and limited generalizability of pre-approval clinical trials [Ahmad (2003)]. As a result, a lot of additional adverse effects are observed after the drug is made available to public over long periods of time. Therefore, the drug manufacturers are ethically committed and legally obliged to accurately monitor any adverse effects and report them to the drug regulatory authorities for better communication about drug usage in the market. Several examples of drug withdrawal exist due to unbalanced risk-benefit ratio of a drug[5]. Adverse effects pose major socioeconomic challenges. For instance, in the year 2008, the National Health System of UK reported an expenditure of £2 billion for treating patients due to adverse drug reactions[6]. In the recent years, beyond the national bodies, international organizations such as the Food and Drug Administration (FDA)[7], the World Health Organization, the European Medicines Agency (EMEA)[8], and the Medicines and Healthcare products Regulatory Agency (MHRA)[9] have maintained postmarketing surveillance systems that enable individuals to spontaneously report the adverse effects experienced as a result of using drugs or healthcare products. The reported adverse effects are carefully monitored by drug regulatory experts in order to ensure drug safety and integrity in the market [Bates et al. (2003)].

1.2 Scientific Problems and Research Motivation

In the medical domain, recent progress in the research and development along with advancement in patient healthcare technologies has resulted in generation of enormous amount of data [Zhu et al. (2003), Doukas et al. (2010)]. Medical data include images from diagnostic procedures (*e.g.* X-ray), textual information described in research articles, or laboratory readouts from patient's experimental samples [Mullins et al. (2006), Mikut et al. (2006)]. Amongst various kinds of medical data generated, free-text denote one important data resource due to their abundant existence, rapid rate of generation, as well as valuable information enclosed. Figure 1.1 shows the amount of indexed citations collected in the bibliographic database MEDLINE[10] during each fiscal year since 1995.

Although an ample amount of medical information are analyzed and stored in heterogeneous electronic databases[11], a substantial amount of information remain unexplored in the form of free-text literature [Krallinger et al. (2005), Harmston et al. (2010)]. Apart from their advantages, databases alone cannot capture the richness

[5] http://www.fda.gov/Safety/Recalls/default.htm
[6] http://www.guardian.co.uk/society/2008/apr/03/nhs.drugsandalcohol
[7] http://www.fda.gov/
[8] http://emea.europa.eu/
[9] http://www.mhra.gov.uk/index.htm
[10] http://www.nlm.nih.gov/bsd/stats/cit_added.html
[11] http://www.meddb.info/

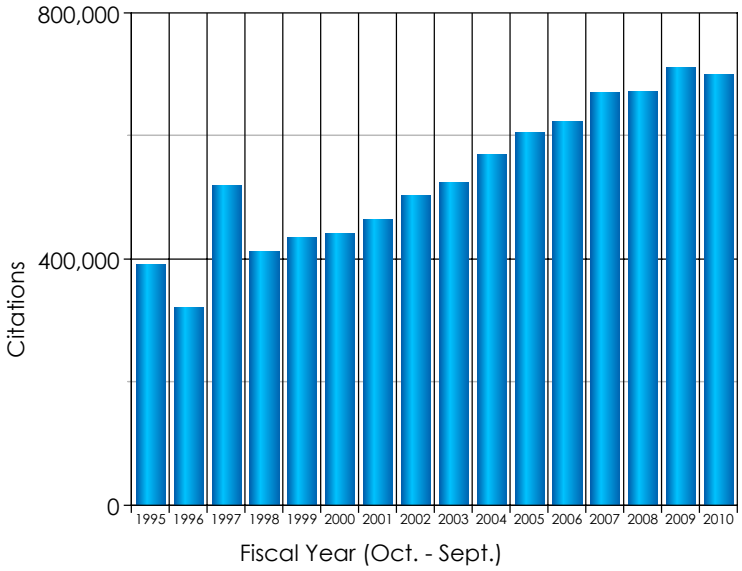

Figure 1.1: Amount of indexed citations added to MEDLINE during each fiscal year since 1995.

of scientific information and argumentation contained in the literature. Complex assumptions, interpretation of novel findings or contradictions, and hypotheses are often expressed using a natural language in free-text. Especially in the medical domain, a major portion of the patient's clinical observations, including radiology reports, operative notes, and discharge summaries are recorded as narrative text (dictated and transcribed, or directly entered into system by care providers) [Demner-Fushman et al. (2009)]. According to Hale (2005) only a small proportion of information is available in structured form manageable by database systems, whereas around 80% is unstructured free-text literature. These include patient health records, electronic medical records, medical case reports, full text research articles, patents, blogs, forums, and news reports [Cohen and Hersh (2005), Kamel Boulos et al. (2010), Van De Belt et al. (2010)]. The study of literature enables the identification of novel facts, hypotheses, new connections between the events occurring at different levels (*i.e.* from microscopic to physiological) and drives the generation of new ideas and clinical decision support.

However, the goal is hard to achieve by reading all the documents since the size of bibliographic space is extremely huge. The enormous growth of literature resources has urged the development of domain specific informatics tools in order to support the analysis of huge amount of unstructured information[12]. Therefore, two fundamental aspects that have gained popularity since over a decade include:

- Information Retrieval (IR): Identification of relevant documents from large collections that satisfy a certain information need.

[12] `http://www.ibm.com/ibm/gio/us/en/clients.mayo.html`

- Information Extraction (IE): Identification of useful facets of information from huge volumes of unstructured document sources.

In the context of medical research, the information retrieval includes identifying the patient records from hospital repositories to find population over which comparative effectiveness studies can be performed [Khoury et al. (2009)]. Similarly, the information extraction includes identifying different classes of medical concepts (*e.g.* disease, treatment, etc.), relationship between the concepts, or events associated with them [Hobbs (2002), Denecke (2008)]. An example of medical information technology is the MedLEE system [Chiang et al. (2010), Sevenster et al. (2011)] that has found various applications in the medical scenario. Collaborative research projects such as EU-ADR[13], and IMI-EHR4CR[14] have fostered the medical research and development by bringing together academia and industries. Competitive assessments such as TRECMED[15] and I2B2[16] have provided ground for open development, evaluation, and benchmarking of medical text mining technologies. In the perspective of healthcare, text mining technologies can avail several benefits such as:[17]

- Provide appropriate access to the key information recorded in free-text such as patient's diagnoses, lab tests preformed, medications prescribed, and their outcome that would facilitate the sound clinical decision making in a timely manner.
- Provide quick access to new and past results (such as patient's response to a therapy) that would increase patient safety and effectiveness of care [O'Connor et al. (2011)].
- Enhance legibility and reduce the redundant experiments or tests performed over patients that can effectively cut down the cost of treatment.
- Generate timely alerts and computerized decision support systems that would improve compliance with best clinical practices and accelerate services to the patients [Haynes et al. (2010)].
- Identify suitable individuals for clinical trails or comparative effectiveness studies [Embi et al. (2005)].
- Facilitate the enrichment of databases and literature-based knowledgebases [Thorn et al. (2005)].
- Perform knowledge discovery and association mining in order to find the association or linkage between different biomedical events [Campillos et al. (2008)].

[13] `http://www.alert-project.org/`, EU-Adverse Drug Reaction
[14] `http://www.ehr4cr.eu/`, Electronic Health Records for Clinical Research
[15] `http://trec.nist.gov/tracks.html`, Text REtrieval Conference Medical Records Track
[16] `https://www.i2b2.org/`, Informatics for Integrating Biology and the Bedside
[17] `http://www.openclinical.org/`

- Likewise, the patent mining can help in updating recent scientific advancements in the market, policy and investment making, portfolio analysis, and so-forth.

1.3 Goals of the Thesis

With an emphasis on huge bibliographic space and growing amount of literature in the medical domain, this work aims at development and validation of automated strategies for efficient information retrieval and information extraction with dedicated focus on patient health records, scientific abstracts, and patents.

A preliminary step during text mining is the recognition of information carries *i.e.* denominations of domain specific named entities or concepts. Continuative information extraction tasks rely on this basic step and essentially depend on their performance. Since over a decade, variety of named entity recognition, and concept identification techniques have been proposed. Similarly, different terminological resources are available that can serve as a backbone for the named entity recognition. In the medical domain, named entity recognition poses an extraordinary challenge since medical terms are written in various forms. For example, medical condition of a patient such as a disease can be mentioned as in accordance to a standard diagnostic code (*e.g.* ICD-9[18]), as a complex description, or as an abbreviation. Although significant amount of work has been done in the biomedical domain for the recognition of entities of biological interest [Yeh et al. (2005)], less efforts have been invested for the recognition of medical disorders such as diseases and adverse effects.

> Goal-1: A systematic development and evaluation of different strategies for recognition of medical disorders in scientific articles. The adapted systems needs to be tested for their generalizability to recognize other classes of medical entities such as medical treatments, and diagnostic tests. Performances of applied systems will be studied by their evaluation over different types of corpora such as scientific articles, and patient health records.

Information retrieval serves as a baseline rationale for selecting the important documents that can be subjected to information extraction either manually or automatically. For a user-defined query, the information retrieval not only focusses on the retrieval of relevant documents (*i.e. documents that best answer the user's question*) but also needs prioritization of highly relevant documents in comparison to less relevant or irrelevant ones. In the medical domain, document retrieval using keywords, manual query expansion, or based on MeSH[19] indexing have been popularly used [Lu et al. (2009), Shetty and Dalal (2011), Trieschnigg et al. (2009)]. This elucidates the need for advanced and sophisticated techniques that can yield improved results in comparison to the existing methodologies.

[18] http://www.cdc.gov/nchs/icd/icd9.htm
[19] http://www.nlm.nih.gov/mesh/

> Goal-2: A systematic development and evaluation of a strategy for information retrieval from biomedical document collections. Performance of the newly implemented retrieval strategy needs to be tested in comparison to the conventional keyword-based retrieval. Generalizability of the adapted retrieval technique will be studied over patient health records and patents.

Extraction of information about adverse effects of drugs can greatly support accelerated pharmacovigilance and patient safety. There are sparsely available general purpose medical information extraction tools that can be apprehensively applied for extraction of adverse events from text. In addition, there are couple of commercial systems that provide facility for adverse event extraction (*e.g.* Luxid[20]). But, there are no popular publicly available technologies that are tailored to support adverse event extraction from text.

> Goal-3: Development of strategies for automatic extraction of information about drug-related adverse effects from the medical literature. The developed technique needs to be qualitatively evaluated as well as carefully studied for its ability to support real world pharmacovigilance studies.

Development of a precise information retrieval or information extraction system requires manually annotated corpora. A manually annotated corpus serves multiple purposes. First, it provides necessary data for development or optimization of automatic systems irrespective of the underlying methodologies. It serves as a gold standard against which the performances of automatic systems can be evaluated and compared. Finally, the annotated corpora can be used as a curated information source for construction of literature-based knowledgebases (such as MetaCore[21]). Unlike the biological domain, the availability of annotated corpora is restricted in the medical domain. This is partly due to the proprietary nature of patient health care systems as well as the safety and legacy issues that health care organizations commit in order to protect the patient privacy data [Sweeney et al. (2005)].

> Goal-4: Aggregation of publicly available medical free-text resources and construction of corpora that are systematically annotated. Annotation and corpus development are manual labor intensive tasks that are often prone to errors. Therefore, strategies to control the quality and minimize errors during annotation needs to be strongly considered. Different sets of corpora will be developed for different tasks such as the named entity recognition (see Chapter 4), and the pharmacovigilance study (see Chapter 10).

[20] `http://www.temis.com/?id=94&selt=16`
[21] `http://www.genego.com/metacore.php`

1.4 Outline of the Thesis

Chapter 2 provides a brief introduction to how medical information is communicated through electronic medical resources. It provides an overview on different terminological resources, knowledgebases, and bibliographic databases in the medical domain.

Chapter 3 gives a general introduction to fundamental aspects of text mining including information retrieval, information extraction, machine learning, evaluation protocols, and state-of-the-art biomedical text mining technologies.

Chapter 4 describes the methodological aspects of identification of medical disorders in scientific articles using various domain specific terminological resources.

Chapter 5 discusses the application of machine learning technique for identification of medical disorders. It also provides a comparative evaluation of different techniques for disorder recognition.

Chapter 6 describes the techniques implemented for identification of medical concepts in patient health records. It also presents an approach developed for the classification of assertions made on medical problems in health records.

Chapters 7, 8, and 9 provide details on adaptation of in-house semantic search platform for retrieval of patient health records and patents. Under both the scenarios, the system is qualitatively evaluated as a part of public assessment.

Chapter 10 describes systematic corpus generation and methodological approach developed for the identification of drug-related adverse events in medical text. Finally, Chapter 11 provides conclusions and outlook onto future perspectives.

Chapter 2

Medical Information Resources

An advancement in research and development in the medical domain as well as patient healthcare technologies has led to the generation of huge amount of data. They can be empirical quantitative observations, short descriptive notes, or full-length documentations (e.g. research articles, patents). Structuring the massive information obtained from these data can ease their accessibility and provide better understanding of the underlying knowledge. Therefore, industries and academia have come forward since past few decades to organize the valuable information by development of databases, thesauri, ontologies, and knowledgebases. This chapter presents few medical information resources as well as the scope of information they cover.

2.1 Terminological Resources

Terminological resources in the medical domain are distributed in the form of thesauri, ontologies, and hierarchies depending on the purpose of development. The scope of information they cover vary across different resources. Broad medical resources such as UMLS[1] or SNOMED[2] cover wide category of aspects such as demography, anatomy, clinical, pharmaceutical, and many more. Whereas, the focussed resources such as MedDRA[3] or ICD[4] cover specific medical sub-domains. The following subsections provide an overview on few popular terminological resources.

2.1.1 Medical Subject Headings

Medical Subject Headings (MeSH)[5] [Sewell (1964)] is a controlled vocabulary thesaurus designed and maintained by the National Library of Medicine (NLM)[6]. MeSH was developed for the purpose of indexing journal articles and books in life sciences. Currently, the NLM uses MeSH for indexing articles in the MEDLINE database. MeSH

[1] `http://www.nlm.nih.gov/research/umls/`, Unified Medical Language System
[2] `http://www.fmrc.org.au/snomed/`, Systematized Nomenclature of Medicine
[3] `http://www.meddramsso.com/`, Medical Dictionary for Regulatory Activities
[4] `http://www.who.int/classifications/icd/en/`, International Classification of Diseases
[5] `http://www.nlm.nih.gov/mesh/`
[6] `http://www.nlm.nih.gov/`

has been extensively used for searching and retrieval of articles in MEDLINE [Lowe and Barnett (1994)].

Currently, the MeSH contains a total of 26,142 subject headings known as descriptors that are hierarchically organized. Most of the descriptors have a short free-text description, links to related descriptors, and a list of synonyms (known as entry terms). The most upper level of the hierarchical structure contains very broad headings such as *Anatomy* or *Mental Disorders*. There are 16 top level descriptors in MeSH. More specific headings (*e.g.* such as *Brain* or *Alzheimer disease*) are found at lower levels of the hierarchy. In addition to descriptors, MeSH contains a small number of qualifiers known as subheadings. Subheading are added to descriptors to narrow down the scope of a topic. For example, *Measles* is a descriptor and *epidemiology* is a qualifier. *Measles/epidemiology* describes the subheading of epidemiological articles about Measles.[7]. In addition to descriptors and qualifiers, MeSH contains over 199,000 Supplementary Concept Records (formerly known as Supplementary Chemical Records) that are present within a separate thesaurus. Although originally developed in English, MeSH has been translated into other languages to support indexing and retrieval of non-English documents [Muench (1971)].

2.1.2 Unified Medical Language System

Unified Medical Language System (UMLS) [Lindberg et al. (1993)] is a metathesaurus designed and maintained by the NLM. UMLS covers numerous thesauri and controlled vocabularies in the biomedical domain. The Metathesaurus denotes a very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health-related concepts and the relationships among them.[8]. Currently, the metathesaurus contains nearly two million concepts and 10 million synonyms of the concepts incorporated from over 100 source vocabularies. UMLS preserves names, meanings, attributes, and relationships of concepts derived from source vocabularies and integrates them into a common semantic framework.

Semantic Network

The semantic network [Kashyap (2003)] consists of a set of broad subject categories (known as semantic types) that provide a consistent categorization of all concepts represented in the UMLS metathesaurus, and a set of useful and important relationships (known as semantic relations) that exist between the semantic types.[9] Major semantic types in UMLS include *organisms*, *anatomical structure*, *biological function*, *chemicals*, *events*, *physical objects*, and *concepts or ideas*. The semantic types are the nodes in the network and the relationship between them are the links. This network kind of representation of the semantic types and concepts aids an easy interpretation of the medical knowledge.

[7] http://www.diabetesdaily.com/wiki/Medical_Subject_Headings
[8] http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html
[9] http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html

2.1.3 Systematized Nomenclature of Medicine

Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [Côté and Robboy (1980)] is a standardized healthcare terminology owned and maintained by International Health Terminology Standards Development Organization (IHTSDO)[10]. SNOMED-CT provides comprehensive coverage of diseases, clinical findings, therapies, procedures, and outcomes.[11] It contains about 357,000 concepts and formal logic-based definitions that are organized into hierarchies and allows consistent way to index, store, retrieve, and aggregate clinical data across various healthcare settings. Some of the computerized applications of SNOMED-CT include electronic health record systems, Computerized Provider Order Entry (CPOE) such as e-Prescribing, catalogues of clinical services; *e.g.* for diagnostic Imaging procedures, knowledge databases used in clinical decision support systems, remote intensive care unit monitoring, laboratory reporting, and many more.[12]

2.1.4 Anatomical Therapeutic Chemical Classification System

The Anatomical Therapeutic Chemical (ATC)[13] classification system is maintained by WHO Collaborating Centre for Drug Statistics Methodology (WHOCC) [Miller and Britt (1995)]. ATC classifies over 3000 drugs from pharmacopeias of different countries. The classification system provides a global standard for categorizing medical substances and serves as a source for drug utilization research. In an application point of view, the ATC classification system has been adopted by several countries as a national standard for medical products. In nordic countries, the ATC is used as a national classification system to identify the marketed medical substances[14].

ATC divides drugs into different groups according to the organ system they act on as well as their therapeutic, pharmacological, and chemical characteristics. In this system, drugs are classified into groups at five different levels. The first level of the ATC indicates the anatomical main group and there are 14 such main groups (*e.g. Cardiovascular system*). The second level indicates the therapeutic group (*e.g. Diuretics*). The third level indicates the therapeutic or pharmacological group (*e.g. High-ceiling Diuretics*). The forth level indicates the pharmacological or chemical group (*e.g. Sulfonamides*). The fifth level indicates the chemical substance. For example, from a top-down view, the drug *Dithranol* is classified under *Dermatologicals*, *Antifungals*, *Antipsoriatic*, *Antipsoriatics for tropical use*, and *Antracen derivatives*.

[10] http://www.ihtsdo.org/
[11] http://www.fmrc.org.au/snomed/
[12] SNOMED CT Technical Implementation Guide, 2009 release
[13] http://www.whocc.no/atc_ddd_index/
[14] http://www.whocc.no/atc_ddd_methodology/history/

2.1.5 Medical Dictionary for Regulatory Activities

The Medical Dictionary for Regulatory Activities (MedDRA)[15] is a clinically validated and standardized medical terminology developed by the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH)[16] to facilitate the sharing of regulatory information internationally for medical products used by humans [Brown et al. (1999)]. MedDRA is managed by the Maintenance and Support Services Organization (MSSO)[17], an organization that reports to International Federation of Pharmaceutical Manufacturers and Associations (IFPMA)[18]. It is used for registration, documentation, and safety monitoring of medical products both before and after a product has been authorized for sale[19]. FDA uses MedDRA as a standard source for coding adverse effects in the Adverse Event Reporting System maintained in the USA.

MedDRA hierarchically organizes medical conditions at four different levels. The first level is called as the System Organ Class (SOC) that contains 26 main groups (*e.g. Endocrine disorders*). The groups at second, third, and forth levels are called as High-Level Group Terms (HLGTs), High-Level Terms (HLTs), and Preferred Terms (PTs) respectively. Each preferred term is associated with one or more Low-Level Terms (LLTs) that are synonyms of the respective preferred term. For example, from a top-down view, the preferred term *Breast cancer* is classified under *Reproductive system and breast disorders*, *Breast disorders*, and *Benign and malignant neoplasms*.

2.1.6 WHO Family of International Classifications

The WHO maintains the Family of Classifications (FIC)[20] on health that provides a consensus and meaningful language framework for communication between governments, providers and consumers. The purpose of WHOFIC to provide a common framework and language to report, compile, use, and compare health information at the national and international level [Jakob et al. (2007)]. The WHOFIC is composed of reference classifications and derived classifications. Reference classifications are those prepared by the WHO and approved by the organization's governing bodies for international use. They include the International Classification of Diseases (ICD), the International Classification of Functioning, Disability and Health (ICF)[21], and the International Classification of Health Interventions (ICHI)[22]. Derived classifications are based on the reference classifications and they include International Classification of Diseases for Oncology 3rd Edition (ICD-O-3)[23], International Classification of Primary

[15] http://www.meddramsso.com/
[16] http://www.ich.org/
[17] http://www.meddramsso.com/
[18] http://www.ifpma.org/
[19] http://www.ich.org/products/meddra.html
[20] http://www.who.int/classifications/en/
[21] http://www.who.int/classifications/icf/en/index.html
[22] http://www.who.int/classifications/ichi/en/index.html
[23] http://www.who.int/classifications/icd/adaptations/oncology/en/index.html

Care (ICPC-2)[24], International Classification for Nursing Practice (ICNP)[25], and many more.

2.2 Knowledgebases

2.2.1 DrugBank

DrugBank is a multipurpose medico-pharmaceutical resource that provides comprehensive information about drugs including their chemical, pharmacological, and medicinal characteristics [Wishart et al. (2008)]. The 2011 version of database contains 6796 drug entries including 1437 FDA-approved small molecule drugs, 134 FDA-approved biotech (protein or peptide) drugs, 83 nutraceuticals and 5174 experimental drugs[26]. DrugBank is unique in terms of the depth of knowledge it covers and levels of integration is provides [Wishart (2007)]. It intends to cover a wide range of knowledge to support research and development at various levels such as academia, industry, and clinic.

DrugBank covers a broad spectrum of drug-related information including their nomenclature (such as *brand names*, *systemic and semi-systemic names*), chemistry (such as *chemical structure*, and *formula*), physico-chemical properties (such as *molecular weight*, and *melting point*), pharmacology (such as *pharmacokinetics*, and *ADME*[27] *properties*), medico-therapeutic properties (such as *disease indications*, *dosages*, and *adverse effects*), and cross-references to related databases (*e.g.* protein database such as *UniProt*[28]). DrugBank is fully searchable and web-enabled resource with built-in tools and features for visualization, searching, and extracting any drug-related information. The database also supports text as well as chemical structure and sub-structure searching with a vision to support drug discovery, pharmacology, and toxicology studies.

2.2.2 PharmaPendium

Pharmapendium is a comprehensive commercial knowledgebase maintained by Elsevier[29] that covers drug-related information. It is the first and only database that enables efficient search and retrieval of FDA and EMA drug approval documents dating back to 1938. Some key features and applications of PharmaPendium are as follows:[30]

- An extensive access to comparative safety, efficacy, and PK data gives potential insights to prioritize the safest and most promising drug candidates for further development.

[24] http://www.who.int/classifications/icd/adaptations/icpc2/en/index.html
[25] http://www.who.int/classifications/icd/adaptations/icnp/en/index.html
[26] http://drugbank.ca/about
[27] Absorption, Distribution, Metabolism, and Elimination
[28] http://www.uniprot.org/
[29] http://www.elsevier.com/
[30] http://www.info.pharmapendium.com/why-use-pharmapendium

- Carefully extracted adverse event and toxicity data from preclinical and clinical studies, as well as a database of post-market event reports allows careful longitudinal investigation of benefit-risk profile of drugs.
- Data search function by text or chemical structures or data pathway browsing by drug, adverse event/toxicity and target can facilitate thorough understanding of pharmacological behavior of drugs and their mechanism of action from physiological to phenotypic levels.

2.2.3 MedicineNet

MedicineNet is a comprehensive medical knowledgebase that aims to bridge the health-related knowledge gap between doctors and the public. It provides in-depth information for consumers through a robust, user-friendly, and interactive website[31]. MedicineNet provides detailed information about various aspects of healthcare including diseases, medical conditions, procedures, tests, and medications. MedicineNet also provides a dictionary called MedTerms[32] that contains easy-to-understand explanations to over 16,000 medical terms.

2.2.4 Side Effect Resource

Side Effect Resource (SIDER)[33] is a public and machine-readable resource for provides information about adverse effects of drugs. Currently, the SIDER contains 62,269 drug-adverse effect pairs and covers a total of 888 drugs and 1450 distinct adverse effects [Kuhn et al. (2010)]. The database incorporates information from drug packet inserts from several public sources, in particular, from the FDA Structured Product Labels (SPL). The standardized Coding Symbols for a Thesaurus of Adverse Reactions Terms (COSTART) is used as the basic lexicon for coding the adverse effects. In order to facilitate the linkage to related databases and reuse for research, the drug names are mapped to the PubChem[34] database. In addition, the SIDER provides the users with facility to explore the package inserts through the concept of *augmented browsing* [Pafilis et al. (2009)].

2.2.5 MedlinePlus

MedlinePlus is a free web-based resource that provides health-related information to patients, and healthcare providers[35]. It incorporates information from the National

[31] http://www.medicinenet.com/
[32] http://www.medterms.com/
[33] http://sideeffects.embl.de/
[34] http://pubchem.ncbi.nlm.nih.gov/
[35] http://www.nlm.nih.gov/medlineplus/

Institute of Health (NIH)[36], several U.S. governmental agencies, and healthcare organizations. Currently, the MedlinePlus contains information about 800 health-related topics (such as medications, therapies, diseases, and diagnostics) in English and Spanish languages. It also provides links to relevant health-related information in 22 other languages. Key features of MedlinePlus include:

- Health information and an encyclopedia covering hundreds of diseases, conditions, and wellness issues.
- Drug-related information including their brand names, dosage forms, and indications.
- Information about herbal medicines and dietary supplements.
- Health news from reputed press releases (*e.g.* Reuters[37]), videos of diagnostic procedures, and tutorials for understanding medical conditions and procedures.

2.3 Bibliographic Resources

2.3.1 MEDLINE

MEDLINE[38] is a premier bibliographic database maintained by the National Library of Medicine. It includes bibliographic information of articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and healthcare. Currently, the database contains more than 19 million records from approximately 5,000 selected journals in the fields of biomedicine and healthcare from 1950 to present. The majority of publications covered in MEDLINE are scholarly journals, whereas a small number of newspapers, magazines, and newsletters considered useful to particular segments of the NLM broad user community are also included[39]. Articles in MEDLINE are indexed using MeSH. This provides a facility for search engines such as PubMed[40] to search over free-text part of articles or MeSH-indexed terms.

2.3.2 TOXLINE

TOXLINE[41] is a bibliographic database of covering toxicological information since 1972 [Schultheisz (1981)]. TOXLINE records provide information covering biochemical, pharmacological, physiological, and toxicological effects of various chemicals including drugs[42]. Currently, the database contains nearly 4 million citations where most of them

[36] http://www.nih.gov/
[37] http://de.reuters.com/
[38] http://www.nlm.nih.gov/bsd/pmresources.html
[39] http://www.nlm.nih.gov/pubs/factsheets/medline.html
[40] http://www.ncbi.nlm.nih.gov/pubmed
[41] http://toxnet.nlm.nih.gov/
[42] http://www.nlm.nih.gov/pubs/factsheets/toxlinfs.html

are abstracts indexed with MeSH terms and CAS registry numbers. A large portion of the database covers standard journal literature in the toxicology domain as well as technical reports and research project reports (such as *Toxicology Document and Data Depository*), and archival collections (such as *Health Aspects of Pesticides Abstract Bulletin (HAPAB)* and *Poisonous Plants Bibliography (PPBIB)*)

2.3.3 PubMed Central

PubMed Central (PMC)[43] is a bibliographic database containing full-text scientific literature covering the biomedical and life science domain. PMC was developed by NLM and is currently managed by NLM's National Centre for Biotechnology Information (NCBI)[44]. The database currently contains nearly 2.2 million articles and serves as an open archive of biomedical journal articles.

2.3.4 ClinicalTrials Database

ClinicalTrials database (popularly known as ClinicalTrials.gov)[45] is a registry of clinical trails. It is maintained by NLM and is the largest database of clinical trials openly available. ClinicalTrials.gov currently contains nearly 130,000 trials from more than 170 countries across the world. The database provides up-to-date information about federally and privately conducted clinical trials for a wide range of diseases and conditions. The primary purpose of the database is to improve public access to drug efficacy studies resulting from approved Investigational New Drug (IND) applications. Every clinical trial record in the database provides comprehensive information including the summary of the clinical study protocol (such as the study design, eligibility criteria, location of the trial, etc) as well as clinical study results (such as the overall outcome, summary of adverse effects, etc).

2.3.5 DailyMed

DailyMed[46] is a database of drug regulatory reports. The database is maintained by NLM and contains information about nearly 28,000 drugs including human and animal drugs. DailyMed aims to provide healthcare providers and the public with comprehensive and up-to-date information about the regulatory reports submitted to the FDA by drug manufacturers. A drug regulatory report is also called as drug label, packet insert, Structured Drug Label (SPL), or Summary of Product Characteristics (SPC). It provides information about the product (such as generic names, ingredients, ingredient strengths, dosage forms, routes of administration, appearance, etc) as well as the packaging (such as package quantities and types).

[43] http://www.ncbi.nlm.nih.gov/pmc/
[44] http://www.ncbi.nlm.nih.gov/
[45] http://clinicaltrials.gov/
[46] http://dailymed.nlm.nih.gov/dailymed/about.cfm

2.3.6 Patent Databases

Patents contain valuable information about intellectual and scientific aspects of inventions. In order to maintain the integrity of the intellectual property, patent documents are made available as images of text documents such as TIFF or PDF provided by their respective patent offices (*e.g.* The United States Patent and Trademark Office (USPTO)[47] publishes patents in TIFF format). The number and importance of patents and patent applications are increasing at a rapid rate worldwide. More than 35 million patent documents have been published so far around the world and the number of inventions since 1968 have been estimated in excess of 8 million. The USPTO patent database includes full-text patents from 1790 - present and also provides TIFF images of most of them. The European Patent Office (EPO)[48] maintains a free database of worldwide patents (including the U.S. patents) called esp@cenet . Here the images of patents are provided in PDF format. National patent offices of various countries (*e.g.* Deutsches Patent und Markenamt (DPMA)[49], and the Japan Patent Office (JPO)[50]) maintain their in-house patent databases.

[47] http://www.uspto.gov/
[48] http://www.epo.org/
[49] http://www.dpma.de/
[50] http://www.jpo.go.jp/

Chapter 3

Foundational Aspects of Biomedical Text Mining

3.1 Fundamentals of Text Processing

With rapidly expanding bibliographic space in the medical domain, there is no surprise in the need for techniques that can identify, extract, and manage important information from this massive amount of data. The primary goal of text mining is to extract the knowledge that is hidden in text and to present it in a concise form to medical professionals or researchers. "Text mining applications integrate a broad spectrum of heterogeneous data resources, providing tools for the analysis, extraction and visualization of information, with the aim of helping biologists to transform available data into usable information and knowledge" [Krallinger et al. (2005)]. Text mining comprises three major activities *i.e.* the information retrieval, to gather relevant text; the information extraction, to identify and extract specific information from the text of interest; and the knowledge discovery, to find associations among pieces of information extracted from various text sources. Sections 3.2-3.6 provide brief introduction to techniques employed in the fields of information retrieval and information extraction.

3.2 Information Retrieval

Information Retrieval is an area of study that deals with searching in large document collections for information within documents or metadata of documents to satisfy certain user needs [Manning et al. (2009)]. The process of information retrieval begins with user's query to the system. The query can be in form of keywords or formal statements. The system automatically interprets the user's request and returns a list of documents that can best answer user's question. PubMed is a popular example of biomedical search engine for searching in MEDLINE abstracts. Following subsections introduce popular aspects of information retrieval including the models for storing, analyzing, and searching the documents.

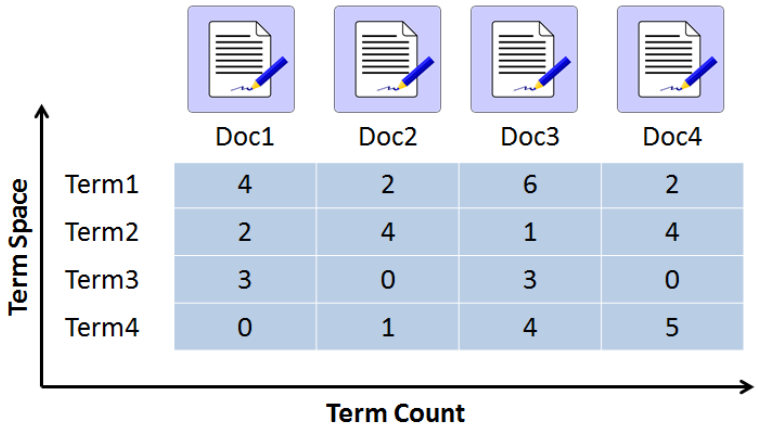

Figure 3.1: Illustration of an example document index. The index contains four documents (*i.e.* Doc1–Doc4) and overall four terms (*i.e.* Term1–Term4). Non-zero numbers indicate the frequency of term occurrence in the respective document.

3.2.1 Vector Space Model

Vector Space Model (VSM) [Salton et al. (1997)] is a model for representation of documents in a multidimensional space. Documents are stored in the form of vectors whose dimensions are determined by terms used to build the index. Figure 3.1 shows an illustration of example document index. Each term occurring in a document forms a separate dimension. If a term occurs in a document, its value for the respective document is set to a non-zero value in the index. There are several ways of weighting the term occurrence in a document. Popular examples include term frequency or term frequency-inverse document frequency (refer Section 3.2.2). Given a query as a term or set of terms, the model enables determination of similarity (refer Section 3.2.2) between the documents and the query, and finally returns a ranked document hit list.

3.2.2 Similarity Scoring

A similarity score is a function that measures the quantitative similarity between a document and user-defined query [Zobel (1998)]. It provides a rationale to rank the documents with respect to the relevance to the query. The quality of retrieval is dependent on the strength and suitability of the scoring function. To date, several scoring functions have been proposed and applied with varying success. They include cosine similarity [Garcia (2006)], term frequency-inverse document frequency similarity [Zobel and Moffat (1998)], and many more.

TF-IDF Similarity

Term frequency-inverse document frequency (TF-IDF) is a term weighting scheme often used for document indexing and retrieval purposes [Spärck Jones (1972)]. It is a statistical measure of how important a term is *w.r.t.* a document collection. In recent years, TF-IDF scheme has been applied by search engines to score the relevancy of documents to a query.

Term frequency (denoted as $tf(t,d)$) counts the number of times a term (t) occurs in a document (d). Inverse document frequency[1] (denoted as $idf(t) = \frac{N}{df(t)}$) measures the importance of a term (t) in a document collection taking into account the size of collection (N) and number of documents containing the term ($df(t)$). Finally, the term frequency and inverse document frequency are combined to produce a composite weight for each term in every document as

$$tf\text{-}idf(t,d) = tf(t,d) \times idf(t). \tag{3.1}$$

Each document can be viewed as a vector of terms weighted by their TF-IDFs. For a given query (q) comprising one or more terms, the similarity score between the document and query terms is calculated as

$$Score(q,d) = \sum_{t \in q} tf\text{-}idf(t,d). \tag{3.2}$$

Although TF-IDF similarity has several advantages such as computing simplicity and document ranking capability, its disadvantages include poor scalability to lengthy documents (such as full-texts) and assumptions of statistical term independence.

Okapi BM25

Okapi BM25 is a scoring function applied by document search engines to measure and rank the documents according to their relevance to a user-defined query [Robertson et al. (1994)]. BM25 is based on a probabilistic framework and is known to overcome several limitations associated with conventional TF-IDF similarity scoring function.

For a given query (q) containing one or more terms, the BM25 score between the query and document (d) is determined as

$$BM25(q,d) = \sum_{t \in q} idf(t) \times \frac{tf(t,d) \times (k_1 + 1)}{tf(t,d) + k_1(1 - b + b \times \frac{dl}{avg(dl)})}. \tag{3.3}$$

In the above equation, $tf(t,d)$ and $idf(d)$ denote term frequency and inverse document frequency respectively, k_1 is a free parameter usually with a value of 2, and $b = 0, .., 1$ is a document length normalization parameter (usually valued 0.75). Assigning b to 0 indicates no document length normalization and assigning it to 1 will carry out full length normalization. Parameter dl indicates the length of document (d) and $avg(dl)$ is

[1] http://nlp.stanford.edu/IR-book/html/htmledition/

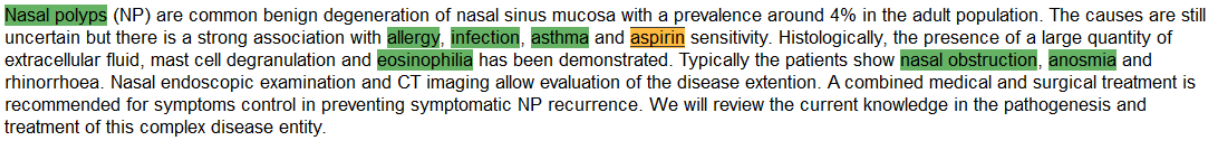

Figure 3.2: Example of named entity recognition performed over a MEDLINE abstract. The recognized entities are highlighted in different colors indicating their semantic classes such as diseases (green) and drugs (yellow).

the average length of documents in a collection. **BM25F**[2] is a variant of BM25 in which documents are considered to be composed of different fields (such as title, abstract, full-text, etc.) with possibly different levels of importance. BM25F function allows searching in one or multiple fields of documents.

3.3 Named Entity Recognition

In the biomedical domain, in order to mine the useful knowledge from literature, the primary requisite is to recognize the named entities such as drugs, diagnostics, and diseases. Named Entity Recognition (NER) refers to the task of recognizing the terms which denote biologically or medically important entities in free-text literature [Hirschman et al. (2005), Cohen and Hersh (2005), Jimeno et al. (2008), Jiang et al. (2011)]. The goal of NER is to relate each named entity of importance in a free-text document to an instance in the real world [Cohen and Hunter (2008)]. In practical text mining applications, the NER is followed by the normalization of recognized entities to biomedical database or ontology entries. Figure 3.2[3] illustrates an example of named entity recognition performed over a MEDLINE abstract.

NER in biomedicine is a non-trivial task despite the availability of many nomenclatures of biomedical entities. Several issues need to be addressed when dealing with biomedical named entities. Some of them are ambiguous names, synonyms, term variations, newly discovered entities not mentioned in curated nomenclatures, and many more. Several approaches have been proposed in the past few decades for efficient NER in biomedical literature. However, the biomedical NER started with recognition of gene/protein names at the beginning and later found applications for recognition of several other biomedical entities such as diseases and drugs. Current NER approaches can be classified as dictionary-based, rule-based, machine learning-based, and their combinations that are briefly introduced in the following subsections.

[2] http://nlp.uned.es/~jperezi/Lucene-BM25/
[3] Adapted from http://www.scaiview.com/

3.3.1 Dictionary-based Approach

Information about several biomedical entities including their nomenclature are well-maintained by leading organizations such as the NLM, WHO, and NCBI. Dictionary-based NER approaches rely on existing domain-specific dictionaries to identify the names of entities in free-text. Dictionaries are derived either manually or automatically from the representative terminological resources. Often string matching or string similarity algorithms are applied as backbone mechanism for identifying the named entities. Dictionary-based approaches provide several benefits such as ability to link database entries to free-text snippets, and normalization of entities at semantic and lexicosyntactic levels. Nevertheless, the performance of dictionary-based approaches are strongly dependent on the comprehensiveness and clarity of information provided by the underlying terminological resource.

There are several public and commercial applications that perform dictionary-based NER in biomedical free-text with varying capabilities and success rates. A publicly available system EbiMed [Rebholz-Schuhmann et al. (2007)] recognizes drug names using a drug dictionary compiled from MedlinePlus. However, they do not provide an evaluation of their system. Applications such as ProMiner [Hanisch et al. (2005), Gurulingappa et al. (2010)] and Peregrine [Hettne et al. (2009)] have demonstrated success in identifying several classes of entities under both biological and medical settings such as genes, drugs, and diseases. Dedicated medical NER system such as MedLEE (Medical Language Extraction and Encoding System) [Friedman et al. (2004)] and c-TAKES (clinical Text Analysis and Knowledge Extraction System) [Savova et al. (2010)] has been successfully applied in clinical settings for identifying and encoding patient-related information such as diagnostics, diseases, and treatments.

3.3.2 Rule-based Approach

Rule-based NER approaches apply manually generated rules for identifying the named entities in free-text [Cohen and Hunter (2004)]. Such systems consist of a set of rules describing term formation patterns using grammatical (e.g. Parts-Of-Speech), syntactic (*e.g.* word precedence), lexical, morphological and orthographic features (*e.g.* capitalization) as well as domain knowledge in combination with dictionaries. They rely on a combination of regular expressions, heuristics, and hand-crafted rules. However, the generation and maintenance of such rules is bound to high costs and domain expertise. Furthermore, rule-based NER systems lack the adaptability to other domains and they are often task and language-specific.

Hamon and Grabar (2010) applied a linguistic rule-based system for analyzing narrative clinical documents to extract medication names and medication-related information. The system attempts to extract medications not covered by the dictionaries. The system was evaluated as a part of open clinical natural language processing assessment and it demonstrated noticeable success. An example of commercial rule-based system is

BioTeKS [Mack et al. (2004)] that is based on UIMA[4] framework. BioTeKS can analyze biomedical text such as MEDLINE abstracts, medical records, and patents; and automatically identify biomedical entities (*e.g.* genes, proteins, compounds, and drugs) and concepts or facts related to them.

3.3.3 Machine Learning-based Approach

Machine Learning (ML)-based NER approaches apply different learning algorithms to train statistical models for performing the task [Zhou et al. (2004)]. The model can be applied for recognition and extraction of entities in various corpora. The strength of a ML-based system depends on the quality and discriminative power of textual features applied as well as the chosen classification algorithm [Krauthammer and Nenadic (2004)]. ML approaches formulate the NER task as a text classification and boundary detection problem. These approaches provide advantages of better adaptability to different domains in comparison to rule-based approaches. However, training an accurate NER model requires manually annotated corpora that can be labor and cost-intensive.

ML-based techniques have been successfully applied in the past for recognition of gene names [Klinger et al. (2007)], and chemical names [Klinger et al. (2008)] including drugs. Hawizy et al. (2011) have developed a open source software (ChemicalTagger) for identifying chemistry-specific names in text. Amongst several ML-based approaches that are available for biomedical NER, Conditional Random Fields (CRF, *see* Section 3.8.6) have been one of the favorite choice and most commonly applied technique that has demonstrated substantial success. Leaman and Gonzalez (2008) and Mahbub Chowdhury and Lavelli (2010) have developed open source re-trainable CRF-based applications named BANNER and BioEnEx respectively. These applications have demonstrated successful examples for biological as well as medical entity recognition (such as diseases).

3.3.4 Hybrid Approach

Hybrid approaches for NER applies combination of techniques derived from dictionary-based, rule-based, or ML-based systems. Several examples exist where hybrid approaches have demonstrated success at various levels [Tsai (2006), Tikk and Solt (2010)].

3.4 Context Disambiguation

A successful identification of named entities in free-text may sometimes require additional processing to communicate the information more precisely. Typical biological examples of such scenarios include disambiguation of gene and protein names [Schuemie et al. (2005)]. Whereas in the medical domain, the disambiguation of medical entities is

[4] `http://uima.apache.org/`, Unstructured Information Management Architecture

gaining popularity since last few years. Context disambiguation in medicine can be in the form of entity disambiguation or assertion classification.

3.4.1 Entity Disambiguation

Entity disambiguation (also popularly known as word sense disambiguation) is a process of identifying and classifying the sense of an occurrence of a named entity in text when it can be associated with multiple meanings (polysemy). For example, *nausea*, a medical condition may occur as a clinical history of a patient or a side effect of drug administration. Several approaches have been taken in the past to solve the entity disambiguation problem.

Stevenson et al. (2011) have proposed an approach that relies on key terms extracted from UMLS and domain information of target document automatically learned from text. They developed and evaluated statistical models that apply information extracted from local context and domain context in order to disambiguate terms in medical documents. Similarly, Savova et al. (2008) developed a machine learning-based approach for disambiguating assertions across two domains, *i.e.* biomedical literature and clinical notes. The system was evaluated in comparison to manually annotated word senses that showed convincing results for both domains. Jimeno-Yepes et al. (2011) presented a method for an automatic development a word sense disambiguation test collection using the UMLS metathesaurus and the manual MeSH indexing of MEDLINE. The dataset is named as MSH WSD. The MSH WSD dataset contains altogether 203 entities and it allows the evaluation of WSD algorithms in the biomedical domain.

3.4.2 Assertion Classification

Named entity recognition in medical documents is often confronted with a requirement to classify the assertion. Assertions are the often made by physicians on medical problems such as symptoms or diseases indicting them as present, absent, or hypothetical. Assertion classification differs from entity disambiguation since it aims at classifying the author's opinion made over the entity.

A popular and classical example of assertion classification tool in medical records is the NegEx program. NegEx was developed by Chapman et al. (2001) and it is a rule-based system that determines negations made over the medical problems in clinical notes. It applies hand-crafted patterns as well as terms extracted from UMLS and identifies negation based on relative distance of occurrence of negation pattern and a medical condition. In 2009, Harkema et al. (2009) extended the features of NegEx to cover temporality and experiencer modules. The program was named as ConText and it classifies assertions made over medical problems as present, absent, or hypothetical. The temporality module of ConText classifies if the medical problem occurs as a patient's history or current event. The experiencer module classifies if the medical problem was observed in the patient or third person (*e.g.* father, mother, son, *etc.*)

3.5 Relationship Extraction

The aim of relationship extraction is to identify and extract semantic relations between different classes of named entities in text [Rink et al. (2011), Bundschus et al. (2008)]. In the biomedical domain, relationship extraction finds several applications including knowledge discovery, hypothesis generation, and question-answering (*e.g. which medical conditions can be treated with paracetamol?*). A classical relationship extraction task requires that the named entities are already known in text. In such cases, the performance of relationship extraction strongly depends on the performance at which named entities are correctly tagged in text.

Relationship extraction has gained immense popularity since past few years. A lot of investigations on relationship extraction are focussed on biological relationships such as protein-protein interactions (PPIs) due to the availability of annotated corpora. A popular example include BioCreative challenge II.5 that focussed on PPI identification [Leitner et al. (2010)]. A recent effort on mining medical relationships was addressed by the I2B2 (Informatics for Integrating Biology and the Bedside)[5] challenge 2010 [Uzuner et al. (2011)]. The relationship extraction task focussed on identifying over 10 different relationships between three classes of medical concepts (*i.e.* problem, treatment, and test). A recent challenge in 2011 focussed on extraction of drug-drug interactions (DDIs) from sentences [Segura-Bedmar et al. (2011)] obtained from free-text fields of the DrugBank database.

Relationship extraction frameworks can be broadly categorized into rule-based and machine learning (ML)-based approaches. Rule-based approaches apply complex linguistic technologies (such as sentence parsing *see Section* 3.6) , hand-crafted domain-specific patterns, or both to capture various types of relationships expressed in text. Generation of such manual rules can be labor and cost-intensive as well as require substantial domain expertise. An example of open source relationship extraction program is the RelEx [Fundel et al. (2007)]. RelEx parses sentences to generate dependency parse tree structures that can be interpreted using simple rules and thereby extract relationships. An investigation was performed by applying RelEx over one million MEDLINE abstracts dealing with gene and protein relationships. The system extracted nearly 150,000 relationships with an estimated performance of 80% overall reliability. Efforts have been invested by various research groups to identify relationships between several classes of biomedical entities using different rule-based approaches [Verspoora et al. (2009), Corney et al. (2004), Morante and Daelemans (2009)]. A recent work on rule-based relationship extraction was performed by Ben Abacha and Zweigenbaum (2011). They presented a platform named MeTAE (Medical Texts Annotation and Exploration) that allows extraction and annotation of entities and relationships in medical text. Their linguistic-rule based approach for extracting relationships between treatments and problems demonstrated competitive results. Commercial applications such as LUXID Skill Cartridge[6] provide facilities for extracting relations between various classes of

[5] https://www.i2b2.org/
[6] http://www.temis.com/

biomedical entities in life sciences and have exhibited success scenarios[7].

As opposed to rule-based approaches, ML-approaches rely on supervised learning algorithms for training and identifying the relationships in text. However, ML-approaches require consistently annotated training data in order to build a reliable classifier. ML-approaches for relationship extraction can be broadly classified into feature-based and kernel-based approaches. Feature-based techniques extract textual features from input text (*e.g.* words occurring between entities) based on which supervised algorithms are trained. Kernel-based methods encode structural representation of text such as the word sequence and a kernel function is designed to capture and differentiate between the meaningful structures [Hofmann et al. (2008)]. Giuliano et al. (2007) have developed an open source kernel-based relationship extraction toolkit called JSRE (Java Simple Relationship Extraction). The JSRE platform allows classification on binary relationships using three different kernels. The system showed competitive results during the DDI-extraction challenge 2011 [Mahbub Chowdhury and Lavelli (2011)]. Yang et al. (2010) have developed a PPI-extraction system BioPPISVMExtractor that is based on Support Vector Machine (SVM)[8]. The system extracts sets of features from text such as surface words, keywords, and distance between the entities to train the system for relation classification. Bundschus et al. (2008) applied CRF-based technique for the identification and classification of relationships between treatments and diseases from PubMed abstracts. They also demonstrated the stability of their approach by relationship extraction between genes and diseases. In the clinical domain, Roberts et al. (2008) applied SVM to detect clinically important relationships (such as has finding, has indication, has location, *etc.*). The system was trained and tested on a corpus of 77 patient narratives which were manually annotated by two clinically trained annotators. The system showed overall high reliability.

3.6 Natural Language Processing Techniques

Text processing requires application of Natural Language Processing (NLP) techniques to transform input textual data into simple structures that can be handled by humans or machines. Such techniques may involve splitting the documents into sentences or words, and so forth. Text processing techniques have been extensively applied for NER, context disambiguation, relationship extraction, etc. Features, rules, or patterns are generated from textual segments of documents processed by NLP techniques that serve as a basis for development of systems for information retrieval or information extraction. Frequently applied text processing techniques are discussed in the following subsections.

[7] https://clara.uib.no/files/2010/09/Geissler.pdf
[8] http://www.support-vector.net/

3.6.1 Sentence Splitting

Sentence splitting is a task of decomposing a document into constitutive sentences. Sentences denote important elements in natural language since they are the smallest units that may express a complete thought or an event. Sentence detection is not a trivial task since the punctuation "." does not always occur at the end of sentences. Biomedical text contains named entities and abbreviations having punctuations as a part of standard nomenclature (*e.g E. coli, W.H.O., etc.*). Correct recognition of sentence boundaries is crucial for several IE tasks.

Several approaches have been proposed for sentence splitting based on various methodologies. Most of them rely on rules and regular expressions for performing the task (*e.g.* GeniaSS[9]). Machine learning-based approaches have been developed and evaluated with considerable success [Tomanek et al. (2007b)].

3.6.2 Tokenization

Tokenization is a process of segmentation of a stream of text into smallest units called tokens such as words, punctuations, and separators. Tokenization can be preformed directly over documents or composite sentences that results in a sequence of tokens. This process often depends on simple heuristics such as separation of tokens on whitespace characters such as spaces or line breaks, and punctuations. An overview on different tokenization techniques is given by Quint (2000) and Jiang and Zhai (2007).

3.6.3 Word Normalization

Word normalization is a process of reducing inflected words to their base forms. Normalization can be performed through stemming or lemmatization. Stemming involves reducing words to their stems. Examples include reduction to words *increasing*, *increased*, or *increases* to *increas*. Programs that perform stemming are called as stemmers. Porter stemmer [Porter (1980)] and Snowball stemmer [Porter (2001)] are popular examples of stemming programs for English language.

Lemmatization is a process of reducing words to their lemmas [Plisson et al. (2004)]. Examples include reduction to words *increasing*, *increased*, or *increases* to *increase*. Lemmatization may sometimes require complex tasks such as understanding the context or parts-of-speech tagging. Lemmatization is closely related to stemming but differs in a way that it assumes the context of appearance of a word in a sentence. MorphAdorner [Burns (2006)] and Dilemma-2 [Facult et al. (1992)] are few examples of lemmatization programs for English language.

[9] http://www-tsujii.is.s.u-tokyo.ac.jp/~y-matsu/geniass/

Word	Base Form	Parts-Of-Speech	Chunk	Named Entity
Antibiotics	Antibiotic	NNS	B-NP	O
treat	treat	VBP	B-VP	O
bacterial	bacterial	JJ	B-NP	O
Antibiotics	Antibiotic	NNS	B-NP	O
infections	infection	NNS	I-NP	O

Table 3.1: Illustration of a sentence tagged by the Genia tagger. Lemmas, parts-of-speech tags, chunk tags, and entity tags are assigned to tokens in the input sentence.

3.6.4 Parts-Of-Speech Tagging

Parts-Of-Speech tagging (POS tagging) is the process of assignment of words in a sentence to the corresponding parts-of-speech based on their context of occurrence. For instance, POS tagging may assign a word to a noun or a verb, and so-forth. POS tagging is not a trivial task since a word can have multiple parts-of-speech depending on the context of occurrence. For example, *antibiotic* appears as a noun in the phrase *antibiotic treats bacterial infections*, whereas it appears as an adjective in the phrase *antibiotic agent*.

POS taggers are generally based on machine learning algorithms such as Hidden Markov Models trained over manually POS-annotated corpora [Marcus et al. (1993)]. Examples exist where rule-based POS-taggers [Brill (1992)] have been developed with considerable success. This has motivated the implementation of specialized taggers optimized for the biomedical domain, such as the MedPost tagger [Smith et al. (2004)], the dTagger [Divita et al. (2006)], and the Genia tagger [Tsuruoka et al. (2005)]. Table 3.1 shows an example of a sentence tagged by the Genia tagger.

3.6.5 Parsing

Parsing involves the application of linguistic knowledge to understand the grammatical structure of a sentence. Parsing is most often performed over sentences rather than directly over documents. A sentence parser typically chunks the sentence into tokens, performs POS tagging, and generates a tree-like data structure with tokens as nodes and directed edges connecting the inter-related tokens. Two nodes are connected if they possess a pre-defined grammatical relation between them (*e.g.* an adjective describing a noun). An example of a sentence parsing outcome is illustrated in Figure 3.3[10].

Several parsers are available for processing the general English and non-English text. Examples of parsers that have been successfully applied in the biomedical domain include the Stanford parser [Klein and Manning (2003)], the McCloskey parser [Mcclosky et al. (2006)], and the Carnegie-Mellon Link Grammar parser [Grinberg et al. (1995)].

[10]Adapted from `http://www.biomedcentral.com/1471-2105/11/101`

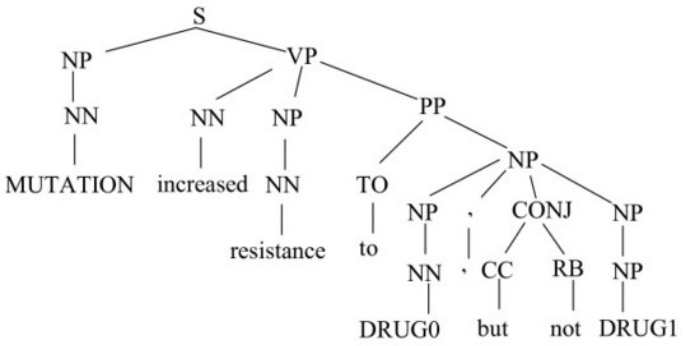

Figure 3.3: Example of a parsed tree structure of a sentence.

3.7 Fundamentals of Machine Learning

Machine Learning evolved from the broad field of Artificial Intelligence, which aims to mimic intelligent abilities of humans by machines [Ratsch (2004)]. The main question addressed in machine learning is how to make the machines able to “learn". Several approaches have been developed for processing the biological and medical data with the help of machine learning techniques. In text mining, an overwhelming amount of textual data and a need for automated procedures to handle this massive data has paved a way for machine leaning to be integrated with classical linguistic techniques. It has been widely used for document classification, information extraction, term classification, knowledge discovery, and so forth.

Machine Learning can be broadly classified into supervised, unsupervised and reinforcement learning. Supervised learning is learning on the labeled data and utilizing this learned knowledge to determine the label for unlabeled/new data. Classical examples of classification and regression fall under the category of supervised learning. Unsupervised learning involves the task of clustering, partitioning or grouping the data under predefined conditions. Reinforcement learning [Kaelbling et al. (1996)] is concerned with how an autonomous system learns or adapts by receiving global feedbacks from an environment. Sections 3.8-3.11 give brief overview on well established algorithms used for machine learning in the field of text categorization.

3.8 Supervised Classification

Classification, also referred to as class prediction, is a process of determining appropriate class labels for unclassified or novel instances. It mainly involves a machine learning technique for learning a function from the training data. The problem of classification and regression has found wide applicability in text categorization stating from term classification up to document classification and email filtering [Larrañaga et al. (2006)]. Several machine learning algorithms have been proposed for performing classification

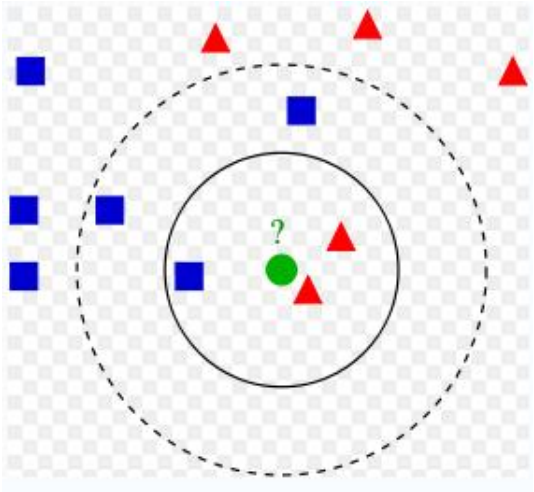

Figure 3.4: Illustration of a nearest neighbor classification.

and each one adopt their own underlying principle of learning. Amongst them are distance-based classifiers (e.g. k-Nearest Neighbor), probabilistic classifiers (e.g. Naive Bayes), decision-based classifiers (e.g. Decision Tree), and margin-based classifiers (e.g. Support Vector Machine). The following subsections provide an introduction to each of the classifiers that have been used within this thesis and the principles behind their classification mechanisms.

3.8.1 *k*-Nearest Neighbor

The Nearest Neighbor (NN) [Cover and Hart (1967)] is a method for classifying objects based on closest training examples in the feature space. An extension of NN is the *k*-NN where a test instance is assigned to the label which is most frequently represented among the *k* nearest training instances [Ratsch (2004)]. Figure 3.4[11] shows an example for simple nearest neighbor classification of an instance.

In the Figure 3.4, an instance in color green indicates a test instance and it has a red instance at nearest distance. Therefore, prediction of class for this test instances according to NN rule is red. A most common way to decide a nearest neighbor of an instance is based on its distance measurement from its neighbors. The Nearest Neighbor method is highly intuitive, simple and produces remarkably low classification errors. The only parameter that controls the performance of the classifier is factor *k* *i.e.* the number of nearest neighbors preferred for classification. If $k = 1$, it represents a simple similarity search criterion.

[11] Adapted from `http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm`

3.8.2 Decision Tree

The Decision Tree classifier, developed by Breiman in 1984 [Breiman et al. (1984)], applies a multistage decision making approach for performing the classification. The basic idea in this multistage approach is to break up a complex problem into a union of several simpler decisions, with the aim to achieve the final solution that resembles the intended solution [Safavian and Landgrebe (1991)]. This algorithm solves the classification problem by repeatedly partitioning the input feature space, so as to build a tree whose nodes represent features and their edges form the decision making function [Yang and Liu (1999)]. Several examples exist for building a decision tree[12]. Once a tree has been modeled based on the examples from the training data, the classification of a test instance is achieved by moving from top to bottom along the branches, starting from the root note until a terminal node is reached. The terminal node assigns a class label to the test instance being classified. Figure 3.5[13] shows an example for a simple Decision Tree where the decision to be made is either *Play* or *Don't Play*, whereas the decision making parameters are *Outlook*, *Humidity* and *Windy*.

Several parameters influence the performance of the Decision Tree classification wherein tree pruning is the most important one [Mingers (1989)]. Pre-pruning involves trying to decide during the tree building process when to stop the subtree development process and post-pruning involves building a complete tree first and then pruning it when necessary. However, most of the Decision Tree builders use the post-pruning strategy. Tree pruning helps to make better decisions by neglecting the unnecessary nodes and it also reduces the computational time and complexity. The Decision tree is a simple yet effective classification scheme when the dataset is small. If the dataset is extremely large, it may result in complicated trees, which in turn require large memory for storage and the situation becomes computationally demanding (this holds true in principle for all learning methods).

3.8.3 Naïve Bayes

The Naïve Bayes classifier is a simple probabilistic classifier based on the Bayes rule. The classifier performs predications based on three important assumptions [John and Langley (1995)]:

- It assumes that the predictive attributes are conditionally independent given the class.
- It posits that no hidden or latent attributes influence the classification process.
- A common but not intrinsic to Naïve Bayes approach is that within each class, the values of numeric attributes are normally distributed.

[12] http://www.cis.temple.edu/~giorgio/cis587/readings/id3-c45.html#1

[13] Adapted from http://gautam.lis.illinois.edu/monkmiddleware/public/analytics/decision_tree.png

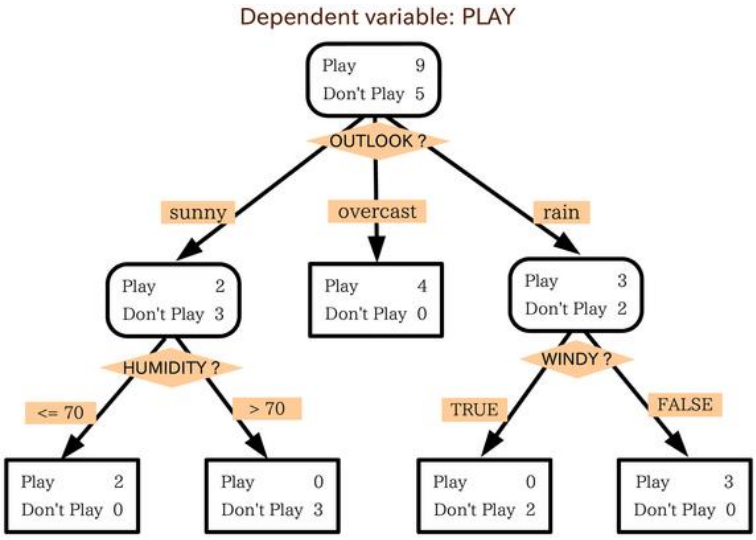

Figure 3.5: Illustration of a Decision Tree classification. Nodes are formed by decision-making features.

Let c denote the class of an instance and let $\vec{x} = (x_1, ..., x_n)$ be a vector of random variables denoting the attribute values of an instance. The Naïve Bayes classifier applies Bayes rule to compute the probability of each class given the vector of observed values and then predicts the most probable class as

$$P(c|\vec{x}) = \frac{P(c) \times P(\vec{x}|c)}{P(\vec{x})}. \tag{3.4}$$

Since the attributes are assumed to be conditionally independent, one arrives at the situation which is simple to compute the class for a test case given the training data and it is represented as

$$P(\vec{x}|c) = \prod_{i=1}^{n} P(x_i|c). \tag{3.5}$$

Naïve Bayes is a simple classifier that needs less or no optimization. McCallum and Nigam (1998) proposed that the use of a kernel density estimation function in order to extrapolate the attribute values into a new higher dimensional space can result in better performance of the classification. Naïve Bayes classifier is pretty easy to implement and robust in solving the classification problem. The Naïve Bayes classifiers won the popularity in recent times for spam mail filtering. In the biomedical text classification domain, the Naïve Bayes classifiers have reported success and have proved to be competitive with other sophisticated classifiers like Support Vector Machine [Huang et al. (2003)].

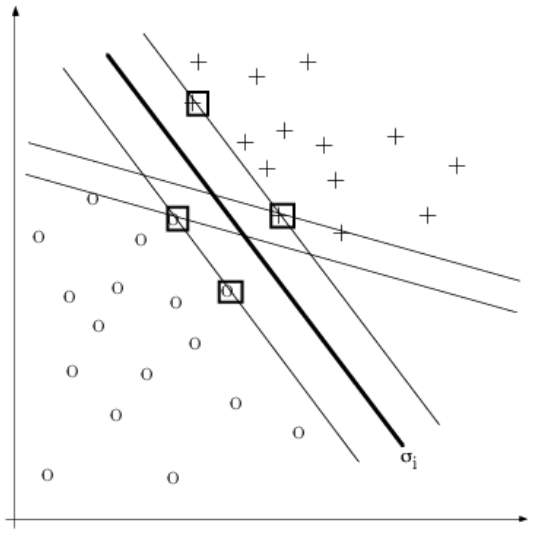

Figure 3.6: Example of SVM-based classification for a binary class problem. The separating hyperplane is indicated by the lines (actual support vectors) that are diagonal in position. The non-diagonal lines indicate second possible solution to separating hyperplane problem where support vectors have lower separation distance when compared to the actual support vectors.

3.8.4 Support Vector Machine

Support Vector Machines (SVM) aim to generate a separating hyperplane that separates the training instances into different groups according to their class labels [Vapnik (1995)]. SVM reply on a data pre-processing strategy wherein the attribute values of labeled instances are projected into a higher dimensional feature space. With an appropriate mapping to a sufficiently high dimension, the data becomes linearly separable by a hyperplane [Joachims (1998)]. The support vectors are training instances that are closest to the hyperplane and they define boundaries for the optimal separating hyperplane. Since the main aim of SVM is to draw a separating hyperplane, larger the margin distance is from support vectors, better is the generalization of the classifier [Larrañaga et al. (2006)]. Figure 3.6[14] shows an example for classification by SVM with a binary class problem where the classes to be separated are 'o' and '+'.

"Given a training set of instance-label pairs (x_i, y_i) where i = 1,...,l, $x_i \in \Re^n$ and $y_i \in \{-1, 1\}^l$, the SVM require the solution of the following optimization problem

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \xi_i, \tag{3.6}$$

and subject to

[14]Adapted from `http://www.cs.technion.ac.il/~pechyony/svm.png`

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i. \tag{3.7}$$

Here the training vector x_i is mapped into a higher dimensional space by a function ϕ . ξ_i is called as a slack variable which measures the degree of misclassification of the datum x_i. The vector w is a normal vector and it is perpendicular to hyperplane indicated as w^T. The parameter b determines the offset of the hyperplane from the origin along the normal vector w. Then SVM finds a liner separating hyperplane with a maximal margin in the higher dimensional space". [Hsu et al. (2010)]

The SVMs were developed in 1992 by Vapnik to initially solve binary classification problems. Now the SVMs have been applied in several areas to solve multi-class problems [Yang and Liu (1999)]. In such cases, the most appropriate way to implement SVM is by classifying one class against the rest of the classes performed for every class individually. SVMs have reported several success stories in the text categorization problem but they are difficult to implement, computationally expensive, and need several optimization steps in order to generate a promising classification.

3.8.5 Maximum Entropy

The Maximum Entropy classifier (also known as the MaxEnt) is based on the principles of multinomial logistic regression [Boehning (1992)]. MaxEnt classifier is used as an alternatives to Naïve Bayes classifier since it does not assume the statistical independence of features. The Naïve Bayes classifier utilizes simple counts of occurrences of features and classes to build a probabilistic predictive model, whereas the MaxEnt applies weights to features upon iterative training that are maximized using maximum-a-posteriori[15] estimation.

"Assuming the presence of dependent variable categories $0, 1, .., J$ with 0 being the reference category, one regression is run for each category $1, 2, .., J$ to predict the probability of y_i belonging to the respective category. Then, the probability of y_i belonging to category 0 is determined by adding-up constraint that the sum of probabilities of y_i belonging to other categories equals one. The regression for $k = 1, 2, .., J$ are performed according to

$$Pr(y_i = k) = \frac{exp(X_i \times \beta_k)}{1 + \sum_{j=1}^{J} exp(X_i \times \beta_k)}, \tag{3.8}$$

and add-up constrains according to

$$Pr(y_i = 0) = \frac{1}{1 + \sum_{j=1}^{J} exp(X_i \times \beta_k)} \tag{3.9}$$

where y_i is the observed outcome for the i^{th} observation on the dependent variable, X_i is a vector of the i^{th} observations of all the explanatory variables, and β_k is a vector

[15] http://www.cs.utah.edu/~suyash/Dissertation_html/node8.html

of all the regression coefficients in the j^{th} regression".[16] The MaxEnt classifier has been successfully applied in the biomedical domain [Yeo and Burge (2004)] including for text classification [Nigam et al. (1999)].

3.8.6 Conditional Random Fields

Linear chain Conditional Random Fields (CRF) are statistical modeling methods applied for sequential data. Klinger and Tomanek (2007) provides a detailed report on the principles of CRF. CRF is a probabilistic model for computing the conditional probability $P(\vec{y}|\vec{x})$ of a possible label sequence $\vec{y}$ given an input sequence $\vec{x}$. In CRF, the conditional probability of a label sequence can depend on arbitrary, non-independent features of the observation sequence, whereas the model does not need to take the distribution of those dependencies into account. In contrast, Maximum Entropy Markov Models (MEMMs) and other Markov models have a theoretical weakness of 'label bias' problem. This makes CRF more powerful modeling technique in comparison to conventional Markov models.

When applied for text modeling, tokens can be described by several features representing their characteristic attributes (*e.g.* string affixes). CRF provides an advantage over other models such as it exploits arbitrary feature sets along with the dependency in the labels of neighboring tokens as indicated by $f_j(y_{i-n}, y_i, \vec{x}, i)$ in Equation 3.10. This results in a feature vector representation of every token in the form

$$f_j(y_{i-n}, y_i, \vec{x}, i) \begin{cases} 1 & if\ y_{i-n} \neq O \text{ and} \\ & \quad y_i \neq O \text{ and} \\ & \quad x_i \text{ has feature } m_i \\ 0, & \end{cases} \tag{3.10}$$

where $i = 1, ..., n$ with $n \in \mathbb{N}$ denotes the label for a token at position i in the sequence $\vec{x}$, and $j = 1, ..., m$ with $m \in \mathbb{N}$ is the number of features.

In general, a linear-chain CRF is an undirected probabilistic graphical model

$$P_{\vec{\lambda}}(\vec{y} \mid \vec{x}) = \frac{1}{Z(\vec{x})} \cdot \prod_{i=1}^{n} \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_j f_j\left(y_{i-1}, y_i, \vec{x}, i\right)\right) \tag{3.11}$$

with the observation-dependent normalization to $[0, 1]$ given by

$$Z(\vec{x}) = \sum_{\vec{y} \in \mathcal{Y}} \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_j f_j\left(y_{i-1}, y_i, \vec{x}, i\right)\right). \tag{3.12}$$

Here, $\mathcal{Y}$ is the set of all possible label sequences over which is summed up, so that a feasible probability is obtained. The weighting factors $\lambda_j \in \mathbb{R}$ are model parameters and define the contribution of single features to the entire model. The goal of model training is to estimate λ_j of the weight vector $\vec{\lambda}$ so that the probability of the output

[16] Section adapted from `http://en.wikipedia.org/wiki/Maximum_entropy_classifier`

label sequence given the training data is maximized. The following likelihood function is maximized where τ represents the training data.

$$\bar{\text{Ł}}(\tau) = \sum_{(\vec{x},\vec{y})\in\tau} \log P_{\vec{\lambda}}(\vec{y}|\vec{x}) \tag{3.13}$$

Maximum-a-posteriori training is performed efficiently using hill-climbing methods such as conjugate gradient or limited memory BFGS (L-BFGS) [Sha and Pereira (2003)]. If a model is given, the task is the determination of the most likely sequence of states $\vec{y}$ for a given observation sequence $\vec{x}$. This means identifying the label sequence that maximizes the joint probability. The most likely sequence is calculated using Viterbi's algorithm [Forney (1973)], a dynamic programming method.

3.9 Active Learning

Active learning is a strategy applied in the machine learning domain to interactively assemble training data. Active learning can help to overcome the limitations associated with human annotation efforts by careful training data selection that can yield a high performing system. It is based on an assumption that not all elements are equally informative and easy to label. An informative instance is one whose contribution to the system leads to significant improvement in its performance. According to Engelson and Dagan (1996), active learning can avoid redundant annotation of non-informative elements that do not contribute to the performance of system.

Active learning is an iterative process composed to three main phases namely training, selective sampling, and human annotation. A learning algorithm examines many unlabeled elements and selects only those for labeling that are most informative for the learner at during each stage of training. The stopping criterion can either be the number of iterations performed or a desired performance measure of the system. The two most popular active learning methods used in NLP are uncertainty-based sampling [Cohn et al. (1994)] and query by committee [Freund et al. (1997)]. In uncertainty-based learning, new instances are selected for annotation based on the system's classification uncertainty. The assumption is that instances which are harder to classify are more useful for training. In case of probabilistic models, the uncertainty of a classifier is commonly estimated using the entropy. For non-probabilistic models, the classification margin is used, as in the case of support vector machines. In query by committee, a body of classifiers is trained on small training data and subsequently applied over pre-selected unlabeled instances. Instances for which the classifiers yield the highest disagreement are considered to be the most informative. The strategy of active learning has been successfully applied to build consistent models for information extraction [Tomanek et al. (2007a)].

3.10 Performance Evaluation Techniques

Performance evaluation provides a platform for systemic assessment of the quality of statistical models. It helps in understanding the generalizability of models over large collections of unseen instances. Metrics such as precision, recall, F_1 score, and accuracy (*see* Section 3.11) are often used for measuring the system's performance. Nevertheless, as a preliminary measure or in the absence of an independent test data, strategies such as cross-validation or bootstrapping are applied for a systematic decomposition and re-utilization of sample data for the assessment of performance of the system.

3.10.1 Cross-Validation

Cross-Validation [Kohavi (1995)] provides means for assessing the performance of statistical models. It is often applied to estimate the performance of a system to solve the class prediction problem. One round of cross-validation involves decomposition of sample data into complementary subsets where few subsets (known as training data) are applied for training a model and the remaining subsets (known as validation data) serve the purpose of performance validation. To reduce variability, multiple rounds of cross-validation are performed using different subset partitions, and the results are averaged over different rounds performed.

k-Fold Cross-Validation

During k-fold cross-validation, the original data is partitioned into k equally sized subsets. Out of k subsets, a single subset is retained as a validation set for testing the model's performance, whereas the remaining $k-1$ subsets are merged to form one large training set. The cross-validation is repeated k times with each of the k subsets used exactly once for validation. The results obtained from k rounds of validation are averaged to generate a single estimate of the system's performance. If the value of k is same as the size of complete dataset, the validation step is called as the leave-one-out cross-validation.

3.10.2 Bootstrapping

Bootstrapping uses the principle of sampling with replacement on the original dataset to partition it into training set and a validation set [Efron (1979)]. The number of drawing equals to number of data points in the dataset, where the drawn samples build the training set and the remaining sample form a validation set. The probability of a variable to be drawn k-times is defined as

$$P(k|n,p) = {}^{n}C_k \times p^k \times (1-p)^{n-k}. \tag{3.14}$$

In Equation 3.14, n is the number of data points and p the probability of a single element to be drawn. The probability to be sampled is the same for all data points *i.e.* $p = \frac{1}{n}$.

With increasing number of data points, the probability of a variable to be not included in a training set approximates e^{-1}. The derivation is described as

$$P(0|n,p) = {}^{n}C_0 \times \left(\frac{1}{n}\right)^0 \times \left(1-\frac{1}{n}\right)^{n-0} \tag{3.15}$$

$$P(0|n,p) \lim_{n\to\infty} = \left(1-\frac{1}{n}\right)^{n} = e^{-1} \approx 0.3678. \tag{3.16}$$

This means that the training data contains about 64% of all the elements of the original data. The sampling is performed t times leading to t training and t validation sets. This allows estimating the performance of the system over t sets leading to a robust estimation. Advantages of bootstrapping is that during each round, the size of training data is same as the size of original dataset (containing duplicates) representing about 65% of the total dataset. Whereas, the remaining about 35% forms the validation set for testing. In this way, it is possible to perform several rounds of bootstrapping and generate the system's performance estimate.

3.11 Performance Evaluation Metrics

Performance evaluation metrics provide means to access the quality of any IR or IE system. It provides a rationale for comparing the outcomes of different systems against one another. Furthermore, an evaluation measures if any changes made to the system (such as the parameter optimization) leads to an improvement of the system's performance. Performance evaluation requires an availability of gold standard that is often based on manual judgment against which the quality of system's output is compared. For example, an evaluation of NER system would be performed against manually annotated named entities in text and similarly an IR system would be compared against the manual judgement of documents. For an evaluation of both IR and IE systems, several measures have been proposed and applied depending upon the user-community needs. Popular examples of evaluation measures include the F_1 score or the accuracy, and so forth.

3.11.1 F_1 score

F_1 score (also referred to as F-score) is one the widely applied evaluation measure for IR as well as IE. It measures the overall completeness and correctness of a system. The F_1 score is calculated by comparing the system's output against manual judgements. Elements (such as documents or named entities) that are correctly identified by the system in comparison to the gold standard are 'true positives'. Elements that are identified by the system but not present within the gold standard are 'false positives'. Whereas, the elements present within the gold standard that are not identified by the system are 'false negatives'. Table 3.2 provides an overview on the basic truth measures.

		Gold Standard	
		Positive	Negative
System Output	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

Table 3.2: Overview on the basic truth measures of information retrieval or information extraction systems.

The basic truth measures are used to determine the system's precision and recall which are systematically combined to generate a final F_1 score. Precision measures the correctness of a system by measuring the proportion of correct outputs amongst all the outputs generated by the system. Precision is calculated as

$$Precision = \frac{TP}{TP + FP}. \tag{3.17}$$

Recall measures the completeness of a system by measuring the proportion of correct outputs made by the system in comparison to ground truths within the gold standard. Recall is calculated as

$$Recall = \frac{TP}{TP + FN}. \tag{3.18}$$

The F_1 score is a harmonic mean of the precision and recall, and is calculated as

$$F_1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{3.19}$$

Precision, Recall, and F_1 score lies between 0 and 1 where 1 indicates the best and 0 indicates the worst.

3.11.2 Accuracy

The accuracy measures the proportion of correct outputs in comparison to the total number of cases evaluated by the system. It measures the fraction of correct answers, i.e. true positives and true negatives, with respect to the total number of cases tested. Accuracy is calculated as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.20}$$

In the machine learning domain, the accuracy measure is sometimes used as favorite choice to evaluate the system's predictions [Kotsiantis (2007)]. Nevertheless, considering the shortcomings of this measure, F_1 score has been applied as a standard alternative.

3.11.3 Mean Average Precision

Mean Average Precision score (MAP score) [Voorhees (2000)] is a performance measure used for the evaluation of IR systems. IR systems often output a ranked list of documents for a user-defined query. In such cases, it is desirable to measure the truth values of returned documents considering the order in which they are presented. For a given query (q), the average precision ($AveP$) is computed as

$$AveP = \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{N} \tag{3.21}$$

where k is the rank in the list of retrieved documents, n is the number of retrieved documents, N is the total number of relevant documents, $P(k)$ is the precision at cut-off k in the list, and $rel(k)$ is a binary variable that assumes value 1 if the retrieved document at rank k is relevant.

MAP for a set of queries (Q) is measured by mean of the average precision scores for each query (q). MAP score is determined as

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}. \tag{3.22}$$

The score ranges between 0 and 1 where 1 indicates the best score whereas 0 indicates the worst.

3.11.4 Binary Preference Score

Binary preference (bpref) score [Buckley and Voorhees (2004)] is another performance measure applied for the evaluation of IR systems. Conventional measure such as precision, recall, and MAP scores are based on an assumption that all the relevant documents within a test collection are known and present in the collection which is not always true in real world scenarios. Therefore, the function bpref score measures an enrichment of relevant documents over irrelevant documents present within a retrieved set of documents. For a query (q) having N relevant documents within the collection, bpref score is calculated as

$$bpref = \frac{1}{N} \sum_{r} 1 - \frac{|n \text{ ranked higher than } r|}{N} \tag{3.23}$$

where r is a relevant document and n is a member of the first N judged irrelevant documents as retrieved by the system.

3.11.5 R-Precision

R-Precision measures the precision of retrieval at R^{th} position in the ranked result set for a query that has R relevant documents. This measure is highly correlated to the Average Precision (see Equation 3.21). In principle, the precision is equal to recall at the

R^{th} position. Although there are several performance evaluation measures available, every measure has its own advantages and shortcomings. Therefore, open assessments such as the TREC, or I2B2 provide multiple evaluation results based on different scores. This provides a common showcase for the comparison of performances of different systems under different test conditions.

3.12 Text Mining Technologies and Scenarios in Biomedicine

With an overwhelming need for sophisticated search engines and tools for processing the biomedical literature, there have been several public and commercial tools developed for addressing this issue. Different tools have their own advantages and have been developed to facilitate the dedicated user needs. Section 3.13-3.15 provide an overview on different information retrieval and information extraction tools available to support biomedical and healthcare information needs. Not all the introduced tools have been used within this thesis considering their commercial nature or suitability, but the author focusses on shedding light on technologies that can be applied for the biomedical text mining application.

3.13 Information Retrieval Technologies

3.13.1 SCAIVIEW

SCAIVIEW [Hofmann-Apitius et al. (2008), Friedrich et al. (2008)] is an advanced semantic search and knowledge discovery environment developed by Fraunhofer SCAI. It was initially developed within the framework of @neurIST project[17] (Integrated Biomedical Informatics for the Management of Cerebral Aneurysms). SCAIVIEW provides functionalities to search using various biomedical terminologies within MEDLINE abstracts, full-text, and patents. SCAIVIEW allows retrieval of relevant documents as defined according to user-needs (by queries) as well as the retrieval of statistically associated entities. Named entities occurring in the retrieved documents are colorfully highlighted in order to aid easy document lookup and quick-tracking of interesting information. Figure 3.7[18] shows an illustration of the graphical user interface of SCAIVIEW. An example search scenario in the SCAIVIEW is *find all the diseases ($d_1, d_2, .., d_n$) related to a drug (D_1) and find all the documents where d_2 and D_1 are closely associated*.

The technical functionalities of SCAIVIEW are based on the Apache Lucene[19] system. It allows robust indexing of several gigabytes of documents and meta information in a reasonable amount of time. SCAIVIEW is implemented as easily scalable system for processing various corpora including abstracts, full-text, and patents. SCAIVIEW has

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

[18] Adapted from http://www.scaiview.com/scaiview/

[19] http://lucene.apache.org/java/docs/index.html

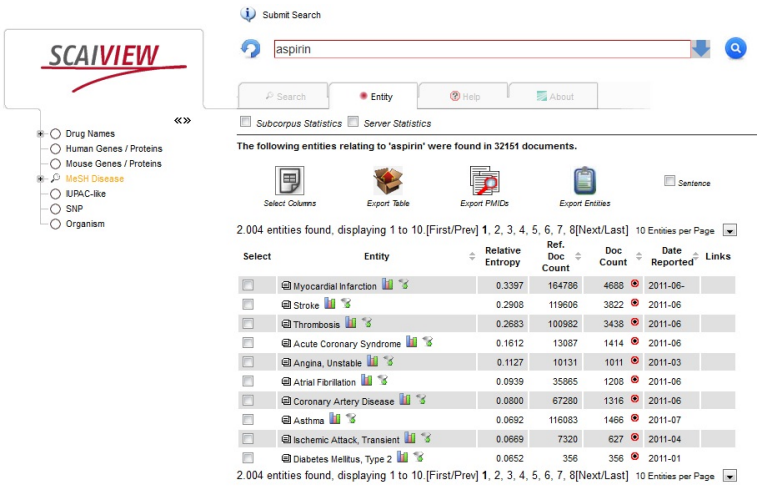

Figure 3.7: Illustration of the user interface of the MEDLINE version of SCAIVIEW.

demonstrated successful scenarios for prior art search in patents [Gurulingappa et al. (2009)], and the retrieval of medical health records.

3.13.2 FACTA

FACTA (Finding Associated Concepts with Text Analysis) [Tsuruoka et al. (2008)] is an open-source search engine developed by NaCTeM[20]. FACTA allows users to retrieve and browse various biomedical concepts (*e.g.* proteins, diseases, chemicals, etc.) appearing in MEDLINE articles in accordance to the user-defined query. The retrieved concepts are ranked according to the co-occurrence statistics and therefore allows users to determine the associations between concepts and the query. FACTA pre-indexes text articles as well as the concepts appearing in them and allows various search strategies such as keywords, concepts, and their boolean combinations. The system allows visualization of snippets within articles to aid easy look-up and evidence tracking. Figure 3.8[21] shows an illustration of the user interface of FACTA search engine. Although the system possesses various text analytic functionalities, it lacks some features such as the search-results export, application programming interface (API), and searching in full-text or patents.

3.13.3 MedSearch

MedSearch [Luo et al. (2008)] is a specialized medical web search engine. The system allows users to search using medical keywords as well as long English descriptions of the information needs. This facilitates several internet users who have limited medical

[20]The National Centre for Text Mining, `www.nactem.ac.uk/`
[21]Adapted from `http://text0.mib.man.ac.uk/software/facta/main.html`

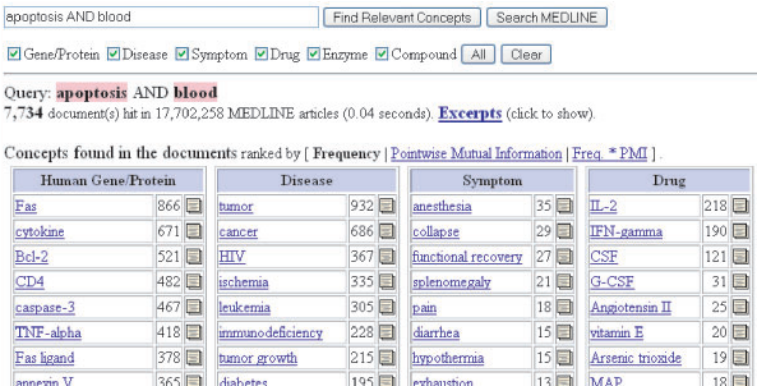

Figure 3.8: Illustration of the user interface of FACTA search engine.

knowledge. MedSearch transforms long descriptive queries into a set of important keywords. An investigation performed over such a query processing indicated improvements in an overall processing time as well as the search results. The system returns diversified web search results that allows users to see various aspects related to their information requirement. Finally, the system also suggests query-related medical phrases that are obtained from standard terminologies such as the MeSH. This helps users to refine the query or easily understand the search results.

3.13.4 Curbside.MD

Curbside.MD[22] is a medical search engine that allows searching in professional resources such as MEDLINE, National Guidelines Clearinghouse[23], FDA, and many more. It allows searches using medical concepts as well as descriptive clinical queries. Key functionalities of the system include differentiation of highly relevant hits across various evidence-based content sources, searching in the image captions of peer-reviewed articles and visualization of informative images, and finally the specialized search facilities within ClinicalTrials database as well as the articles from ACP Journal Club[24], Cochrane Collaboration[25], and many more.

3.13.5 MedicoPort

MedicoPort [Can and Baykal (2007)] is a medical web search engine designed for users with no medical expertise. The backbone of MedicoPort is enriched with medical concepts from the UMLS metathesaurus. The system has been designed to carefully capture the medical semantics in webpages as well as the user-defined queries. Therefore, the

[22] http://www.curbside.md/
[23] http://www.guideline.gov/
[24] http://acpjc.acponline.org/
[25] http://www.cochrane.org/

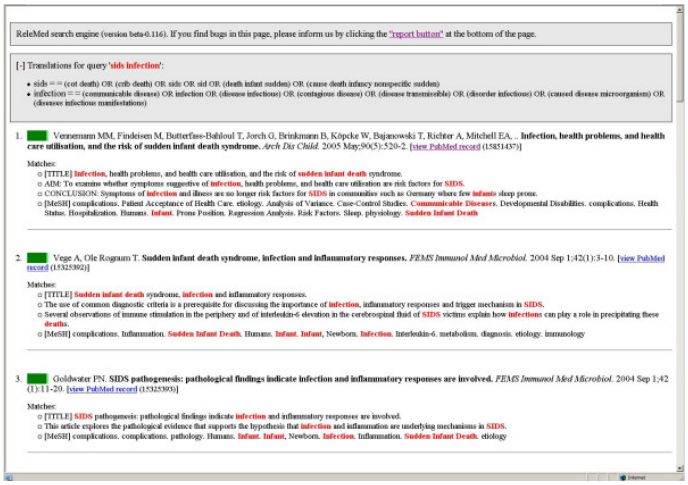

Figure 3.9: Illustration of the search results obtained by Relemed.

system is targeted to help user-groups with minimum or no medical knowledge who seek information about health on web. The system has an ability to mimic the medical expert's domain knowledge by transforming general user queries into domain specific medical concepts that can increase the relevance of query and the retrieved document hits. An experimental investigation made on the performance of system showed the retrieval of relevant document sets that can satisfy the user's request.

3.13.6 Relemed

Relemed[26] [Siadaty et al. (2007)] is a biomedical search engine for searching in MEDLINE. Relemed aims to generate high precision document sets by searching for user-defined query terms within sentences or adjacent sentences. The principle behind Relemed is that when users pose multi-term queries to a system, it is necessary that all terms appear in the articles as well as they are closely associated with one another. Therefore, Relemed applies the sentence-level criteria to judge the relevancy of documents to user-defined queries. Experimental results showed that the system can deliver highly relevant articles at the top of result sets and can outperform the performance of conventional PubMed search in terms of specificity. Additional observations also showed that Relemed can fetch relevant articles that are not retrieved by PubMed search due to 'automatic term-to-concept mapping' to the UMLS implemented within the system. Figure 3.9[27] shows an illustration of the search results obtained by Relemed.

[26] http://bmlsearch.com/
[27] Adapted from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1780044/figure/F1/

3.13.7 EMERSE

EMERSE (Electronic MEdical Record Search Engine) Hanauer (2006) is a medical search engine for searching in free-text fields of electronic medical record. It was developed in order to address the needs for searching in medical records for research and data abstraction. The system allows using complex boolean queries over a easy-to-use interface and the search results are context sensitive. Results are displayed consistently in accordance to the structure of medical records with separate categories for the problem summary list, patient notes, and pathology reports. The system offers automatic spelling correction and robust search across multiple patient records at once.

3.14 Information Extraction Technologies

3.14.1 ProMiner

ProMiner is a tool that can be used to identify potential named entity occurrences in biomedical text and associate database identifiers to the detected terms [Hanisch et al. (2005)]. The ProMiner system uses a pre-processed synonym dictionary, and it follows a combined dictionary based and rule based approaches for biomedical named entity recognition. Its search algorithm is powerful enough to recognize multi-word terms, synonyms, and their variants in text. During the BioCreative open-assessment challenges in 2004 and 2006, the ProMiner demonstrated competitive results for the identification of gene and protein names. Several examples exist where the system has been successfully applied on patents, medical reports, and various other forms of free-text literature. ProMiner can address the following problems:

- Recognition of biological, medical, and chemical named entities in scientific text and their spelling variants depending on the dictionaries used.
- Ability to work with voluminous dictionaries derived from large controlled vocabularies, thesauri and databases.
- Context-dependent disambiguation of biomedical entities and resolution of acronyms[28].
- Mapping of found entities to reference names in the respective data sources.

3.14.2 MedLEE

MedLEE (Medical Language Extraction and Encoding system) is a medical NLP system to extract, structure, and encode clinical information in free-text patient reports so

[28] ProMiner system allows flagging of tokens within a separate list named *Questions*. If a token appearing in *Questions* list is found in a document it is considered as an entity only if atleast one of its synonym co-occurs in the same document. Another list named *AsIsTok* contains tokens that have to be found as-is. These features are helpful for case-sensitive matching and acronym disambiguation.

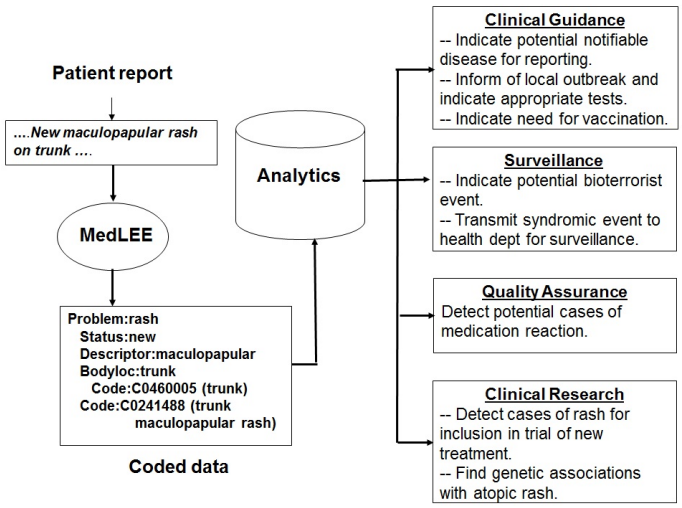

Figure 3.10: Illustration of applications of the MedLEE system.

that the information can be used for subsequent automated processing [Friedman et al. (1996)]. Initially, the MedLEE was developed by Department of Biomedical Informatics at Columbia University but later on commercialized by the NLP International Corporation[29]. The system has an ability to process various kinds of medical records such as radiology reports, cardiology reports, discharge summaries, progress notes, and many more. A qualitative evaluation of the performance of the system indicated superior results in comparison to the physician's ability to handle the medical records. The system has demonstrated various applications in the past including biosurveillance, syndromic surveillance, adverse drug event detection, clinical decision support, clinical research, quality assurance, automated encoding, patient management, and data mining *i.e.* finding trends and associations [Chiang et al. (2010)]. Figure 3.10[30] shows an illustration of applications of the MedLEE system.

3.14.3 MedEx

MedEx is a system for the extraction of medication information from free-text medical records [Xu et al. (2010)]. It can identify medication names and signatures such as the dose, route of administration, frequency, and duration. The medication extraction approach works similar to MedLEE but with finer granularity of the semantics of extracted information. The workflow includes first pre-processing the records to generate the sentences. The sentences are subjected to semantic tagging where dictionary-lookup and RegEx-based methods are applied for the identification of drug names using the RxNorm[31] lexicon. A rule-based parser links the drug names to their respective

[29] http://www.nlpapplications.com/
[30] Adapted from cdc.confex.com/cdc/phin2008/recordingredirect.cgi/id/4165
[31] http://www.nlm.nih.gov/research/umls/rxnorm/

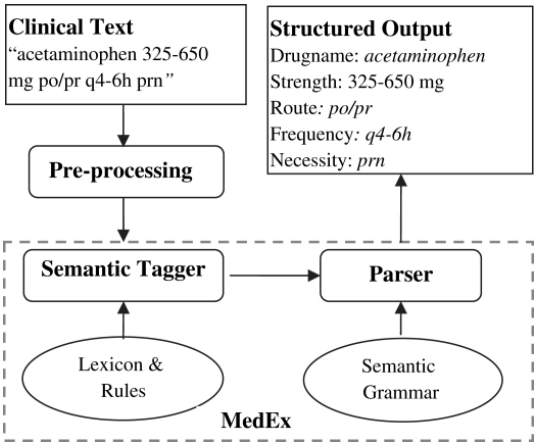

Figure 3.11: An overview of the MedEx system.

signatures such as dose, frequency, etc. The system lacks ability to link the drug names and signatures not occurring within a single sentence. An evaluation of the system indicated overall F_1 score of nearly 92% in comparison of expert annotation for both drug names and signatures. Figure 3.11[32] illustrates an overview of the MedEx system.

3.14.4 MERKI

MERKI[33] (Medication Extraction and Reconciliation Knowledge Instrument) is an open source medical text parser for the extraction of medication information [Gold et al. (2008)]. The system recognizes drug names and related information such as the dose, frequency, strength, and duration. The principle behind working of MERKI parser is similar to that of MedEx. The program relies on parsing rules written as a set of regular expressions and an user-configurable lexicon. It has been developed on discharge summaries from hospitals and an evaluation showed an overall F_1 score of nearly 87%. An evaluation as a part of I2B2 medical extraction challenge showed highly competitive results (*i.e.* ranked fifth out of several participating systems).

3.14.5 cTAKES

cTAKES[34] (Clinical Text Analysis and Knowledge Extraction System) [Savova et al. (2010)] is an open source NLP platform for information extraction from electronic health records. It was developed by Mayo Clinic[35] as a part of OHNLP (Open Health Natural Language Processing)[36] consortium. cTAKES builds on existing open source

[32] Adapted from `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995636/figure/fig1/`
[33] `http://projects.dbmi.columbia.edu/merki/`
[34] `https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/CTAKES_1.2`
[35] `http://www.mayoclinic.com/`
[36] `https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/OHNLP`

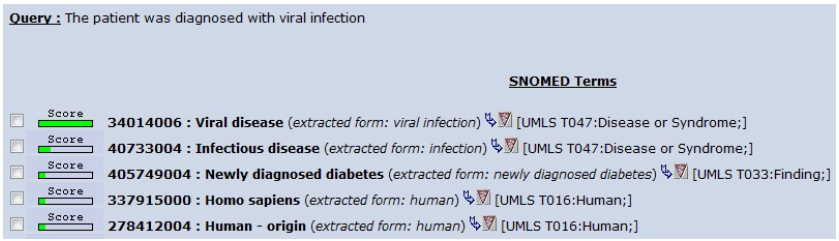

Figure 3.12: Illustration of the output of SNOcat for a user-defined query.

technologies such as the UIMA (Unstructured Information Management Architecture) and OpenNLP toolkit[37]. Its components are specifically trained to handle medical text and create rich linguistic and semantic annotations. The components of the system include sentence splitter, tokenizer, parts-of-speech tagger, shallow parser, and named entity recognizer. The named entity recognizer is based on dictionary-lookup from UMLS and in addition can handle negations and status of named entities. The negation recognizer implements the NegEx algorithm for finding the negations made over the named entities. Similarly, the status annotator uses a regular expression-based approach to determine weather the named entity occurs as a history, current, or a family event. An evaluation of individual components of cTAKES showed convincing results. The upcoming improvements to the cTAKES architecture include co-reference annotation, temporal relationship discovery, and certainty assertion detection.

3.14.6 SNOcat

SNOcat[38] (SNOMED Categorizer) [Ruch et al. (2008)] is an open-assess tool for the identification of SNOMED-CT concepts in biomedical free-text. The system allows online submission of a textual record (such as an abstract, full-text, or medical report) and it returns a ranked list of possible matches to the SNOMED terminology. The system combines pattern-matching based on regular expressions of terms, vector-space indexing and retrieval engine, and *tf-idf* weighting schema with a cosine normalization. An evaluation of the top retrieved concepts showed nearly 80% precision indicating sufficient means for consistently recognizing the SNOMED concepts in free-text. The system has an ability to consistently retrieve documents through the SNOMED indexing that can be better in comparison to the conventional MeSH-based retrieval. Figure 3.12[39] shows an example of the output of SNOcat for a user-defined query.

[37] http://incubator.apache.org/opennlp/
[38] http://eagl.unige.ch/SNOCat/
[39] Adapted from http://eagl.unige.ch/SNOCat/

```
Phrase: "sodium channel inhibitor"
>>>>> Phrase
sodium channel inhibitor
<<<<< Phrase
>>>>> Candidates
Meta Candidates (10):
   981 C0872271:Sodium Channel Inhibitors (Sodium Channel Blockers) [Pharmacologic Substance]
   827 C1999216:Inhibitor [Qualitative Concept]
   793 C0243077:inhibitors [Chemical Viewed Functionally]
   755 C0021469:Inhibitory (Metabolic Inhibition) [Molecular Function]
   734 C0037492:Sodium Channel [Amino Acid, Peptide, or Protein,Biologically Active Substance]
   660 C0037473:Sodium [Biologically Active Substance,Element, Ion, or Isotope]
   660 C0037570:Sodium (Sodium, Dietary) [Food,Inorganic Chemical]
   660 C0439799:Channel [Spatial Concept]
   660 C0597484:Sodium+ (Sodium Cation) [Element, Ion, or Isotope]
   660 C1706095:Channel (Channel Object) [Conceptual Entity]
<<<<< Candidates
>>>>> Mappings
Meta Mapping (981):
   981 C0872271:Sodium Channel Inhibitors (Sodium Channel Blockers) [Pharmacologic Substance]
<<<<< Mappings
```

Figure 3.13: Example of result of mapping performed by the MetaMap program.

3.14.7 Textractor

Textractor[40] [Meystre et al. (2010)] is a medical NLP system for the extraction of medication information from free-text medical documents. The system can extract drug names, including their dosage, frequency, and the reasons for their prescription. Textractor is based on the Unstructured Information Management Architecture (UIMA) framework, and uses methods that are a hybrid between machine learning and pattern matching. Two modules in the system are based on machine learning algorithms, while other modules use regular expressions, rules, and dictionaries, and one module uses the MetaMap program. An evaluation of the system showed convincing results for the recognition of drug names, dosage, and route of administration. However, the system attained poor results for the identification of reasons for drug administration.

3.14.8 MetaMap

MetaMap[41] is an open source software that maps biomedical free text to concepts in the UMLS metathesaurus. It uses a knowledge intensive approach based on symbolic, natural language processing, and computational linguistic techniques [Aronson (2001)].

There are several options that control the input and output behavior of the program. Given an arbitrary text, it is parsed into simple noun phrases performed by the SPECIALIST[42] minimal commitment parser which produces a shallow syntactic analysis of the text. For each phrase, variants are generated using the knowledge in SPECIALIST lexicon and a supplementary database of synonyms. A variant consists of a phrase word together with all its synonyms, abbreviations, derivational variants, inflections, and spelling variants. The candidate set of metathesaurus strings containing at least one of the variant is retrieved and evaluated against the input text by first computing a mapping from the phrase words to the candidate's words and then calculating the

[40] http://icb.med.cornell.edu/crt/textractor/
[41] http://metamap.nlm.nih.gov/
[42] http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html

(1)	We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia.
(2)	Hemofiltration-TREATS-Patients Digoxin overdose-PROCESS_OF-Patients hyperkalemia-COMPLICATES-Digoxin overdose Hemofiltration-TREATS(INFER)-Digoxin overdose

Figure 3.14: Illustration of an arbitrary sentence processed by the SemRep program.

strength of mapping by a linguistically principled function. Figure 3.13[43] shows an example for a result of mapping process performed by the MetaMap program. The noun phrase “sodium channel inhibitor" is mapped to UMLS concept “Sodium Channel Inhibitors" with semantic type “Pharmacological Substance".

3.14.9 SemRep

SemRep[44] (Semantic Representation) is an open source knowledge extraction and representation framework developed by the NLM. The SemRep program identifies UMLS concepts and relationships in any arbitrary text. SemRep utilizes the MetaMap for first identifying the semantic concepts in input text. Later on, a rule-based approach determines the relationships between concepts occurring within a sentence.

SemRep can handle any form of free-text data including abstracts, full-texts, and medical records. SemRep is available in the interactive mode as well as the batch mode. The interactive mode allows immediate submission and retrieval of results, and is suitable when the size of data is small. The batch mode is designed for large scale processing of documents. Using the batch mode, users can submit as many as millions of documents to the serves in NLM. However, there is no evaluation of the performance of system available. Figure 3.14[45] shows an example of an arbitrary sentence processed by the SemRep program.

3.15 Text Mining Scenarios

3.15.1 I2B2 Challenge

I2B2[46] (Informatics for Integrating Biology and the Bedside) is an open-assessment challenge for the evaluation of state-of-the-art systems for information retrieval and information extraction in medicine. The competition is primarily organized the National Centres for Biomedical Computing[47] and the event is held annually. The first I2B2 challenge was held in 2007 and was known as the *Smoking Challenge* [Uzuner et al.

[43] Adapted from `http://skr.nlm.nih.gov/interactive/metamap.shtml`
[44] `http://skr.nlm.nih.gov/`
[45] Adapted from `http://skr.nlm.nih.gov/`
[46] `https://www.i2b2.org/NLP/Coreference/PreviousChallenges.php`
[47] `http://www.ncbcs.org/`

(2007)]. This challenge aimed at classification of de-identified patient health records into five possible categories based on the information contained in records and based on their medical intuitions. The pre-defined categories were *past smoker*, *current smoker*, *smoker*, *non-smoker*, and *unknown*. The second challenge held in 2008 was known as the *Obesity Challenge* [Uzuner (2008)]. The obesity challenge was a multi-class, multi-label classification task focused on obesity and its co-morbidities. The task was to classify the obesity information and co-morbidities at a document level as present, absent, questionable, or unmentioned in the documents. The goal of the challenge was to evaluate systems on their ability to recognize whether a patient is obese and what co-morbidities they exhibit. The third competition was held in 2009 and known as the *Medication Extraction Challenge* [Uzuner et al. (2010)]. The challenge aimed to encourage development of natural language processing systems for the extraction of medication-related information from narrative patient records. Information to be targeted included medication names, dosages, modes of administration, frequency of administration, and the reason for administration. The recent challenge in 2010 was known as the *Relations Challenge* [Uzuner et al. (2011)]. This challenge aimed at evaluation of systems for identifying the medical concepts in patient health records. The categories of concepts include treatments, tests, and problems. Two additional triers aimed as classification of assertions made over medical problems, and the identification of relationships between different categories of medical concepts. All the challenges demonstrated various levels of success with the participants from industries as well as academia. Finally, the I2B2 makes medical corpora, ground truth annotations, and evaluation protocols publicly available under certain licensing agreements.

3.15.2 TREC

Text Retrieval Conference (TREC)[48] aims at open evaluation of state-of-the-art systems for information retrieval in different domains. Its purpose is to promote and encourage research within the information retrieval community by providing a workbench necessary for large-scale evaluation of information retrieval techniques as well as to speedup the transfer of technology from lab-to-product. A common platform allows researchers coming from different domains to learn about the state-of-the-art problem solving approaches, reduce redundancy in the research, and promote academia-industry collaborations.

TREC began in 1992 and is co-sponsored by the National Institute of Standards and Technology[49]. TREC runs several tracks every year including the Genomics tracks [Hersh and Bhupatiraju (2003)] held from 2003 to 2007. Genomics tracks focussed on ad-hoc retrieval of genomics full text literature. Another set of challenging tracks were held under the banner of Chemical track (also referred to as TRECCHEM) from 2009 to 2011 [Lupu et al. (2009)]. The TRECCHEM provides a platform for evaluation of information retrieval from patents and full-text literature in the biomedical and

[48] http://trec.nist.gov/
[49] http://www.nist.gov/index.html

chemistry domains. In 2009 and 2010, two independent tasks namely *Technology Survey* and *Prior Art Search* were conducted. In the final year 2011, a new task for chemical image to structure conversion (I2S) was introduced in additional to the earlier tasks [Lupu et al. (2011)].

In 2011, a new track for retrieval of information from patient health records was started (known as Medical Records Track) [Voorhees and Tong (2011)]. "The goal of the Medical Records track is to foster research on providing content-based access to the free-text fields of electronic medical records. In the initial year, the track focuses on a task that models the real-world task of finding a population over which comparative effectiveness studies can be performed"[50].

3.15.3 CMC Challenge

The CMC challenge was held in 2007 and co-organized by the Computational Medicine Centre[51] [Pestian et al. (2007), Farkas and Szarvas (2008)]. The goal of the competition was to create and train computational intelligence algorithms that automate the assignment of ICD-9-CM codes to clinical free text. The task involved over 25 participants worldwide with participants from industries as well as academia. Similar to the principles of TREC, the CMC challenge aimed at bringing together academia and industries to work together on a common platform and promote joint research interests in the medical arena. Different participants of the CMC challenge competed with a variety of challenging approaches including statistical, machine learning-based, and rule-based systems. The final outcome of the CMC challenge demonstrated that expert rule-based approaches perform competitively or even outperform purely statistical approaches for the ICD-9-CM coding of radiology reports.

3.15.4 TMMR

In 2008, the Canadian National Research Council[52] conducted a research on knowledge discovery from free-text medical records through a project called Text Mining of Medical Records (TMMR)[53]. The project aimed to establish text mining tools that are flexible and adaptable so that they can be applied by the end-user for processing electronic health records and other medical text. The text mining tool development was motivated for an improvement of health and wellness, by increasing the efficiency and effectiveness of medical researchers and other health professionals. The project addressed two principal scenarios namely *producing alerts* and *extracting medical facts* for processing the text from medical records.

[50] http://www-nlpir.nist.gov/projects/trecmed/2011/tm2011.html
[51] http://computationalmedicine.org/challenge/previous
[52] http://www.nrc-cnrc.gc.ca/index.html
[53] http://www.nrc-cnrc.gc.ca/eng/projects/iit/text-medical.html

Chapter 4

Evaluation of Terminologies for Medical Disorder Recognition

A disease in the context of human health is an abnormal condition that impairs the bodily functions and is associated with physiological discomfort or dysfunction. Similarly, an adverse effect is a health impairment that occurs as a result of intervention of a drug, treatment or therapy [Ahmad (2003)]. Mentions of both diseases and adverse effects (constitutively known as *medical disorders*) in free-text denote special entity classes for the medical experts, clinical professionals as well as health care companies [Hauben and Bate (2009), Forster et al. (2005)]. This not only helps in understanding the underlying hypothetical cause but also provide rationale means to prevent or diagnose such abnormal medical conditions. Furthermore, from the text mining perspective, precise identification of medical disorders can promote the development of semantic medical document search engines to obtain disorder-centric information (*see* Chapter 7) which in-turn can support disease surveillance, epidemiological studies and so-forth. Furthermore, it helps in identifying relationships with other classes of entities such as treatments or diagnostics that can support knowledge discovery, hypothesis generation and medical decision-making.

Identification of medical disorders in free-text is a challenging issue due to the existence of various forms in denoting their mentions as well as the ambiguous nature. Examples include synonyms (*e.g. cancer*, *carcinoma*, *malignancy*, *etc.*), abbreviations (*e.g. multiple sclerosis* abbreviated as *MS*), ambiguity (*MS* can mean *multiple sclerosis* or *mitral stenosis*), writing variations (*e.g. anemia*, and *anaemia*), and descriptiveness (*e.g. thrombocytopenia* written as *subnormal levels in blood platelet count*). Several terminological resources exist that provide information about diseases and adverse effects such as the MeSH, UMLS, ICD, and many more. Most of these medical terminological resources have been developed to serve different purposes. For example, the MeSH is used for indexing the documents within bibliographic databases such as PubMed, and NLM's clinical trials database. Similarly, the ICD is used within clinical settings by physicians for coding the diseases associated with patients. Although, these resources serve as a good basis for the dictionary-based named entity recognition in text, not all of them essentially suit the text mining needs. According to the author's knowledge, there are extremely limited noticeable efforts in the past aiming at evaluation of different medical terminologies on a common workbench in order to support the natural

language processing. Therefore, the aim of this work is to provide an overview on different data sources and evaluate the usability of the contained disease and adverse effect terminologies for named entity recognition in biomedical free-text.

4.1 Terminological Resources

Dictionary-based named entity recognition approaches rely on comprehensive terminologies containing frequently used synonyms and spelling variants. Such resources include databases, ontologies, controlled vocabularies and thesauri. This section gives an overview of the available data sources for diseases and adverse effects. Examples of synonyms and term variants associated with the MeSH disease concepts are provided in Table 4.1

Different resources have been designed to meet the needs of different user groups whereas some of them include certain disease specific information. For example, the NCI thesaurus serves as a reference terminology and an ontology providing a broad coverage of cancer domain including cancer related diseases, findings, abnormalities, gene products, drugs, and chemicals. Similarly, there are databases that include very specific organ or disease class related information such as the autoimmune disease database [Karopka et al. (2006)] or the DSM-IV codes[1] which is specific to mental disorders. On the other hand, sources such as the ICD-10, the UMLS and the MedDRA provide a wider coverage of diseases, signs, symptoms, and abnormal findings irrespective of any kind of disease or any affected organ system. All these resources have their own advantages and areas of applicability. Therefore, the survey made here includes only those resources that encompass information about medical abnormalities that are associated with the entire human physiology. From all the resources introduced here, individual dictionaries were generated and evaluated over a manually annotated corpus. Although, the MeSH, ICD-10, MedDRA, and SNOMED-CT are already included as source vocabularies within the UMLS, these resources were separately downloaded from their respective official websites. The main reason is because when the terms from the source vocabularies are imported into the UMLS, they undergo a series of term modification steps[2]. This provides an impression that the terms present in the UMLS may not be identical to the terms present in the source vocabularies. Therefore, in order to validate the hypothesis of suitability of the individual resources for text mining, they were treated as independent terminologies. The analyzed terminologies are as follows:

MeSH contains concepts that are arranged in a hierarchical order and associated with synonyms and term variants (*see* Section 2.1.1). A subset of MeSH that corresponds to the category *Diseases* (tree node identifiers starting with 'C') was extracted to generate a dictionary. The MeSH dictionary contains over 4,300 entries.

[1] Diagnostic and Statistical Manual of Mental Disorders (DSM) 4th Edition, `http://www.psych.org/mainmenu/research/dsmiv/dsmivtr.aspx`

[2] `http://www.nlm.nih.gov/research/umls/knowledgesources/metathesaurus/sourcefaq.html`

ID	Concept	Synonyms
D000292	Pelvic Inflammatory Disease	Adnexitis, Inflammatory Disease; Pelvic, Inflammatory Pelvic Disease; Pelvic Disease, Inflammatory
D002534	Brain Hypoxia	Anoxia, Brain; Anoxic Brain Damage; Brain Anoxia; Brain Hypoxia; Cerebral Hypoxia; Encephalopathy, Hypoxic; Hypoxic Brain Damage; Hypoxic Encephalopathy

Table 4.1: Examples of synonyms and term variants associated with concepts in the MeSH database.

MedDRA provides a hierarchical structure of concepts that include signs, symptoms, diseases, diagnosis, therapeutic indications, medical procedures, and familial histories (*see* Section 2.1.5). The MedDRA dictionary contains over 20,000 entries associated with synonyms and term variants.

ICD-10 provides concepts that are hierarchically ordered according to the organ system that is being affected (*see* Section 2.1.6). Unlike other resources, the ICD provides a flat list of terms and does not include synonyms or term variants. The complete ICD-10 was used for generating the dictionary and it contains over 70,000 entries altogether.

SNOMED-CT concepts are organized into hierarchies and the sub-hierarchy that corresponds to *Disorder* was used to generate a dictionary (*see* Section 2.1.3). The SNOMED-CT dictionary contains over 90,000 concepts associated with synonyms and term variants.

UMLS concepts are categorized according to semantic groups[3] (*see* Section 2.1.2). The semantic group *Disorders* contains semantic subgroups such as *Acquired Abnormality*, *Disease or Syndrome*, *Mental or Behavioral Dysfunction*, *Sign or Symptom*, etc. All concepts in the *Disorders* semantic group of the UMLS were used to generate a dictionary. This dictionary contains over 110,000 entries altogether.

4.2 Dictionary Characteristics

The dictionaries generated for the recognition of diseases and adverse effects were analyzed with regard to the total number of entries, number of synonyms provided, and availability of mappings to other data sources.

Table 4.2 provides a quantitative estimate of the entities present in the raw dictionaries. The UMLS has the largest collection of disease and adverse effect data followed by the SNOMED-CT. Figure 4.1 shows the distribution of synonyms for all the analyzed dictionaries. Since the ICD-10 does not provide synonyms and term variants, it is

[3] http://semanticnetwork.nlm.nih.gov/SemGroups/

	MeSH	MedDRA	ICD-10	SNOMED-CT	UMLS
No. of entries	4,350	20,515	74,830	92,376	112,341
No. of synonyms (incl. concepts)	42,631	69,121	74,830	170,561	295,773
Cross mappings	no	yes	no	yes	yes

Table 4.2: A quantitative analysis of the dictionaries generated for the disease and adverse effect named entity recognition. Total number of entries, number of synonyms, and the availability of inter-data source mappings for individual dictionaries are reported.

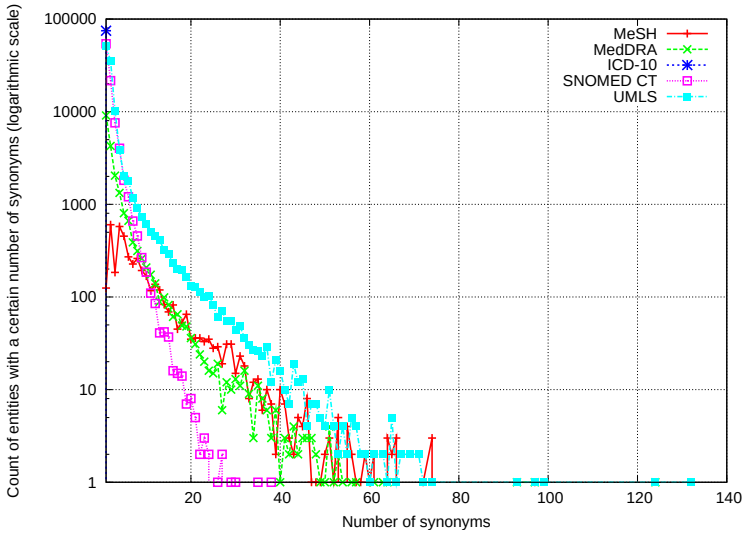

Figure 4.1: Plot of the synonym count distribution for all the analyzed dictionaries.

visible only as a point in Figure 4.1. A large part of all the dictionaries contain less than 20 synonyms. Few entries in the UMLS, MeSH, and MedDRA are associated with as much as more than 60 synonyms. Resources with high number of synonyms are of great value for dictionary-based named entity recognition approaches. They help to overcome a high false negative rate but may pose a risk of high number of false positives requiring a dedicated curation.

Since UMLS is the largest resource, a survey was conducted to check the percentage of synonyms that overlap with synonyms in rest of the resources. The synonym comparison between the different resources was performed using a simple case-insensitive string match (*i.e.* only complete string matches were accepted). About 96% of the MeSH and 23% of the MedDRA synonyms are present in UMLS. Only 4% of the ICD-10 and 9% of the SNOMED-CT synonyms are covered by UMLS. Hence, the outcome of this survey showed that integrating the smaller resources with UMLS would account for an enhanced terminology coverage.

Although, there is an enormous variation in sizes of the dictionaries used, their adaptability for finding terms in the text is questionable. A manual survey was performed concerning the quality of information contained in each of these dictionaries. The UMLS and SNOMED-CT contained over 20,000 terms each that had special characters such as '@', '#&', '[X]', etc. enclosed within the terms. Examples of such ambiguous terms found in the UMLS are *5-@FLUOROURACIL TOXICITY* and *Congestive heart failure #&124*. A large subset of terms were too long and descriptive composed of more than 10 words. Such synonyms are seldom found in the text. An example of such descriptive term found in ICD-10 is *Nondisplaced fracture of lateral condyle of right femur, initial encounter for closed fracture*. ICD-10 has nearly 35,000 long descriptive terms which constitutes nearly 50% of the entire dictionary. According to the experience of curators, MeSH and MedDRA were regarded as the specialized resources with considerably low level of ambiguity. Nevertheless, few vague entries such as *Acting out*, *Alcohol Consumption*, and *Childhood* were encountered in these dictionaries.

4.3 Corpus Characteristics and Annotation

For evaluating the performance of named entity recognition systems, an annotated corpus is necessary. Since, there is no freely available corpus that contains annotations of disease and adverse effect entities, a corpus containing 400 randomly selected MEDLINE abstracts was generated using 'Disease OR Adverse effect' as a PubMed query. This evaluation corpus was annotated by two individuals who hold a Master's degree in life sciences. All the abstracts were annotated with two entity classes, *i.e. disease* and *adverse effect*. In order to obtain a good estimate of the level of agreement between the annotators, they were insisted to carry out the task independently. First, one annotator participated in the development of a guideline for annotation. The corpus was iteratively annotated by this person along with the standardization of the annotation rules. Later, the second person annotated the whole corpus based on the annotation guideline generated by the first annotator. This procedure formed an evaluation corpus of 400 abstracts containing 1428 disease and 813 adverse effect annotations. Recognizing the boundaries without considering the different classes in the evaluation corpus, the inter-annotator agreement F_1 score and *kappa* (κ) between the two annotators are 84% and 89% respectively which indicates a substantial agreement. The annotation of disease and adverse effect entities were performed very sensitively taking the context into account. Several instances occurred where the disease names and adverse effect names were the same. For example, in the sentence *Before and after MS therapy, patients had no early dumping symptoms, while patients after MS therapy clearly had fewer symptoms such as reflux esophagitis, nausea, and abdominal pain compared with before MS therapy (PMID: 18613449)*, the term *reflux esophagitis* occurs as an adverse effect associated with MS therapy. In contrary, the sentence *A total of 122 patients were receiving PPI treatment for either peptic ulcer disease or reflux esophagitis and were included as the study group (PMID: 20123595)* contains *reflux esophagitis* as a disease being treated. In such cases, the annotators were strictly insisted to use the contextual information

for annotating the entities. Entities that overlap with semantic classes disease and adverse effect are difficult to be recognized unless a context-based disambiguation is performed. Altogether, there were 178 annotated entities that have an overlap with the classes *disease* and *adverse effect*.

4.4 Results of Dictionary Performance

For the identification of named entities in text, the ProMiner system was used along with different dictionaries. The text searching with ProMiner was performed using the raw or unprocessed dictionaries as well as with the processed dictionaries. The search was performed using case-insensitive, word order-sensitive, and the longest string match as constraints. The performance of the ProMiner runs with different dictionaries was evaluated using the Precision and Recall (*see* Section 3.11.1). Evaluations were performed for the complete match as well as partial match between the annotated entities and the dictionary-based entity matches. A partial match is a situation where either the left boundary or the right boundary of the annotated entity and the ProMiner search result are matched. The results with raw dictionaries and such a simple search strategy gives a rough estimate of the coverage of different dictionaries and the effort that has to be invested to curate them. Table 4.4 shows the search results obtained with every individual dictionary when complete matches and partial matches were considered. The highest recall for complete matches were achieved by the MedDRA dictionary (0.62) and the UMLS dictionary (0.58). The recall of ICD-10 was the lowest of all dictionaries covering only 10% of the entities annotated in the corpus. Unlike other dictionaries, ICD-10 lacks information about the synonyms and term variants which hinders it from covering different types of variants mentioned in the text. The combination of results of all the dictionaries lead to a promising recall of 0.75. Another important observation is the low recall (0.18) attained by the SNOMED-CT dictionary. Although, this dictionary contains over 90,000 entries with 170,561 different terms, its usability for finding entities in the text seems extremely limited. One reason is because of the descriptive nature of most of the terms present in the SNOMED-CT vocabulary such as *Spastic paraplegia associated with T-cell lymphotropic virus-1 infection*. Although such long descriptive terms provide substantial information about the medical condition, they are not quite often used in the literature. Additional reasons are the perception of named entities in annotator's mind as well as the style adopted by the annotation guideline. Perhaps, our principle annotators would annotate such a textual description with *Spastic paraplegia* and *T-cell lymphotropic virus-1 infection* as two distinct entities rather than annotating the entire phrase as one single entity.

Comparison of the results of complete matches and partial matches in Table 4.4 shows the granularity of information covered by different data sources and the textual explications. The UMLS and MedDRA achieved an overall recall of 0.73 and 0.72 respectively for the partial matches whereas the combined results of all the dictionaries achieved a highest recall of 0.92. This provides an indication that the terms contained in these dictionaries cover the head nouns associated with the disease and adverse effect

	MeSH	MedDRA	ICD-10	SNOMED CT	UMLS
No. of entries	4,335	18,273	37,263	84,292	100,871
No. of synonyms (incl. concepts)	42,531	57,017	37,263	146,545	243,602

Table 4.3: A quantitative analysis of the *curated* dictionaries applied for the disease and adverse effect named entity recognition. Total number of entries and number of synonyms present within the individual dictionaries are reported.

entities but does not include different enumerations used in the literature. For example, in the case of *progressive neurodegenerative disorder*, only *neurodegenerative disorder* was identified whereas the adjective *progressive* was not covered. Based on the experience of the curators and the results from Table 4.4, nearly 10% of the mismatches are caused by the medical adjectives such as *chronic*, *acute*, and *idiopathic* that are frequently used in text but not provided by the resources. Another source of mismatch is the anatomical information often attached to the disease entity in text. For example, in the case of *vaginal squamous cell carcinoma*, only the *squamous cell carcinoma* was recognized whereas the remaining anatomical substring remained unidentified.

The highest precision rates for the complete matches were achieved by the MeSH dictionary (0.54) and the MedDRA dictionary (0.48) hence validating the curator's opinion about the quality of these resources. The lowest precision of 0.18 was achieved by the UMLS dictionary. The precision after combining the results of different dictionaries was considerably low due to the overlapping false positives generated by different dictionaries. The low precision is due to the presence of noisy terms such as *disease* and *response* within the dictionaries. Amount of such noisy terms considerably varies amongst different resources with UMLS having the highest. Therefore, the curation of dictionaries is necessary in order to achieve better performance. Experiences from the previously reported dictionary-based named entity recognition approaches let us assume that the precision could be greatly improved by the dictionary curation.

4.4.1 Dictionary Curation

The dictionaries were processed and filtered based on a subset of pre-defined rules in order to reduce the level of ambiguity associated with them. Most of the rules were adapted from Hanisch et al. (2005), Hettne et al. (2009), and Aronson (1999). The rules that were applied for processing the dictionaries are listed below. All the rules were used in common to all the analyzed dictionaries.

Remove very short tokens: Single character alphanumericals that appear as individual synonyms were removed. For example, *5* was mentioned as a synonym of the concept *Death Related to Adverse Event* in the UMLS.

Remove terms containing special characters: Remove all the terms that contain unusual special characters such as '@', ':' and '&#'. An examples of such term in

		Raw			Curated			Disambiguation		
Dictionary	Match type	All	DIS	AE	All	DIS	AE	All	DIS	AE
MeSH	*Complete*	0.54/0.43	0.46	0.40	0.61/0.43	0.46	0.40	0.61/0.43	0.46	0.40
	Partial	0.73/0.58	0.64	0.51	0.80/0.57	0.62	0.51	0.80/0.57	0.62	0.51
MedDRA	*Complete*	0.48/0.62	0.64	0.59	0.57/0.61	0.63	0.59	0.60/0.61	0.62	0.59
	Partial	0.55/0.72	0.76	0.68	0.67/0.72	0.75	0.68	0.69/0.71	0.74	0.68
ICD-10	*Complete*	0.46/0.10	0.10	0.10	0.57/0.15	0.10	0.19	0.57/0.15	0.10	0.19
	Partial	0.59/0.15	0.15	0.14	0.66/0.19	0.14	0.23	0.57/0.19	0.14	0.23
SNOMED	*Complete*	0.38/0.18	0.18	0.18	0.40/0.20	0.22	0.18	0.43/0.18	0.20	0.15
	Partial	0.66/0.28	0.33	0.23	0.69/0.34	0.39	0.28	0.71/0.34	0.39	0.28
UMLS	*Complete*	0.18/0.58	0.60	0.55	0.33/0.57	0.60	0.54	0.36/0.57	0.60	0.54
	Partial	0.25/0.73	0.74	0.71	0.43/0.72	0.73	0.71	0.46/0.72	0.73	0.71
Combined	*Complete*	0.12/0.75	0.80	0.70	0.18/0.76	0.81	0.71	0.19/0.76	0.80	0.71
	Partial	0.14/0.92	0.92	0.91	0.21/0.91	0.92	0.89	0.22/0.91	0.92	0.89

Table 4.4: Comparison of performances of different dictionaries tested over the evaluation corpus. The results are reported for the *complete matches* and *partial matches* of annotated classes disease (DIS), adverse effect (AE) and a combination of both the classes (All). For a combination of both the classes, *i.e. All*, the precision and recall values are reported. For the classes DIS and AE, only the recall values are reported. 'Combined' indicates the performance achieved by combining the results of all the dictionaries.

SNOMED-CT is *Heart anomalies: [bulbus/septum] [patent foramen ovale]*.

Remove under-specifications: Substrings such as NOS, NES, and not elsewhere classified were removed away from the terms. Such strings were often encountered at endings of the dictionary terms. An example of such a term from MedDRA is *Congenital limb malformation, NOS*

Remove very long terms: Very long and descriptive terms that contains more than 10 words were removed. An example of such a term found in SNOMED-CT is *Pancreas multiple or unspecified site injury without mention of open wound into cavity*. Although such long terms do not appear in the text, filtering them from the dictionary gradually reduces the run time of the process.

Remove unusual brackets: Unusual substrings that often appear within the brackets were removed from the terms. Examples of such terms found in SNOMED-CT include *[X]Papulosquamous disorders* and *[D]Trismus*.

Remove noisy terms: ProMiner with different dictionaries was run over an independent corpus of 100,000 abstracts that were randomly selected from MEDLINE. The 500 most frequently occurring terms matched with the individual dictionaries were manually investigated to remove the most frequently occurring false positives. This

process will improve the precision of entity recognition during the subsequent runs.

In addition to dictionary curation, the configuration of the ProMiner system was readjusted to match the possessive terms (*e.g.* Alzheimer's disease) that contain *'s* substring at the word endings. After the end of the dictionary processing and filtering, the number of entries and synonyms that remained in the individual dictionaries can be found in Table 4.3. The MeSH dictionary sustained minimum changes with only 15 entries being removed whereas ICD-10 underwent a large noticeable change. The size of the ICD-10 dictionary was reduced to nearly half of the previously used raw dictionary. The search results obtained with every individual curated dictionary can be found in Table 4.4. As the result of dictionary curation, the performance of all the dictionaries improved remarkably well. For the complete matches, the precision of UMLS dictionary raised by 15% with a drop in recall by just 1%. Other dictionaries that benefited well from the curation process are ICD-10 and MedDRA with raise in their precision by 11% and 9% respectively.

4.4.2 Acronym Disambiguation

In spite of processing the dictionaries by removing the noisy terms as well as lexical modification of the synonyms, the acronyms present in the dictionaries turned out to be another source of frequent false positives. For example, *ALL* which is an acronym for *Acute Lymphoid Leukemia* generated a considerable noise. Therefore, acronyms present in all the dictionaries that have two to four characters were collected in a separate acronym list. Whenever there is a match between the term in the acronym list and the text tokens, a rule was defined in order to accept or neglect the match. This disambiguation facility is available within the ProMiner system. The acronym disambiguation rule accepts the match based on two criteria and they are:

- The match should be case sensitive.
- The acronym as well as any one of its synonym in the respective dictionary should co-occur anywhere within the same document.

For example, the term *ALL* is associated with 17 synonyms in the MedDRA dictionary. Any case-sensitive match between the *ALL* and tokens in the text would be accepted if any one synonym of the *ALL* occurs within the same abstract. The search results obtained with the individual curated dictionaries in addition to the acronym disambiguation can be found in Table 4.4. Considering the complete matches, the acronym disambiguation raised the precision of MedDRA, SNOMED-CT and UMLS dictionaries by about 3% each. The performance of MeSH and ICD-10 remain unaffected indicating the presence of less acronyms within them. There was a marginal decline (less than 2%) in the recall of the dictionaries after applying the disambiguation rule. This indicates the success of applying the rule-based acronym disambiguation for effective filtering of noisy acronyms.

4.5 Discussion

This chapter describes the challenges associated with the identification of diseases and adverse effects in free-text articles using the standard medical terminologies. A corpus containing 400 systematically annotated MEDLINE abstracts was generated that was used within a common workbench to study the performances of different resources. An outcome of this survey upheld the MedDRA as a compatible resource for the text mining needs having its recall competitive to the UMLS meta-thesaurus with considerably fair precision upon processing. The UMLS being the largest resource does not include all the names that are covered by the smaller resources. Hence, the combination of the search results from all the terminologies lead to an increase in the recall. This indicates a need for intelligent ways to integrate and merge the information spread across different resources. The amount of work that needs to be invested to curate very large resources such as the SNOMED-CT and UMLS in order to make them applicable for the text mining is also shown.

In addition to the performance comparison, the effect of dictionary curation and a limited manual investigation of the noisy terms showed to be very effective with a significant improvement in the precision of entity recognition with different resources. A rule-based processing coupled with the dictionary curation can substantially improve the performance of recognition of diseases and adverse effects.

The applied strategy for the identification of medical disorders can promote the development of semantic search engine for capturing the disorder-centric knowledge that has been systematically dealt in Chapter 7. It also supports the pharmacovigilance studies (*see* Chapter 10) by spotlighting the zones in documents (*e.g.* sentences) that can be further analyzed.

Although, the dictionary-based recognition of medical disorders is a forthright and comprehensive approach, an ability of the machine learning-based approach to perform this task needs to be tested. Previous experiments have shown that machine learning techniques for disorder recognition can perform competitively [Li et al. (2008)]. Therefore, an adaptation and evaluation of the machine learning-based approach is discussed in the upcoming chapter.

Chapter 5

Machine Learning Strategy for Medical Disorder Recognition

Availability of comprehensive and expert-modeled medical terminologies provide an adequate backbone for the dictionary-based named entity recognition approaches to support recognition of medical disorders in free-text. But, an inherent drawback associated with the dictionary-based approach is its dependence on the quality and coverage of the data source used for generating the dictionary. For example, if a disease name is not covered within any standard terminologies, its mentioning in text cannot be detected by dictionary-based techniques. In contrary, the approaches that depend on rule-based or machine learning strategies benefit from the dictionary independence and help in covering the different term variants and enumerations that are not covered by the dictionaries. Therefore, the work reported here presents a Disorder-Recognizer which uses a machine learning strategy based on Conditional Random Fields (CRF) for identifying the mentions of medical disorders in free-text. The impact of active learning, feature selection, and the use of additional information from the domain dictionary on the learning process is discussed in the following sections.

5.1 Corpus Generation

Any supervised learning problem requires independent training and test sets. For the purpose of training, a seed corpus (referred to as Disorder-train) containing 300 randomly selected MEDLINE abstracts was used. Disorder-train corpus was annotated for the mentions of medical disorders with an entity class label **DISORDER**. This seed training corpus contains 1,194 annotated entities. For testing the trained model, previously generated corpus described in Section 4.3 (referred to as Disorder-test) was employed. This corpus contains the medical disorders annotated with two different class labels **DISEASE** and **ADVERSE EFFECT**. Both these annotated classes were merged to form one main class (*i.e.* **DISORDER**). The test corpus contains 400 MEDLINE abstracts having 2,241 annotated entities. For the purpose of active learning, an additional corpus of 100,000 randomly selected MEDLINE abstracts (referred to as Disorder-al) was used.

Labels:	**O**	**B-DISORDER**	**I-DISORDER**	**O**	**O**
Tokens:	The	rectal	cancer	was	diagnosed

Table 5.1: Example of observation and label sequence for a text snippet after its tokenization.

5.2 Training with Conditional Random Fields

Conditional Random fields (CRF, *refer* Section 3.8.6) is a machine learning technique that has been widely applied for modeling the sequential data. In the context of NER, the input sequence corresponds to the tokenized text where the text is split at white spaces, punctuation marks, and parenthesis in general. The label sequence is coded using the label alphabet:

$$\mathcal{L} = \{\textbf{I-DISORDER}, \textbf{O}, \textbf{B-DISORDER}\}$$

where $y_i =$ **B-DISORDER** means that x_i is the beginning token of the medical disorder, $y_i =$ **I-DISORDER** means that x_i is the continuation token of medical disorder and $y_i =$ **O** means that x_i is a token not of interest. An example of the tokenized and labeled text snippet is provided in Table 5.1

The technical details of CRF can be found in the report of Roman and Tomanek Klinger and Tomanek (2007). The implementation of Disorder-Recognizer is based on MALLET McCallum (2002), a widely used system for liner-chain CRF.

5.2.1 Feature Extraction

The features used for training the CRF can be broadly categorized into *morphological*, *context-based*, and *ProMiner-based* features. An overview of different features used is depicted in Table 5.2. *Morphological* features are concerned with the internal structure of the tokens. They include *static morphological* features as well as *automatically generated morphological* features. *Context-based* features use information about the surrounding elements for every token. *ProMiner-based* features use information from the named entities recognized by ProMiner in both training and test sets (*e.g.* in this case, check for the matches between prefixes/suffixes of the token and the ProMiner identified terms). MedDRA was used as a dictionary for the ProMiner-based NER.

5.3 Performance Evaluation Criteria

The evaluation of NER was performed using exact match as a criterion. An exact match is a situation where both the left and right boundaries of the annotated disorder name is correctly recognized by the system. The performance of the system was judged based on the Precision, Recall, and F_1 score. Under the preliminary settings (Section 5.4) as well as during the active learning (Section 5.5), and feature selection (Section 5.6), the

Name	Explanation
Static morphol. features	
AllCaps	Match regex: [A-Z]+
IsSlash	Match regex: [\/]
IsQuote	Match regex: [“ ” ‘ ’]
IsDash	Match regex: [-]
Autom. generated morphol. features	
Affixes	Autom. generation of a feature for every token: match that prefix or suffix (lengths: 2, 3, 4)
WordAsClass	Autom. generation of a feature for every token: match that token
POS	Parts-of-speech tag of a token
Lemma	Lemmatized form of a token
Context-based features	
Spaces	Is a token preceded or succeeded by white space
OffsetConjunction	Add features of preceding and succeeding tokens for every token (order: 1, 2, 3)
ProMiner-based features	
EntityAffixes	Prefixes and suffixes (lengths: 3, 4) of intermediate and last words of named entities recognized by ProMiner
LexiconMatch	Check if a token appears in the lexicon of named entities recognized by ProMiner

Table 5.2: Example of features used for training the CRF for disorder recognition.

performance of the system was evaluated by 10-fold cross validation (*see* Section 3.10.1). During the final comparative assessment of different tools (Section 5.7), the evaluations were performed on the stand-off annotations of the complete test set.

5.4 Preliminary Evaluation of NER

In the first step, the CRF was trained and evaluated by 10-fold cross validation over the DISORDER-TRAIN corpus that contains 300 annotated abstracts. Under this preliminary settings, all the *morphological* features, Spaces, and order-1 offset conjunction were used whereas the *dictionary-based* features were set to idle. The performance of the system is shown in Figure 5.1. The system attained the F_1 score of 0.63 ± 0.04.

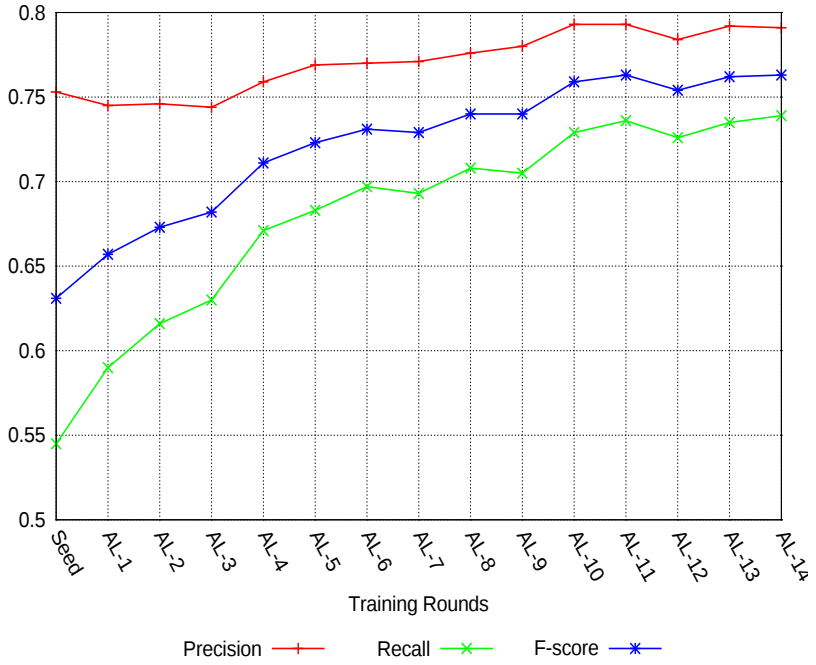

Figure 5.1: Results of the disorder recognition achieved during different rounds of active learning.

5.5 Training Corpus Extension and Evaluation during Active Learning

Since the performance of the preliminary model was below expectation, an immediate step was to extend the training corpus with additional annotated abstracts. For this purpose, the concept of active learning (*see* Section 3.9) was employed. The principle behind active learning is to systematically select the abstracts from DISORDER-AL corpus that can be annotated and added to the seed corpus DISORDER-TRAIN for further training. Active learning is an iterative process that is carried out repeatedly until a stopping criterion is reached. The stopping criterion applied here was the convergence of F_1 score. The process of active learning is shown in Algorithm 1.

Figure 5.1 shows the performance of the system at different rounds of active learning. During each round of active learning, the preliminary feature set described in Section 5.4 was used. Altogether, 14 rounds of active learning were performed in order to observe the convergence in performance of the system. The process of active learning resulted in an extended training corpus (referred to as DISORDER-TRAIN-AL) containing 860 abstracts and having 15,288 annotated entities. Finally, it turned out that training over an extended corpus substantially improved the performance of the system with F_1 score of 0.76 ± 0.04.

Algorithm 1: The process of active learning

Require: The annotated abstracts in DISORDER-TRAIN
Require: The unannotated abstracts in DISORDER-AL
Require: The model M trained on DISORDER-TRAIN
Repeat
1. Apply the trained model classifier M on DISORDER-AL
2. Rank the abstracts of DISORDER-AL according to a performance measure
3. Manually annotate the top ranked 40 abstracts of DISORDER-AL and add them to DISORDER-AL
4. Train the model on the extended DISORDER-TRAIN and evaluate by 10-fold cross validation
Until the stopping criterion is reached

	Precision	Recall	F_1 score
DISEASE	0.64	0.69	0.66
ADVERSE EFFECT	0.74	0.36	0.48
Overall	0.69	0.53	0.60

Table 5.3: Assessment of system's performance for the identification of diseases and adverse effects separately.

Performance Evaluation for Identification of Diseases and Adverse Effects

Documents present in the seed DISORDER-TRAIN corpus as well as articles gathered during active learning were in parallel annotated with **DISEASE** and **ADVERSE EFFECT**. Ability of the system to differentiate between two classes was tested. The DISORDER-TRAIN corpus after active learning contains 12,039 **DISEASE** and 3,249 **ADVERSE EFFECT** annotations. Performance of the system for identification of diseases and adverse effects separately evaluated by 10-fold cross validation is provided in Table 5.3. This provides a rationale supporting the fact that diseases and adverse effects are hard to be differentiated and confronts a challenging scenario for the automatic identification.

5.6 Feature Selection

The fundamental purpose of feature selection is to study the influence of different features on the performance of the learning process. Experiments with CRF in the past have shown that feature selection can improve the performance of the system [Klinger et al. (2008)]. Omitting the non-informative features can reduce the number of features used for training as well as the processing time.

The impact of different features on the system's ability to correctly recognize the

	Precision	Recall	F_1 score
Prel. features	0.75	0.54	0.63
Active learning	0.79	0.74	0.76
Feature selection	0.84	0.76	0.80

Table 5.4: Assessment of system's performance with preliminary (baseline) features, active learning, and feature selection.

disorder mentions was tested systematically. This was performed by setting various features to idle or by adding new features to the preliminary feature set. For every modified feature set, a separate model was trained and validated by 10-fold cross-validation. The results of feature selection is shown in Figure 5.2. Most of the features from class *static morphological* had negligible impact on the performance of the system. Therefore, only the features that showed considerable change in the performance are indicated in Figure 5.2. Leaving out the WordAsClass and Affixes resulted in a substantial decline in recall of the system. This points out that it is necessary for a system to learn the prefixes and suffixes of tokens and whether the tokens appear within the names of medical disorders or not. Increasing the OffsetConjunction to order-2 or order-3 resulted in a slight increase in precision but noticeably decreased the recall. An important observation was that including the *ProMiner-based* features (*i.e.* EntityAffixes) substantially improved the performance of the system. Specially, the EntityAffixes of length 4, and LexiconMatch resulted in an improvement in the system's performance with F_1 score of 0.79. Furthermore, although the POS tags did not contribute effectively, applying Lemma features substantially improved the system's performance with the highest F_1 score of 0.80 ± 0.04. Table 5.4 shows an assessment of system's performances during the baseline test, after the active learning, and after the feature selection.

Therefore, the results of feature selection indicated that employing a combination of preliminary-feature set and ProMiner-based features is optimum for the identification of medical disorders. Using lemmas as features can provide an additional gain. This apparent optimal feature set was applied for tagging the disorders in Disorder-test corpus later during the comparative assessment.

5.7 Comparative Assessment of Disorder NER

The aim of comparative assessment was to test the performance of the developed Disorder-recognizer in contrast to state-of-the-art methods on a common platform. Therefore, the well known tools such as the MetaMap, ProMiner with MedDRA dictionary, BANNER, and JNET were employed in parallel to the Disorder-recognizer to identify the names of medical disorders in Disorder-test corpus. Since the MetaMap, and ProMiner are based on unsupervised techniques, they were applied directly on

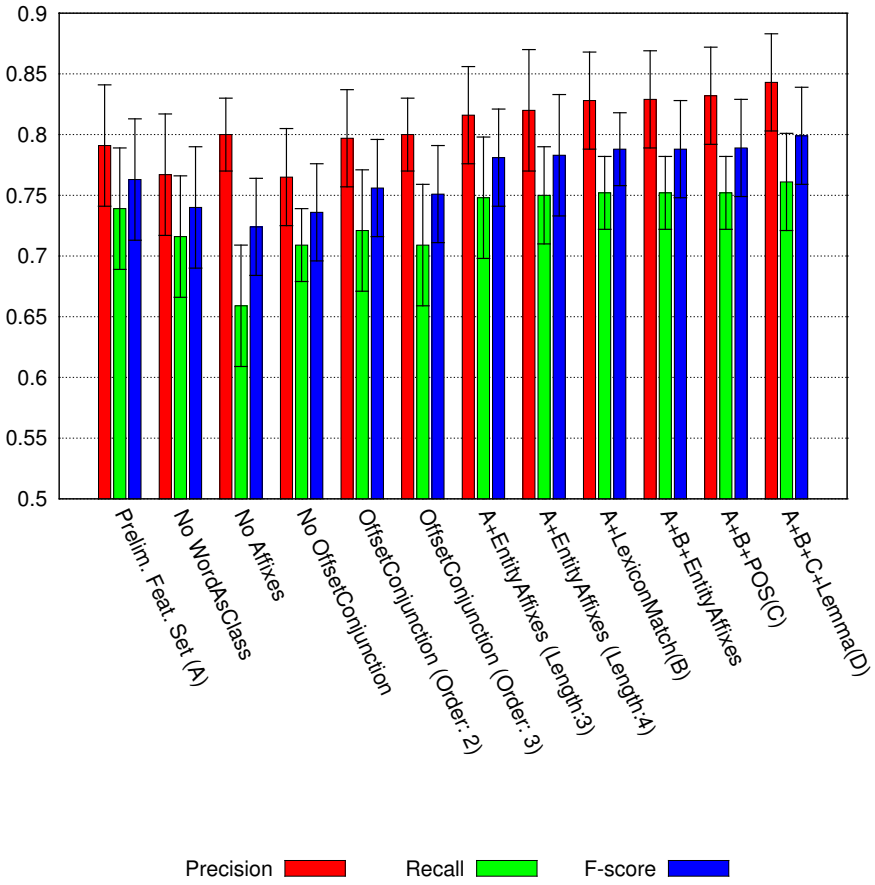

Figure 5.2: Results of feature selection for disorder recognition. The preliminary feature set, LexiconMatch, POS, and Lemma as features have been indicated as A, B, C, and D respectively.

the Disorder-test corpus. The BANNER, JNET and Disorder-recognizer were trained on the Disorder-test-al corpus and applied on the Disorder-test corpus. The performance achieved by different systems is shown in Table 5.5. An overview of these results shows that Disorder-recognizer outperformed the remaining tools by a considerable margin. In general, the machine learning-based techniques performed better than lexical and dictionary-based techniques in identifying the medical disorders. Amongst the three machine learning-based tools used, the JNET achieved the best precision of 0.84 whereas the Disorder-recognizer achieved the best F_1 score of 0.79 balancing its consistency in both precision as well as the recall. JNET and BANNER have been used with best inherent feature sets. However, adding additional features to these systems could improve their performances but this issue is not addressed here.

An evaluation of the performance of Disorder-Recognizer to identify the **DISEASE** and **ADVERSE EFFECT** annotations in Disorder-test corpus was performed thereafter. The system resulted in a recall of 0.76 and 0.74 for the identification of diseases and

	MetaMap	Dictionary-lookup	BANNER	JNET	Disorder-Recognizer
Precision	0.46	0.60	0.80	0.84	0.83
Recall	0.57	0.61	0.69	0.71	0.75
F_1 score	0.51	0.61	0.74	0.77	0.79

Table 5.5: Comparative assessment of disorder named entity recognition.

adverse effects respectively. Currently, the system is not tuned to differentiate between both classes and it identifies one class in general as **DISORDER**.

5.8 Error Analysis

The entities tagged by the Disorder-Recognizer were manually investigated to understand the common errors that a machine learning-based system could experience. The difficulty in identifying the abbreviations was one noticeable source of error. The frequently used abbreviations such as *AD* for *alzheimer's disease* were identified correctly. Whereas, the abbreviations that are in seldom usage such as *T2D* for *Type 2 Diabetes* resulted in observable false negatives. Abbreviations that denote different entity types such as *MCC* that designates *merkel cell cancer* as well as a *gene* generated inevitable errors. An implementation of post-processing strategy for acronym disambiguation may be helpful in overcoming such problems.

Apart from abbreviations, the descriptive enumerations of disorder names caused substantial problems. The system was successfully able to handle several enumerations such as *advanced squamous cell carcinoma of the vagina* that were not completely recognized by rest of the approaches. However, there were few instances such as *metastatic/recurrent squamous cell carcinoma of head and neck* that were not completely recognized by the system. Medical adjectives that are often used to express the severity of a disorder were also erroneous. The system was able to capture most of the entities associated with generic medical adjectives such as *acute pain*, *chronic hypothermia*, *mild headache*, and *severe cardiac attack*. Such expressions are not often covered by the domain terminologies and they are difficult to be captured with dictionary based techniques. However, few instances containing rare adjectives such as *idiosyncratic drug toxicity* were encountered that resulted in partial matches. Finally, the perception of annotators and the resulting annotation errors are also few points that affect the performance of evaluation.

5.9 Discussion

A survey on the performance of CRF-based approach for the identification of medical disorders in text was performed which is one the demanding tasks in the field of biomedical named entity recognition. Training data generation by active learning

followed by systematic enrichment of the feature space and the feature selection showed convincing results. Enhancing the strength of CRF with features from dictionary-based NER showed robust results. The system's performance was compared to state-of-the-art named entity recognizers and found that the so developed Disorder-Recognizer performed superiorly. The system's ability to recognize diseases and adverse effects was also evaluated separately that indicated a challenging scenario.

The Disorder-Recognizer is believed to improve the disease-centric information retrieval as well as the information extraction. The current experiment has demonstrated its ability to successfully identify disorder mentions in MEDLINE abstracts. In the medical domain, apart from MEDLINE abstracts, a huge amount of information is published in the form of e-health records and the ability to identify mentions of medical disorders and other classes of medical entities is crucial to support the development of dedicated search engine for medical text (*see* Chapter 7). This can apparently provide an environment for quick document lookup and support evidence-based medical practices and decision-making [Haynes et al. (2010)]. Therefore, the system's scalability to a new family of corpora (such as medical health records) and its re-trainability to recognize new classes of concepts (such as medical treatments) is essential to be examined and this issue is addressed in the upcoming chapter.

Chapter 6

Concept Identification and Assertion Classification in E-Health Records

The electronic patient health records encompass information about medical problems, diagnosis, and therapeutic interventions associated with the patients. Hence, an automatic processing of the health records helps in understanding the etiologies of medical problems, develop preventive rationales, promote evidence-based medicine, and thereby improve the overall patient's healthcare, safety and effectiveness. Automatic processing of patient health records requires the identification of various categories of medical concepts and the assertions made over them as well as the relationships between different concepts. For example, identifying both new and hidden relations between symptoms, treatments, diagnoses, age, gender, and social situation of the patients can greatly support physicians to take timely decisions, reduce medication errors, and cut down the overall cost of treatment. However, mining the information from health records is not a trivial task since the structure of these articles as well as the writing conventions deviate greatly in comparison to the scientific articles (*see* Section 6.2). Medical text are generated by physicians, healthcare providers, or voice recognition systems that do not strictly adhere to scientific writing standards. Moreover, the medical language used under different clinical settings vary. For example, the radiology department and the surgical pathology department stick to different writing conventions. This poses an additional challenge for the conventional biomedical information extraction technologies to fetch the useful and informative snippets from medical text successfully.

Due to the proprietary nature of health records and patient's private data security policies, obtaining access to health records in order to promote development of automated medical text processing systems is often difficult. Therefore, thanks to the I2B2/VA challenge (*see* Section 3.15.1) in promoting research and development of medical IR and IE systems. This chapter provides an overview on the challenges associated with mining the patient health records through author's participation in I2B2/VA challenge 2010. The work reported here presents a hybrid approach for identifying the medical concepts in patient health records. It utilizes a CRF-based supervised classifier combined with the strength of ProMiner system. For the classification of medical assertions, a Support Vector Machines (SVM)-based system was applied. Workflow details and performance assessments are discussed in the following sections.

6.1 The Fourth I2B2/VA challenge 2010

The fourth I2B2/VA challenge was a three tiered challenge that aimed at evaluation of state-of-the-art technologies for:

- Extraction of medical problems, tests, and treatments.
- Classification of assertions made on medical problems.
- Classification of relations between medical problems, tests, and treatments.

This work addresses only concept identification and assertion classification tasks. The dataset contains de-identified patient records and is composed of Partners HealthCare[1] medical records, Beth Israel Deaconess Medical Center[2] discharge summaries, University of Pittsburgh Medical Center[3] discharge summaries and progress notes. Records from the Pittsburgh Medical Center are subset of records provided by the TREC medical records track (*see* Chapter 7). Medical concepts, assertions, and relations are annotated by medical professionals. Details of corpus characteristics can be found in Section 6.2.

6.2 Goals and the Corpus Characteristics

The dataset provided by I2B2 contains a training set of 349 expert-annotated patient health records (referred to as I2B2-TRAIN) and a test set of 477 unannotated records (referred to as I2B2-TEST). The I2B2 corpus was annotated for the mentions of medical problems, treatments, and tests. Descriptions of the annotated concepts according the guidelines[4] are as follows:

Problem: Phrases that contain observations made by patients or clinicians about the patient's body or mind that are thought to be abnormal or caused by a disease. They are loosely based on the UMLS semantic type *Disorder* but not limited by UMLS coverage. Examples of sentences annotated with concepts belonging this semantic class are shown below.

> The wound was noted to be clean with mild serous drainage.
> An echocardiogram revealed a pericardial effusion and tamponade clinically.

Treatment: Phrases that describe procedures, interventions, and substances given to a patient in an effort to resolve a medical problem. They are loosely based on the UMLS semantic subtypes therapeutic or preventive procedure, medical device, steroid, pharmacologic substance, biomedical or dental material, antibiotic, clinical drug, and

[1] http://www.partners.org/
[2] http://www.bidmc.org/
[3] http://www.upmc.com/Pages/default.aspx
[4] https://www.i2b2.org/NLP/Relations/Documentation.php

	Problem	Treatment	Test
I2B2-TRAIN	11,967	8,497	7,367
I2B2-TEST	18,550	13,560	12,899

Table 6.1: Counts of annotated concepts in the I2B2 corpus.

Assertion	Example
Present	The patient experienced a drop in hematocrit.
Absent	No pneumonia was suspected.
Possible	Doctors suspect an infection of the lungs.
Conditional	Patient had increasing dyspnia on exertion.
Hypothetical	If you experience wheezing or shortness of breath
Not associated with patient	Brother has asthma

Table 6.2: Examples of sentences containing assertions on medical problems (marked in red color).

drug delivery device but not limited to UMLS coverage. Examples of sentences annotated with concepts belonging this semantic class are shown below.

The patient had a bronchoalveolar lavage performed.
After months of physical therapy, the patient gained strength.

Test: Phrases that describe procedures, panels, and measures that are done to a patient or a body fluid or sample in order to discover, rule out, or find more information about a medical problem. They are loosely based on the UMLS semantic types laboratory procedure, diagnostic procedure, but also include instances not covered by UMLS. Examples of sentences annotated with concepts belonging this semantic class are shown below.

An abdominal ultrasound was performed showing no stones.
Cardiac catheterization revealed coronary artery lesions.

The I2B2-TRAIN corpus contains 30,673 sentences, 260,573 tokens and 27,831 annotated entities. The I2B2-test corpus contains 45,053 sentences and 396,173 tokens. A later supplied gold standard for the I2B2-TEST contains 45,009 annotated entities. Table 6.1 shows the number of annotated concepts in training and test sets. The aim of concept identification task is to utilize the information from I2B2-TRAIN corpora in order to automatically tag the mentions of medical problems, treatments and tests in the I2B2-TEST corpus. The expert annotations of the I2B2-TEST (also referred to as gold standard) were made available at the end for assessing the performance of the applied system.

On the other hand, for assertion classification task, only the mentions of medical

Assertion Category	I2B2-TRAIN	I2B2-TEST
Present	8,052	13,025
Absent	2,535	3,609
Possible	535	883
Conditional	103	171
Hypothetical	651	717
Not associated with patient	92	145

Table 6.3: Counts of assertion categories in the I2B2 corpus.

problems in the I2B2 corpus corpus were categorized into six predefined categories. Table 6.2 shows the categories of assertions made over medical problems and examples of sentences containing them. Table 6.3 shows the counts of assertion categories in the I2B2 corpus. The aim of assertion classification task is to utilize the information from the I2B2-TRAIN corpus and automatically classify the mentions of medical problems in the I2B2-TEST corpus into pre-defined categories.

Unlike biomedical scientific articles, the structure of e-health records vary drastically amongst different records. A large portion of e-health records are available in narrative form as a result of transcription of dictations, direct entry by healthcare providers, or use of speech recognition applications [Meystre et al. (2008)]. Spelling errors, and contextual features such as negations and temporality is something that can be frequently found in medical text in comparison to scientific articles. The shortest record in the I2B2 dataset contains 5 sentences whereas the longest record contains 358 sentences. Few records are semi-structured whereas others are completely unstructured. A semi-structured record contains information about patient's illness, medications, and diagnoses written in structured manner whereas some other information is described in free-text natural language expression. A large portion of records have signatures of physicians, note of thanks, and many more meta-data that does not contribute to medical semantics of the reports. The varying structure of records and heterogeneously coded information makes it challenging for information extraction systems for successfully processing this form of text. An example of semi-structured part of an anonymous patient record is shown below.

Allergies :

Patient recorded as having No Known Allergies to Drugs

Chief Complaint :

Shortness of Breath

Major Surgical or Invasive Procedure :

Endotracheal Intubation

Central Venous Catheter and Swan Ganz catheter placement

Medications on Admission :

Levothyroxine 100 mcg PO daily

Labels:	**O**	**B-Prob**	**I-Prob**	**O**	**O**
Tokens:	The	rectal	cancer	was	diagnosed

Table 6.4: Example of text snippet and label sequence after tokenization and IOB conversion.

6.3 Concept Identification with CRF

Considering the previous successful application of CRF for the recognition of medical disorders in biomedical text, the system was re-trained and applied for the identification different categories of medical concepts in patient health records. The I2B2 data was tokenized at whitespaces and converted into IOB sequences before they can be subjected to training or validation. Table 6.4 shows an example of text snippet in the IOB format. The labels **B-Prob**, **B-Treat**, and **B-Test** indicate beginning tokens of the problems, treatments and tests whereas **I-Prob**, **I-Treat**, and **I-Test** correspond to intermediate tokens of problems, treatments and tests respectively. The label **O** corresponds to a token that does not belong to any entity class.

6.3.1 Feature Sets for Concept Identification

The features used for training the CRF can be broadly categorized as *morphological*, *grammatical*, *context-based* and *ProMiner-based* features. The applied feature sets are based on experiences obtained from previous work on disease and adverse effect identification described in Section 5.2.1. *Morphological* features are concerned with the internal structure of the tokens (e.g. suffixes/prefixes, capitalizations, special characters, WordAsClass, etc). *Context-based* features use information about the surrounding elements for every token [e.g. offset conjunction (OC) of order ± 1, ± 2]. *Grammatical* features are Parts-Of-Speech (POS) tags of the tokens. *ProMiner-based* features include lists of candidate named entities that occur in the complete I2B2 corpus that were recognized by the ProMiner. Three separate dictionaries were used for identifying the candidate names of problems, treatments and tests. For identifying the candidate medical problems, the MedDRA dictionary was used. A combined dictionary composed of entries from DrugBank, and KEGG was used for identifying the candidate treatments. A subset of MeSH representing diagnostic procedures was used for identifying the names of tests in the I2B2 dataset.

6.4 Assertion Classification

The principle behind this classification task was to use the contextual information in order to automatically classify the assertions of medical problems. A range of classifiers that include Naïve Bayes, Nearest Neighbor, Decision Tree, and Support Vector Machines (SVM) were preliminarily validated on the training data. Based on

the outcome of this validation, the best suited classifier was subjected to classify the instances in the test set. Weka 3.6[5] platform was used for the assertion classification task.

6.4.1 Feature Sets for Assertion Classification

During the preliminary evaluation over the training data, various feature sets were tested that include words-in-window, lemmas-in-window, bigrams-in-window, positions, family history, and nearest verbs.

- **Words-in-window** of size '±n' includes 'n' number of words that precede and succeed the mentions of medical problems.
- **Lemma-in-window** of size '±n' includes lemmas of 'n' number of words that precede and succeed the mentions of medical problems.
- **Bigrams-in-window** of size '±n' includes bigrams (also referred to as word pairs) of 'n' number of words that precede and succeed the mentions of medical problems.
- **Position** adds information to every token, lemma or bigram whether it precedes or succeeds the mention of medical problem.
- **Family history** adds information to the mention of medical problem whether it occurs in 'family history' subsection of the document or not.
- **Nearest verbs** include the verbs that precede and succeed the mentions of medical problems. The lemmatized forms of verbs were used as features. This feature is independent of the size of the window.

Words-in-window, lemmas-in-window, bigrams-in-window, positions, and family history were modeled as binary features. For every feature, its value was set to '1' if the feature was present or set to '0' if the feature was absent. Nearest verbs denoted two separate features *i.e.* left nearest verb and right nearest verb with their values set to the respective lemmatized stings. Table 6.5 shows an illustration of features associated with an arbitrary problem concept.

6.5 Performance Evaluation Criteria

The evaluation of concept identification was performed using exact match as a criterion. An exact match is a situation where the system identifies both left as well as the right boundaries of the annotated concept correctly. Performances of concept identification and assertion classification were judged based on Precision, Recall and F_1 score. During

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

Features	Values
Window size	±4
Words-in-window	was, noted, to, have, for, which, he, started
Lemma-in-window	be, note, to, have, for, which, he, start
Lemma-in-window + Position	PRE=be, PRE=note, PRE=to, PRE=have, POST=for, POST=which, POST=he, POST=start
Nearest verbs	PREVERB=have, POSTVERB=start

Table 6.5: Features associated with the concept *an erythematous perianal rash* that will be subjected to assertion classification and that is present in the sentence: *He was noted to have an erythematous perianal rash for which he started on Nystatin powder*.

the preliminary experiments, the performances of the systems were evaluated by 10-fold cross validation of the I2B2-TRAIN corpus. Finally, the best performing settings were chosen to tag or classify instances in the I2B2-TEST corpus.

6.6 Evaluation of Concept Identification

Under the preliminary settings, the CRF was trained and evaluated by 10-fold cross validation of the I2B2-TRAIN corpus. All the morphological features and context-based features (OC = ±1) were used. The system attained the F_1 score of 0.78 ± 0.04 (also referred to as baseline).

In order to evaluate the impact of different features, a new set of features were added to the preliminary feature set or the existing ones were set to idle and experimented in a systematic way. For every modified feature set, a separate model was trained and evaluated by 10-fold cross validation. The result of feature evaluation is shown in Table 6.6. Only the features that contributed to an improvement in the performance of baseline result are shown. Table 6.6 implicitly indicates that the perturbation of preliminary feature set (*i.e.* removal of WordAsClass, changing the offset conjunctions) does not contribute to the improvement of the baseline result. The POS tags improved the performance by nearly 1% whereas adding the lemmas and *ProMiner-based* features contributed substantially to the system's performance. Finally, the best model that achieved the F_1 score of 0.83 ± 0.03 was applied to tag the I2B2-TEST corpus.

6.7 Evaluation of Assertion Classification

In the first step, different classifiers were trained and evaluated by 10-fold cross validation of the I2B2- train corpus. Words-in-window of sizes ±2, ±3, ±4, ±5, ±6

Features	F_1 score
Prel. Features	0.78 ± 0.04
Prel. Features + POS	0.79 ± 0.04
Prel. Features + POS + Lemma	0.81 ± 0.05
Prel. Features + POS + Lemma + ProMiner-based	0.83 ± 0.03

Table 6.6: Results of the system's performance (F_1 score) during different stages of feature evaluation experiments. Prel. features indicate all the morphological features and OC of ±1.

Window size	NB	NN	DT	SVM
±2	0.76±0.04	0.78±0.05	0.82±0.04	0.84±0.05
±3	0.77±0.05	0.79±0.04	0.83±0.05	0.85±0.04
±4	0.78±0.05	0.79±0.05	0.83±0.03	0.86±0.03
±5	0.78±0.03	0.79±0.06	0.83±0.06	0.86±0.04
±6	0.77±0.04	0.79±0.05	0.84±0.04	0.86±0.03

Table 6.7: Performance of assertion classification (F_1 score) over the varying window sizes during 10-fold cross-validation.

were used as features in the preliminary settings. The aim was to choose one best classifier and a suitable window size for further evaluations. Table 6.7 shows the performance of different classifiers over the varying window sizes during 10-fold cross-validation.

Based on the results of classifier and window size selection, the SVM and a window size of ±4 were chosen to be optimum for further experimentation. In the second step, different sets of features were used and the performance of SVM was evaluated. For every modified feature set, a separate model was trained and evaluated by 10-fold cross validation. The results of feature evaluation for the assertion classification task are shown in Table 6.8.

The results of feature evaluation indicated that a combination of lemma-in-window, positions and family history coupled with SVM is best suited for classifying the assertions of medical problems in the I2B2-TEST corpus.

6.8 Final Evaluation over the Test Set

For identifying the concepts in I2B2-TEST corpus, a trained CRF that utilizes the best feature set observed in Table 6.6 was applied. The results of concept identification over an independent test set are shown in Table 6.9. The applied system achieved an overall F_1 score of 0.82 for tagging the problems, treatments and tests in the I2B2-TEST corpus. For classifying the assertions in I2B2-TEST corpus, a trained SVM that utilizes the best

Features	F_1 score
Words-in-window (baseline)	0.86 ± 0.03
Bigrams-in-window	0.86 ± 0.04
Lemma-in-window	0.87 ± 0.04
Lemma-in-window + Positions	0.89 ± 0.05
Lemma-in-window + Positions + Nearest verbs	0.89 ± 0.04
Lemma-in-window + Positions + Nearest verbs + Family History	0.90 ± 0.04
Lemma-in-window + Positions + Family History	0.90 ± 0.03

Table 6.8: Results of the system's performance (F_1 score) during different stages of feature evaluation experiments for the assertion classification.

Concept Category	Precision	Recall	F_1 score
Problem	0.84	0.80	0.82
Treatment	0.84	0.77	0.81
Test	0.85	0.80	0.82
Overall	0.84	0.80	0.82

Table 6.9: Assessment of performance of the system for identifying the concepts in I2B2-TEST corpus.

feature set described in Table 6.8 was applied. The results of assertion classification are shown in Table 6.10. The applied system achieved an overall F_1 score of 0.90.

6.9 Error Analysis

The concepts tagged by the CRF during training as well as in the test set were manually investigated to understand some common sources of errors. Examples of frequent sources of errors include abbreviations such as *CXR* that stands for *chest X-ray* and *IVP* that stands for *Intravenous Pyelogram*. Apart from abbreviations, the descriptive enumerations of medical problems caused substantial problems. Long and descriptive mentions of the medical problems such as *subtle decreased flow signal within the sylvian branches* were not recognized completely by the system. Other sources of errors include nested concepts such as *rupture of liver, left renal vein, pancreas, and transverse mesocolon* and anaphors such as *the following medications*. A manual inspection of the results of assertion classification indicated several errors. For example, in the sentence *It was felt that his dementing illness and rigidity was most likely due to some type of cortico-basal ganglia degeneration process, but this was not clarified during this admission*, the medical problem *cortico-basal ganglia degeneration process* that was originally annotated as possible was misclassified as present by the system. This is because the four features in the preceding window could not capture the keyword likely which is a critical feature in this scenario

Assertion Category	Precision	Recall	F_1 score
Present	0.92	0.96	0.94
Absent	0.88	0.85	0.87
Possible	0.71	0.47	0.57
Hypothetical	0.73	0.73	0.73
Conditional	0.70	0.18	0.28
Not associated with patient	0.96	0.66	0.78
Overall	0.90	0.90	0.90

Table 6.10: Assessment of performance of the system for classifying the assertions in I2B2-TEST corpus.

for a correct classification. Other examples include mentions of multiple neighboring problems that render them far from their actual context. For example, in the sentence *She currently denies any fever , chills , night sweats , weight change , blurred vision , headaches , nausea , vomiting , diarrhea , constipation , abdominal pain , changes in vision , shortness of breath , chest pain or pressure , or changes in her bowel habits*, concepts such as *vomiting*, *diarrhea*, *constipation* etc. that belong to the class hypothetical were misclassified as present since their preceding or succeeding features fail to capture the actual context. In the cases of both concept identification as well as assertion classification, the annotation errors induced by human annotators also contribute to the decline in performance of the system.

6.10 Summary on Competing Systems at I2B2 2010

"The performances of competing systems were evaluated on held-out dataset (*i.e.* I2B2-TEST). There were 22 systems competing for concept identification and 21 systems for assertion classification.

For the concept identification task, most effective systems used Conditional Random Fields whereas the only exception was Bruijn et al. (2010) who secured the top position for this task. Bruijn et al. (2010) trained an online Passive-Aggressive algorithm [Crammer et al. (2006)] based on lexicosyntactic textual features that achieved the best results for the concept identification. Our system that used CRF enriched with features from ProMiner was ranked fourth. Roberts et al. (2010) broke the concept extraction task into two steps, so that in the first step they trained CRF on identifying concept boundaries and in the second step they determined the class of the concept. Some others [Jiang et al. (2010), Kang et al. (2010)] utilized CRF in an ensemble, either of existing named entity recognition systems and chunkers, or of different algorithms with input based on knowledge-rich sources [Denny et al. (2003)]. Jonnalagadda and Gonzalez (2010) applied a semi-supervised CRF that utilized distributional semantics-based features.

Most effective assertion classification systems used Support Vector Machines either with contextual information and dictionaries that indicate negation, uncertainty, and

family history, or with the output of external rule-based systems. Roberts et al. (2010) and Chang et al. (2010) used both dictionaries and rule-based systems. Chang et al. (2010) complemented SVM with logistic regression, multi-logistic regression, and boosting, which they combined using voting. Bruijn et al. (2010) created an ensemble whose final output was determined by a multi-class SVM. Clark et al. (2010) used CRF to determine negation and uncertainty with their scope, and added sets of rules to separate documents into zones, to identify cue phrases, to scope cue phrases, and to determine phrase status. They combined the results from the found cues and the phrase status module with a maximum entropy classifier that also used concept and contextual features" [Uzuner et al. (2011)].

6.11 Discussion

This chapter addresses the challenging tasks of identifying the medical concepts in e-health records and to classify the assertions made over medical problems. These are under-addressed challenges due to the proprietary nature of e-health records as well as the unavailability of well annotated data that can support machine training or evaluation. Based the success demonstrated by previously applied approaches for the identification of concepts in scientific text (addressed in previous Chapters 4 and 5), they have been systematically readapted for the identification of concepts in medical text.

The applied strategy for medical concept identification with CRF and assertion classification with SVM achieved competitive results with F_1 scores of 0.82 and 0.90 respectively. In case of concept identification, it was shown that the application of ProMiner enabled features to CRF substantially contributes to the performance of the system. For classifying the assertions on medical problems, the window-based contextual features in combination with SVM were shown to be successful. Nevertheless, several strategies have to be tested in order to improve the performances of the applied systems. The dictionaries used for identifying the candidate named entities had a limited coverage. For example, the dictionary used for treatments had a good coverage of chemical and drug names but did not include names of operative procedures, therapies, etc. Manual curation and quality assurance of the terminological resources is a possible solution.

From an application point of view, the developed strategy allows capturing important categories of medical concepts with high specificity and sensitivity that can support the development of a semantic platform for searching in e-health records (addressed in the next chapter). Furthermore, the ability to find medical concepts and their assertions can also support automatic strategies for finding unsuspected links (between concepts or associated events) from huge volumes of medical literature. As a use case scenario, the developed systems are believed to improve medical literature searches, literature-based knowledge discovery and thus support clinical decision-making for advanced healthcare in the medical arena.

Chapter 7

Semantic Platform for Information Retrieval from E-Health Records

Electronic patient health records encompass valuable information about patient's medical problems, diagnoses, and treatments offered including their outcomes. However, a problem for medical professionals is an ability to efficiently access the information that are documented in the form of free-text. An example for the potential application of searching within e-health records is the strategy applied by the Mayo Clinic for mining the patient health records for identifying suitable subjects for clinical trails[1]. Healthcare sector worldwide has been taking strong initiatives for the development of sophisticated NLP technologies for mining the e-health records[2]. Another example is the public-private partnership EU project EHR4CR[3] that aims at providing adaptable, reusable, and scalable solutions for exploring information from e-health records for clinical research. Therefore, strategies for efficient searching and retrieval of information from e-health records is highly demanding in clinical settings. However, the goal is hard to achieve due to the proprietary nature and ethical issues involved in mining the patient data as well as extremely limited availability of publicly available datasets that can support the development and validation of medical search engines.

In order to address this issue, the Text Retrieval Conference Medical Records Track (TRECMED) 2011 provides an experimental platform for open development, evaluation, and comparison of approaches for efficient information retrieval from e-health records. Based on the successful scenarios exhibited by the previously applied strategies for identifying the concepts in scientific and medical literature (Chapters 4-6), they have been coupled with the strength of foreign NLP tools (such as SemRep, ConText *see* Chapter 3.12) for the development of a semantic platform for searching and retrieval of e-health records. The system offers facilities for keyword searches, semantic searches, and ontological searches. Workflow details and performance assessments are described in the following sections.

[1] http://www.informationweek.com/news/healthcare/EMR/231601559
[2] http://www.openehr.org/home.html
[3] http://www.ehr4cr.eu/about.cfm

Report Type	No. in Repository
Radiology Reports	47,555
History and Physical Exam Reports	15,721
Emergency Department Reports	13,424
Progress Notes	8,538
Discharge Summaries	7,931
Operative Reports	5,032
Surgical Pathology Reports	2,877
Cardiology Reports	632

Table 7.1: Types of electronic health reports present in the TRECMED dataset and their respective counts.

7.1 Task Description

The dataset used for TRECMED 2011 contains approximately 101,711 e-health records from University of Pittsburgh NLP repository[4]. The dataset is composed majorly of radiology reports constituting nearly 50% of the total dataset followed by history and physical exam reports, emergency department reports, and so forth. Table 7.1 provides the counts of different types of reports contained in the TRECMED dataset. Altogether, 35 expert-formulated questions (also referred to as *topics*, *see* Table A.2) were provided and the task was to retrieve sets of records from the collection that can best answer the topic questions. An example of topic question is *find patients with gastroesophageal reflux disease who had an upper endoscopy*. Later on, officially submitted records from different participants were pooled and a group of human evaluators with strong medical background made judgments over relevancy of the retrieved records (*i.e.* records were judged as *irrelevant*, *possibly relevant*, or *relevant* for a given question).

7.2 Data Preprocessing

The TRECMED collection contains 101,711 reports. A notion of "Visit" defines all the reports corresponding to a patient's consult to the hospital. In the current dataset, the smallest visit corresponds to one report and the largest visit corresponds to 418 reports. Mapping between the reports and visits were provided in prior[5]. An official evaluation criteria required participants to return sets of visits for different topics. The pre-processing step combined multiple reports to their representative visits without changing the semantic structure of visits. For example, if a visit contains two radiology reports and two discharge summaries, after report-to-visit merging the final visit would have one radiology report section that is a combination of two constituent radiology reports and one similarly generated discharge summary section. The report-to-visit

[4] http://nlp.dbmi.pitt.edu/nlprepository.html
[5] http://www-nlpir.nist.gov/projects/trecmed/2011/tm2011.html

merging resulted in 17,198 visits that were subjected to further processing. Each visit contains 9 free-text sections that are formed by constituent reports. The sections are complaint (COMP), radiology reports (RAD), history and physical exams (HP), emergency department reports (ER), progress notes (PGN), discharge summaries (DS), operative reports (OP), surgical pathology reports (SP), and cardiology reports (ECHO).

7.3 Patient Demography Identification

Patient demography identification task identifies patient's age and gender indicated within the visit. An age-identifier was developed that is a rule-based and regular-expression based system for the identification of de-identified age groups mentioned in visits. The system finally classifies a visit as *child*, *teen*, *adult*, or *elder*. Identifying the age within patient visits is not a trivial task since a visit may contain ages of patient's relatives such as son, father, mother, etc. Manually crafted rules were applied to filter out ages of non-patients and an evaluation of the system was internally performed that indicated superior results. Visits with ambiguous multiple age groups information were classified into multiple age groups respectively. For example, the visit *ge4U9SGxaDRw* defines the patient as *teen* and *adult*. As a result of age identification, 9185 visits were classified as *adult*, 5747 as *elder*, 581 as *teen*, 273 as *child*, and 3248 had no age information.

A gender-identifier was developed that is a rule-based and regular-expression based system for identification of patient's gender mentioned in visits. The system finally classifies a visit as *male* or *female*. The gender-identifier recognizes gender-specific nouns and pronouns such as male, female, she, her, etc. and based on the frequency of gender mentions it classifies a visit. Visits with ambiguous gender information were classified into both gender categories. As a result of gender identification, 8034 visits were categorized as *male*, 6916 as *female*, and 2248 visits had no gender information.

7.4 Concept and Relation Identification

Different tools were applied for the recognition of concepts and relations in visits. Concept and relation identification was performed on all free-text sections of visits.

MetaMap was applied for the identification of UMLS concepts in visits. UMLS contains over 100 semantic classes of concepts such as the anatomy, physiology, disorder, and many more. All classes of UMLS concepts recognized by the MetaMap were used.

SemRep (Semantic Knowledge Representation)[6] is a tool for the identification of relations in any arbitrary biomedical text. SemRep identifies relationships between UMLS concepts in text within the sentences. Types of relations that SemRep identifies is pre-defined by the UMLS. Table 7.2 shows top five types of frequently occurring relationships. Altogether, 30 different types of relationships were identified in TRECMED visit collection.

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

Type of relation	No. of occurrences
LOCATION-OF	151,225
PROCESS-OF	58,443
TREATS	26,816
IS-A	20,417
PART-OF	20,228

Table 7.2: Top five frequently occurring types of relationships in TRECMED collection.

ProMiner was used along with pre-processed dictionaries for the identification of named entities (referred to as *concepts*) in text. The dictionaries used for concept identification can be broadly categorized as dictionaries for medical problems, treatments, and diagnostic tests. Dictionaries used and the information they contain are as follows:

MedDRA provides a comprehensive terminology for medical problems such as signs, symptoms, diseases, adverse effects, syndromes, and many more. The curated version of applied MedDRA dictionary contains over 15,000 entries with nearly 55,000 synonyms.

MeSH-Disease provides a comprehensive terminology for medical problems covered by the "C" sub-hierarchy of MeSH. However, MeSH is hierarchically organized into 14 levels and provides facilities for ontological searches. The curated version of applied MeSH-Disease dictionary contains over 4,000 entries with nearly 40,000 synonyms.

DrugBank covers names and synonyms of drugs including their brand names, systemic names and registry codes. The curated version of applied DrugBank dictionary contains over 6,800 entries with nearly 64,500 synonyms.

ATC[7] provides a coverage of pharmacological, therapeutic, and chemical class names. Examples include terms such as *adrenergic antagonist*, *anti-bacterial agent*, *Prostaglandin*, etc. Synonyms of ATC terms were extracted from the UMLS. Mappings exist between ATC and DrugBank entries within the DrugBank database. Curated ATC dictionary over 650 entries with nearly 3,500 synonyms.

MeSH-Diagnostic provides a comprehensive terminology for diagnostic tests covered by the "E" sub-hierarchy of MeSH. Applied MeSH-Diagnostic dictionary over 2,500 entries with nearly 22,000 synonyms.

A CRF-based system was trained over manually annotated concepts in approximately 800 e-health records provided by the I2B2 challenge 2010 (*see* Section 6.2). The system was trained for the recognition of medical problems, treatments, and tests in e-health

[7] Anatomical Therapeutic Chemical classification system, `http://www.whocc.no/atc_ddd_index/`

Concept/Rel.	No. of occurrences	No. of unique occurrences
UMLS	9,571,099	36,747
Relations	342,712	82,833
MedDRA	1,298,729	4,605
MeSH-Disease	1,144,267	2,239
DrugBank	239,258	902
ATC	38,140	157
MeSH-Diagnostic	406,711	939
CRF-Prob	1,657,912	294,038
CRF-Treat	630,256	76,341
CRF-Test	632,404	47,836

Table 7.3: Counts of different types of concepts and relations occurring in TRECMED dataset. Total number of occurrences (column 2) and number of unique occurrences after normalization (column 3) are reported.

records (referred to as CRF-Prob, CRF-Treat, CRF-Test respectively). Concepts recognized by the CRF were morphosyntactically normalized[8]. Table 7.3 shows counts of different types of concepts and relations occurring in the TRECMED dataset.

7.4.1 Assertion Classification on Medical Problems

For classification of assertions made over medical problems, the ConText program [Harkema et al. (2009)] was used. ConText program contains three separate modules for the identification of negation, temporality, and experiencer information provided over mentions of medical problems in text. The negation module identifies any negations made over medical problems. The temporality module classifies a medical problem as *history*, *recent*, or *hypothetical*. Similarly, the experiencer module identifies if a medical problems occurs in the patient or patient's relatives (such as father, mother, son, etc.). Context program was applied to identify negations, temporalities, and experiencer information made over mentions of problems that were mapped to MedDRA, MeSH-Disease, UMLS (*Disorder* semantic-type), and, CRF-Prob. The negation and experiencer modules were applied as-is whereas the *history* and *hypothetical* rules associated with temporality module were modified. Examples of such modifications include removal of patterns such as *reported*, *complains*, and *presented* that asserts a medical problem as *history*. Similarly, modifications associated with *hypothetical* assertions include removal of patterns such as *as needed*, *come back for*, and so forth. Using the experiencer module, problem mentions were classified as *in-patent* or *not-in-patient*. Several instances exist where a medical problem can attain multiple assertions. For example, in the sentence *His father had no history of hypertension*, the medical problem *hypertension* belongs to *history*, *negation*, and *not-in-patient*. Table 7.4 shows counts of assertions made over

[8] http://www.ncbi.nlm.nih.gov/books/NBK9680/

	Negation	History	Hypothetical	Not-in-patient
UMLS	609,193	224,077	833	21,447
MedDRA	460,117	164,413	787	14,572
MeSH-Disease	377,913	149,822	749	13,497
CRF-Prob	563,682	192,375	1,029	15,341

Table 7.4: Counts of assertions made over medical problems.

medical problems identified by different concept identification approaches. Nearly 30% to 35% of medical problems recognized by different techniques are negated and this indicates the importance of negation identification in patient health records.

7.5 Indexing

Free-text fields of TRECMED visits including demographies as well as medical concepts, and relationships occurring in different sections of visits were indexed with SCAIVIEW [Hofmann-Apitius et al. (2008)]. SCAIView is a high performing and scalable Information Retrieval (IR) system based on Lucene[9]. It provides a framework for indexing several gigabytes of document data and to quickly perform complex searches over text as well as concepts. Free-text in the form of stemmed tokens appearing in different sections of patient visits were indexed. Meta-data such as ICD-9CM codes appearing in the *admit-diagnosis* and *discharge-diagnosis* fields of visits were expanded before indexing. Concepts and relations occurring in different sections of visits were indexed separately. For example, the current index allows searching for the keyword *diabetes* or the MeSH concept *Diabetes Mellitus* (MeSH-ID:D003920) in *discharge summary* (DS) sections of visits. Figure 7.1 illustrates the workflow adapted for indexing the TRECMED records. The system allows keyword searches, semantic searches, and ontological searches. For a given query, the system retrieves a ranked list of patient visits from the index.

7.6 Querying and Retrieval

Various querying strategies such as semantic search in the concept space, ontological search and text search were performed. Lucene BM25F (*see* Section 3.2.2) was applied as a scoring function to measure the similarity between visits and the query. Descriptions of different runs and the underlying query formulation strategies are discussed in the following subsections.

[9] http://lucene.apache.org/java/docs/index.html

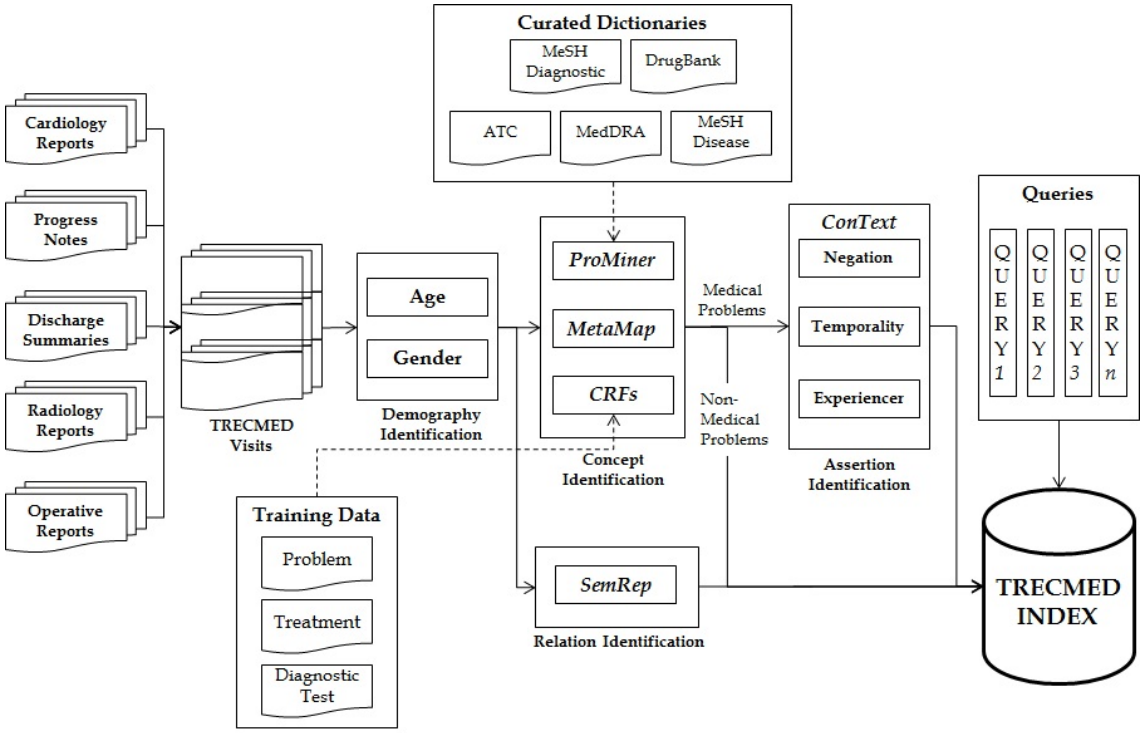

Figure 7.1: Illustration of the workflow adapted for indexing the TRECMED records.

MEDRUN1

MEDRUN1 serves as a baseline run where queries were formed by manual extraction of key terms from the topic questions. Queries were formulated in a way to reflect knowledge-based human queries. This run provides a rationale for the comparison of performances of semantic and ontological searches with knowledge-based human searches.

MEDRUN2

MEDRUN2 applies semantic search strategy to search for UMLS concepts and relations in the index. MetaMap and SemRep programs were applied for the identification of UMLS concepts and relations in topic questions. Automatically identified concepts and relations in topic questions were used for searching in the concept and relation fields of the index. Examples of SemRep found relations in the topic-116: *Patients who received methotrexate for cancer treatment while in hospital* are:
a. (C0920425) Cancer Treatment USES (C0025677) Methotrexate
b. (C0025677) Methotrexate ADMINISTERED_TO (C0030705) Patients

Information about demographies and sections to be searched were extracted from the topic questions. For example, in the topic-110: *Patients being discharged from hospital on hemodialysis*, the system would search in discharge summary (DS) sections of visits with a higher priority in comparison to rest of the sections. A higher priority was assigned to necessary sections by duplicating them in the query. In visits, the concepts referring to medical problems that are negated or those that occurs as family status were omitted during search. No difference was made when searching for problem concepts occurring as *history* or *recent* event. Nevertheless, the system allows searching for negated concepts, concepts referring to family members, and concepts indicating *history*, *hypothetical* or *recent* events.

MEDRUN3

MEDRUN3 applies semantic search strategy to search for ProMiner and CRF identified concepts in the index. ProMiner and pre-trained CRF were applied for the identification of concepts in topic questions. Automatically recognized concepts in topic questions were applied for querying in the concept space of the index. Information about demographies and sections to be searched were extracted from the topic questions. Problem concepts that are negated, historical, or indicating family status were processed as described during MEDRUN2.

MEDRUN4

MEDRUN4 applies ontological search strategy to search for ProMiner and CRF identified concepts in the index. ProMiner and pre-trained CRF were applied for the identification of concepts in topic questions. Automatically recognized concepts in topic questions were applied for querying in the concept space of the index. For the MeSH-Disease and MeSH-Diagnostic concepts, hyponyms (also referred to as child concepts) of the concepts present in topic questions were also used during querying. For example, in the topic-113, MeSH concept *Adenocarcinoma* has several hyponyms such as *Endometrioid Carcinoma*, *Hepatocellular Carcinoma*, and many more. Information about demographies and sections to be searched were extracted from the topic questions. Problem concepts that are negated, historical, or indicating family status were processed as described during MEDRUN2.

Run Combinations

Visits retrieved during two or more runs amongst MEDRUN1, MEDRUN2, MEDRUN3, and MEDRUN4 were systematically merged. If a *Visit* occurs in more than one run, its final score was computed using $\sum \frac{BM25F(Visit_i)}{Rank(Visit_i)}$ where i indicates the run.

Run-ID	bpref	R-Prec
MEDRUN1	0.4852	0.3218
MEDRUN2	0.4470	0.2909
MEDRUN3	**0.5503**	**0.3966**
MEDRUN4	0.5333	0.3774

Table 7.5: Results of retrieval during the preliminary TRECMED runs.

Run Description	bpref	R-Prec
MEDRUN1 + MEDRUN3	0.5732	0.3966
MEDRUN2 + MEDRUN3	0.5410	0.3796
MEDRUN3 + MEDRUN4	0.5517	0.3920
MEDRUN1 + MEDRUN2 + MEDRUN3	0.5658	0.3949
MEDRUN2 + MEDRUN3 + MEDRUN4	0.5487	0.3981
MEDRUN1 + MEDRUN3 + MEDRUN4	**0.5767**	**0.4088**
MEDRUN1 + MEDRUN2 + MEDRUN3 + MEDRUN4	0.5746	0.4079

Table 7.6: Performance measures of merging the retrieved visits from different runs.

7.7 Results

7.7.1 Performance Evaluation

In information retrieval, along with the relevance of the retrieved documents, the order in which they are presented is important. For example, a system that returns maximum relevant documents within top N documents is worthier than the system that returns maximum relevant documents within middle N documents. Therefore, performances of the experimented runs were evaluated using the Binary Preference score (bpref) as a primary metric and R-Precision (R-Prec) as a secondary metric (*see* Sections 3.11.4 and 3.11.5).

7.7.2 Evaluation Results

The reported results of retrieval are based on the bpref, and R-Prec scores. Table 7.5 shows the results of different individual runs and Table 7.6 shows results of run combinations. Table 7.7 shows results of the impact of age, gender, assertions, and relations on the semantic search.

Based on the observations from Table 7.5, semantic search in the concept space generated by ProMiner and CRF achieved good results with the bpref and R-Prec scores of 0.5503 and 0.3966 respectively. Results of semantic search with dictionary concepts and CRF-identified concepts considerably outperformed rest of the preliminary runs (*meaning without any post-processing*). Nevertheless, semantic searches strongly depend

Run Description	bpref	R-Prec
MEDRUN3	0.5503	0.3966
MEDRUN3 (excl. Age)	0.5505	0.3954
MEDRUN3 (excl. Gender)	0.5499	0.3934
MEDRUN3 (excl. Assertions)	0.5356	0.3793
MEDRUN3 (incl. Relations)	0.5494	0.3969

Table 7.7: Impact of age, gender, assertions, and relations on the semantic search.

on the quality of the semantic space generated. Low quality of the semantic space may hinder the performance of retrieval. An example for this instance is searching with MetaMap and SemRep identified UMLS concepts and relations that showed poor results. This indicates potential false recognitions that these systems may perform during concept or relation identification. Section 7.8 provides a detailed study on characteristics of retrieval during text, semantic, and ontological searches. Results of ontological search (MEDRUN4) performed better than manual searching but poorer than normal semantic search. One potential reason for shortcomings of ontological search is that MeSH was used as a primary hierarchy for hyponym extraction. For several MeSH concepts such as *cancer* (in topic-116) or *colonoscopy* (in topic-113), MeSH provides hundreds of hyponym concepts organized at various levels of hierarchy. It may be fuzzy for topic evaluators (coming from medical backgrounds) to accept certain hypernym/hyponym concept relations as described in MeSH.

Post-processing by merging the retrieved visits from different runs showed substantial improvement in the overall performance (*see* Table 7.6). Merging the retrieved visits of runs MEDRUN1, MEDRUN3, and MEDRUN4 which were generated by text search, semantic search, and ontological search respectively outperformed rest of the runs in terms of bpref and R-Prec scores. This indicates the successful use of applied function for merging the retrieval results obtained from different runs. Summarizing the observations from Table 7.6 indicates that coupling the retrieved visits from semantic and text searches can help in maximizing the performance of retrieval with an improved ordering of the relevant documents.

The impact of different factors such as age, gender, assertions, and relations on the semantic search was experimented (*see* Table 7.7). Excluding the age and gender information from the run MEDRUN3 resulted in slight decrease in bpref and R-Prec scores. A potential reason for the low impact of age on retrieval is that all the topic question addressing the ages of patients are associated with adults (*e.g.* Topic-114: *Adult patients discharged home with pallative care/home hospice.*) and the corpus contains over 85% visits belonging to adults (including elders as adults in Section 7.3). There are only two topic questions addressing the gender of patients (*i.e.* both focussing on female) and medical conditions associated with these questions are *breast cancer* and *osteopenia* that are more common in females than in males. Relations contributed extremely little to the R-Prec score but the bpref declined. The potential reason for

Run Description	Gain	No. Diff	Loss
MEDRUN3 & MEDRUN1	19	0	15
MEDRUN3 & MEDRUN2	24	1	9
MEDRUN3 & MEDRUN4	11	12	11

Table 7.8: Counts of topics for which *no-difference*, *gain*, and *loss* were observed by comparison of the run MEDRUN3 with runs MEDRUN1, MEDRUN2, and MEDRUN4.

decrease in performance of the system with relations is that SemRep generates potential false positives with concept recognition and therefore the associated relations applied for searching can hamper the performance of retrieval.

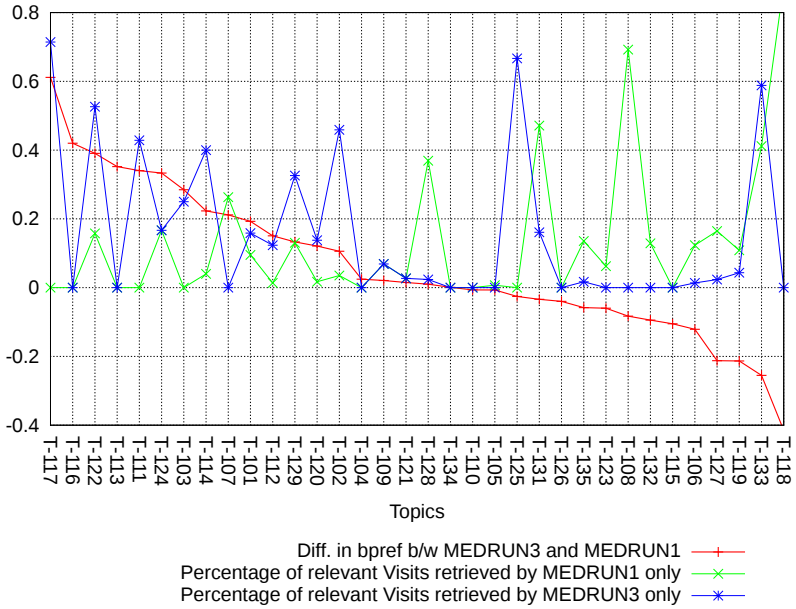

Figure 7.2: Differences in bpref scores between runs MEDRUN3 and MEDRUN1 for different topics. Percentage of unique relevant documents retrieved by both runs have been indicated.

Differences in performance between the run MEDRUN3 and remaining runs without post-processing (*i.e.* MEDRUN1, MEDRUN2 and MEDRUN4) were analyzed over different topic questions. Figure 7.2, Figure 7.3, and Figure 7.4 show analysis of differences in results between different runs. Table 7.8 shows the counts of topics for which *no-difference*, *gain*, and *loss* were observed by comparison of the run MEDRUN3 with the rest.

Table 7.8 shows that semantic search in the concept space generated by in-house NER tools (*i.e.* ProMiner & CRF) resulted in an improvement in retrieval performance over

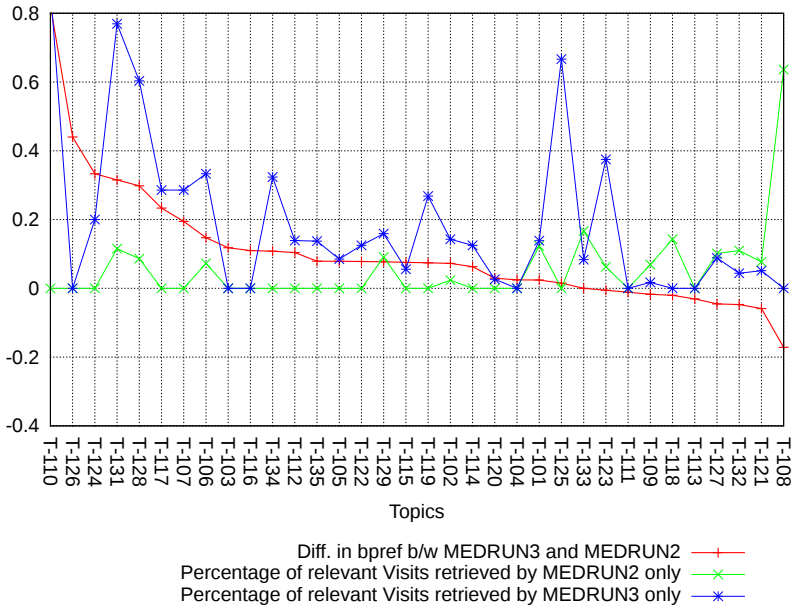

Figure 7.3: Differences in bpref scores between runs MEDRUN3 and MEDRUN2 for different topics. Percentage of unique relevant documents retrieved by both runs have been indicated.

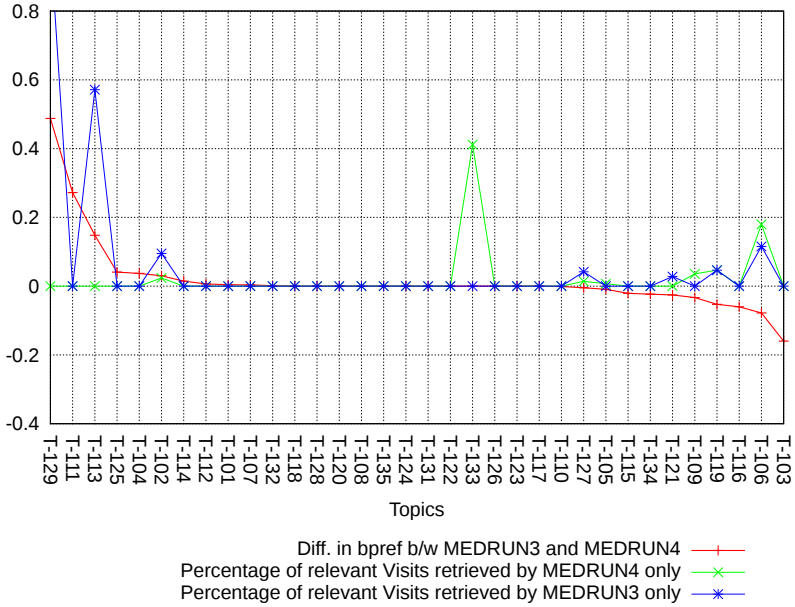

Figure 7.4: Differences in bpref scores between runs MEDRUN3 and MEDRUN4 for different topics. Percentage of unique relevant documents retrieved by both runs have been indicated.

majority of topics in comparison to searching with keywords or in the UMLS space. Although, an overall quantitative comparison showed that the semantic search can perform better than ontological search (see Table 7.5), from Figure 7.4 it was clear that semantic and ontological searches can perform competitively depending on questions of interest.

Evaluation of the retrieval performance depends on several factors and they include:

- Number of highly-relevant or relevant versus number of irrelevant or unjudged documents retrieved.
- Relative ranking of relevant and irrelevant documents.
- Relative ranking of highly relevant and relevant documents.

From Figure 7.2, it can be observed that for topics 116, 113, 104 and 134, bpref scores with semantic search (*i.e.* MEDRUN3) were better than text search (*i.e.* MEDRUN1) but both runs retrieved exactly the same relevant visits with different ranking. On contrary for topic 107, although MEDRUN1 retrieved nearly 25% more relevant visits in comparison to MEDRUN3, the bpref score for MEDRUN3 was higher than MEDRUN1. Similarly for topic 133, although MEDRUN3 retrieved nearly 20% more relevant visits than MEDRUN1, the bpref score of MEDRUN1 was relatively better than MEDRUN3. This indicates that performances of retrieval depend on ability of system to fetch the relevant documents as well as rank the relevant ones with higher priority.

Combining Text-based and Concept-based Searches

According to the observations from Table 7.5, although the overall results of semantic search (*i.e.* MEDRUN3) outperformed the results of text search (*i.e.* MEDRUN1), Figure 7.2 shows that text-based search can perform better than semantic search and deliver unique relevant documents that varies amongst different topics. Therefore, it was essential to understand the performance of retrieval by combining the text search with semantic search. In this context, two experiments were performed and they are:

- Combining text-based and concept-based queries (*i.e.* combine MEDRUN1 and MEDRUN3 queries).
- Combining the results of text-based and concept-based retrieval similar to experiments indicated in Table 7.6.

Table 7.9 shows the performance of retrieval by combining queries as well as retrieval results of MEDRUN1 and MEDRUN3. Combining the queries from text and semantic searches did not greatly contribute to the retrieval performance whereas combining the retrieval results (*referred by TXTSEM*) showed substantial improvement. A per-topic analysis was performed as shown in Figure 7.5 in order to check the performance of run TXTSEM against MEDRUN1 and MEDRUN3. Over majority of topics TXTSEM performed well in terms of bpref scores in comparison to MEDRUN1 and MEDRUN3. A paired t-test[10] [Efron (1969)] for difference in bpref between TXTSEM, MEDRUN1

[10] http://www.graphpad.com/quickcalcs/ttest1.cfm

Run Description	bpref	R-Prec
MEDRUN1 (A)	0.4852	0.3218
MEDRUN3 (B)	0.5503	0.3966
Combine A & B *Queries*	0.5531	0.3906
Combine A & B *Results* (Id: TXTSEM)	**0.5732**	**0.3966**

Table 7.9: Performance measure by combining queries and retrieval results of MEDRUN1 and MEDRUN3.

and MEDRUN3 indicated no significant difference between TXTSEM and MEDRUN3 whereas highly significant difference between TXTSEM and MEDRUN1. Both, combining the queries as well as combining the retrieval results showed improvement in the bpref and R-Prec scores. This again emphasizes on the fact that combining semantic and text searches can deliver improved retrieval when compared to individual searches.

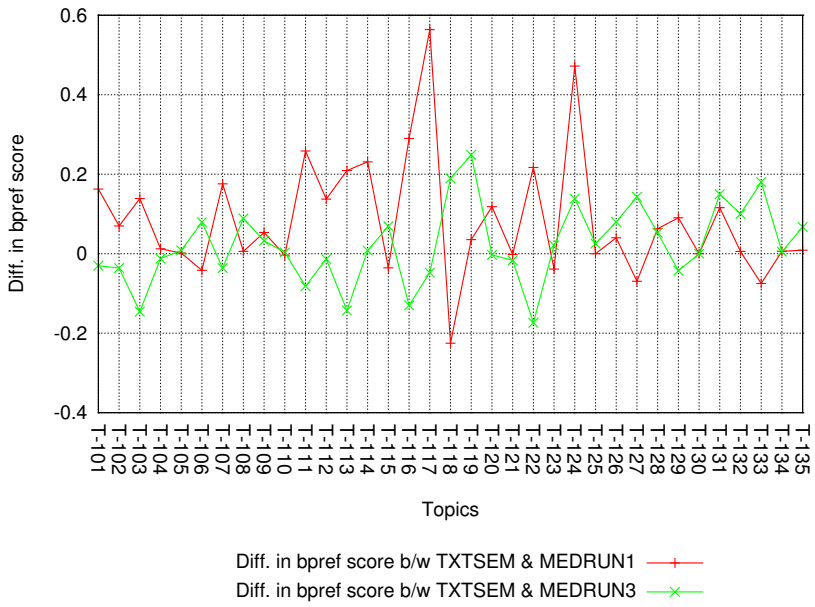

Figure 7.5: Differences in bpref scores between runs TXTSEM and MEDRUN1 as well as between TXTSEM and MEDRUN3 for different topics.

Parameter Optimization of BM25F and its Influence on Retrieval

The similarity scoring function BM25F can be tuned with two free parameters *i.e.* b and k_1 (*see* Section 3.2.2). Experiments were performed using different runs with different values of b and k_1. It was observed that altering the parameter b did not have noticeable influence on the performance of retrieval whereas altering k_1 showed changes in the

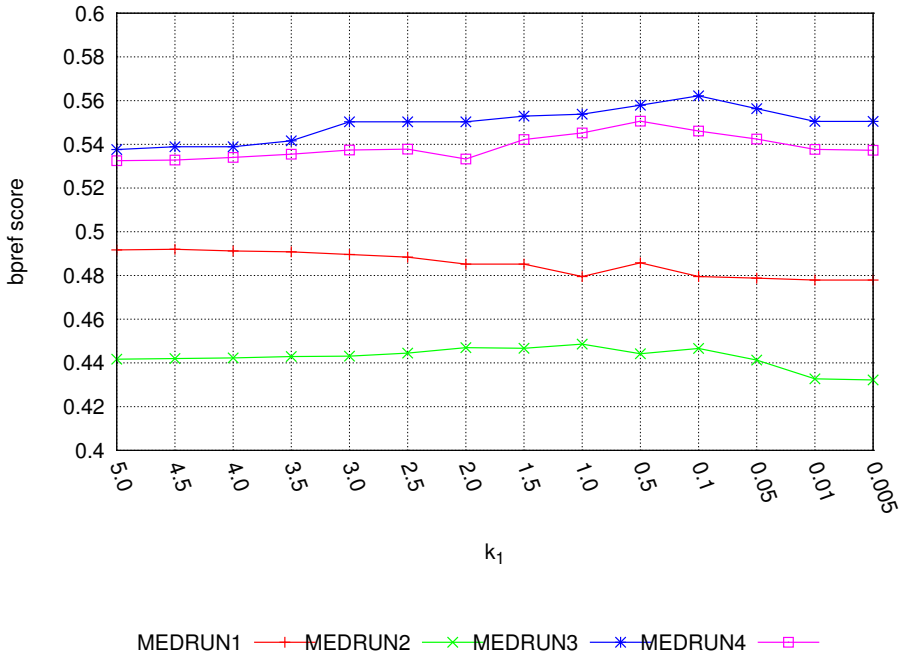

Figure 7.6: Performance of retrieval (bpref scores) for different values of k_1 (for BM25F scoring) for different TRECMED runs.

behavior of the retrieval. By default, BM25F uses k_1=2. Different values of k_1 were chosen between the values 0.005 and 5.0, and its impact on retrieval was measured that can be observed in Figure 7.6.

Table 7.10 shows the retrieval performance with the best chosen parameter k_1 for different runs. At the end of parameter optimization, the best result was obtained by MEDRUN3 with k_1=0.1 with bpref score of 0.5622 and R-Prec score of 0.4107. The best value of parameter k_1 for MEDRUN3 increased its bpref score from 0.5503 to 0.5622. A paired t-test for the difference in performance resulted in P-value of 0.5776 which indicates statistical insignificance of the observed difference. Performances of different runs varied with changes in the parameter k_1. Parameter optimization improved both bpref and R-Prec scores of all runs. Although it was not possible to establish one global maximum value of k_1 that suits different runs, observations showed that searching in concept space (MEDRUN2, MEDRUN3, and MEDRUN4) favored lower k_1 values such as 0.1 to 1.0 whereas the text search favored higher values of k_1 like 4.0 to 5.0.

Comparison of Lucene Vs BM25F

Experiments in the past have shown that Lucene similarity can perform competitively in comparison to the BM25 retrieval [Wang and Hauskrecht (2008), Lin (2009)]. Therefore, a systematic assessment of retrieval performance using different scoring functions was performed. Queries generated during MEDRUN1 to MEDRUN4 were applied

Run-ID	k_1	bpref	R-Prec
MEDRUN1	4.5	0.4920	0.3289
MEDRUN2	1.0	0.4468	0.4107
MEDRUN3	0.1	**0.5622**	**0.4107**
MEDRUN4	0.5	0.5506	0.4081

Table 7.10: Performance measures with the best chosen parameter k_1 (for BM25F scoring) for different TRECMED runs.

	BM25F		Lucene	
Run-ID	bpref	R-Prec	bpref	R-Prec
MEDRUN1	0.4852	0.3218	0.4673	0.3237
MEDRUN2	0.4470	0.2909	0.4307	0.2783
MEDRUN3	0.5503	**0.3966**	0.5521	0.3894
MEDRUN4	0.5333	0.3774	**0.5615**	0.3953

Table 7.11: Comparison of retrieval performances with Lucene and BM25F scoring.

for retrieval using the Lucene similarity scoring function[11]. Lucene uses improved version of Cosine similarity to measure the relevance between the query and documents. Table 7.11 provides a comparison of performances of retrieval using Lucene and default BM25F.

During the text search and semantic search with UMLS concepts, the performance with BM25F was comparatively better than Lucene similarity. However, Lucene exhibited successful results with the ontological search in comparison to BM25F. Systematic optimization of BM25F with the parameter k_1 (Table 7.10) performed better than Lucene and untuned BM25F scoring for majority of runs. A paired t-test for the results of runs with Lucene, BM25F and tuned BM25F indicated less significant differences in their results. Therefore, to summarize the observations from Tables 7.11 and 7.10, Lucene and BM25F can perform competitively and this varies across different search scenarios but a systematic tuning of BM25F parameters can further improve the retrieval.

7.8 Error Analysis

Retrieval results of different runs were analyzed in comparison to gold standard judgements by topic evaluators in order to understand common sources of errors. One potential reason for shortcomings of retrieval performance during the run MEDRUN2 was false positive concept identification by the MetaMap or SemRep programs. An example is *Topic 107: Patients with ductal carcinoma insitu (DCIS)*, where MetaMap

[11] `http://lucene.apache.org/java/2_9_0/api/all/org/apache/lucene/search/Similarity.html`

identified several occurrences of *DCI* in documents (that designates a place) as ductal carcinoma. MEDRUN3 utilized the concepts identified by ProMiner with acronym disambiguation strategy that helped in overcoming various false positive concept recognition that can hamper the performance of retrieval.

The author was able to identify cases where semantic search retrieved documents that were judged as *irrelevant* although they contained relevant information. An example is *Topic 117: Patients with Post-traumatic stress disorder*. The MEDRUN3 run retrieved the visit */6RlgeNinbY+* as one amongst the top 10 visits. This visit was judged as *irrelevant* but a manual investigation of the visit revealed the evidence that the patient had post-traumatic stress disorder. This is exemplified by the statements *The patient no longer works. He was trapped in a house fire several years ago and was extensively burned. He has post-traumatic stress disorder. He has been treated for depression*. Another example is *Topic 101: Patients with hearing loss*. The run MEDRUN3 retrieved the visit *D3PsCRkoq+R8* as one amongst the top 10 visits. This visit was judged as *irrelevant* by topic evaluators. Whereas a manual investigation revealed the evidence that the patient had hearing loss. This is exemplified by statements *Extremities: Negative for clubbing or edema. Skin: No rashes, nodules, or lesions. Neurological: He is awake and alert. His visual fields are intact. He has severe hearing loss, but is otherwise nonfocal*. Such evidences indicate either a non-expert evaluation, or extreme hard cases of judgement for experts.

From Figure 7.2, it can be observed that semantic search failed in several cases compared to text search. The best example for this scenario is topic-118 in Figure 7.2. For topic-118: *Adults who received coronary stent during admission*, the text search retrieved nearly 80% relevant visits that were not retrieved by semantic search. The reason was during MEDRUN3, searching in the concept space was performed using the concept designating *coronary stent*, *coronary artery stent*, and so forth that did not successfully retrieve many relevant visits. A lot of visits mentioned coronary stents administered to patients that were explained descriptively. Examples include an evidence within the visit *kwFRWomsN1Ly: Stenting at two sites of the vien graft of the right coronary artery and mid posterior descending artery with 2.5 mm drug-eluting stent*. Another example of this case was found in visit *r3FTkt2ecEdg: stent placed in the first obtuse marginal branch of the circumflex coronary artery*. These are few examples of relevant visits that were retrieved by MEDRUN1 (text search) and not retrieved by semantic search. This exemplifies some limitations associated with semantic search when the coverage of semantic concept space is not very comprehensive.

MEDRUN4 that uses ontological search performed competitively in comparison to MEDRUN3. Although, the overall results of MEDRUN3 is better than MEDRUN4, Table 7.8 indicates 11 topics where ontological search performed poorer than semantic search. As mentioned perviously, MeSH was used as resource for ontology expansion and this may conflict with topic evaluator's *hypernym-hyponym* acceptability for evaluation. An example is *Topic 116: Patients who received methotrexate for cancer treatment while in hospital*. MEDRUN4 retrieved *PDfRzvZE904q* as one amongst the best 10 retrieved visits. This visit was judged as *irrelevant* by the respective topic evaluator. Upon manual investigation, this visit revealed an evidence that the patient suffered from *T-cell lymphoma* and the patient was administered *high dose methotrexate therapy* while in

the hospital. T-cell lymphoma is a subtype of cancer that was likely to be not addressed during topic evaluation.

7.9 Summary on Competing systems at TRECMED 2011

TRECMED received a total of 127 runs from 29 participants where 109 were automatic runs and 18 were manual runs [Voorhees and Tong (2011)]. Amongst them the best performing systems have been described. King et al. (2011), from both the queries and the medical reports, extracted limiting attributes, such as age, race, and gender, and labeled terms appearing in the UMLS. They also used three different techniques of query expansion *i.e.* UMLS related terms, terms from a network built from UMLS, and terms from their in-house medical reference encyclopedias. They applied pseudorelevence feedback strategy for query enrichment that gave significant improvement in the system's performance and one of their run was ranked as best performing automatic run amongst all submissions. Demner-Fushman et al. (2011) used two search engines used for retrieval *i.e.* Essie and Lucene. In addition to the UMLS synonymy-based query expansion built within Essie and an externally implemented Lucene, they expanded the terms in the documents with their ancestors and children from the MeSH hierarchy. They also expanded query terms for recognized drug names using RxNorm and Google searches. Manual queries submitted to Essie significantly outperformed all the other manual runs submitted for this task. Mayo clinic [Wu et al. (2011)] used cTakes as well as Aho-Corasick dictionary lookup for finding UMLS concepts in reports. Similarly, Aho-Corasick dictionary matching was used to find concepts in query topics. In addition, they used the semantic structure of UMLS to find hyponym concepts for query expansion. The results showed that retrieval based on Aho-Corasick dictionary lookup was better than cTakes based retrieval. Karimi et al. (2011) applied a set of manually constructed patterns to map query terms into query language. Query expansion was performed using UMLS and Dbpedia. Best results were achieved using query transformation *i.e.* breaking the query into different components and mapping these to their uniform representation as used in the documents and query expansion using external resources. The system of Goodwin et al. (2011) builds a query by extracting keywords from a given topic using a Wikipedia-based approach. They used regular expressions to extract age, gender, and negation requirements. Each query was then expanded by relying on UMLS, SNOMED, Wikipedia, and PubMed co-occurrence data for retrieval. Runs were submitted based on Lucene with varying scoring methods, and based on a hybrid approach with varying negation detection techniques.

7.10 Discussion

This chapter reports on the development of a semantic framework for information retrieval from e-health records which has been one of the most challenging issues in

the modern medical informatics domain. Indexing the medical concepts and relations allows semantic searches and ontological searches in the concept space. The system also provides facilities to search for inter-related medical concepts. In addition, the performance of system with different search strategies has been systematically evaluated. Semantic search in the concept space indicated superior results in comparison to the conventional manual text search. During the preliminary experiments, the results of concept-based search outperformed rest of the runs with best bpref score of 0.5503. A strategic combination of results obtained from text search, semantic search, and ontological search yielded the highest scoring bpref score of 0.5767. It was also shown that combining the retrieval results of semantic search and text search can yield improved results in comparison to individual searches.

In the TRECMED scenario, the performance of retrieval has been tested over 35 topic questions. In the future, it is necessary to evaluate the system using more questions with medical expert's evaluation. This minimizes the deviation of results from standard average and gives a better estimation of system's actual performance. The system with comprehensively indexed medical relationships may substantially enhance the search performance. Finally, the developed system is believed to help domain experts and medical professionals to carry out patient record searches more efficiently. This promotes evidence-based medicine and therefore improves the overall quality of patient care and safety.

Apart from e-health records, a lot of information about medical practices and their impact on patient's health are published in forms of scientific articles, webpages, and patents. Considering a real world scenario, ability to retrieve information from various document sources is essential to capture the valuable and novel information that are distributed discretely in the biomedical bibliome beyond the clinical paradigm. Therefore, an adaptation of the developed semantic platform for searching and information retrieval from various forms of free-text data is an interesting issue. An application of the developed retrieval platform to perform semantic searches in biomedical patents is addressed in the upcoming chapters.

Chapter 8

Technology Survey in Patents

Technology survey search deals with querying and retrieval of patents or full-text documents in order to uncover any knowledge that can answer a scientific question. In the medical and healthcare arena, technology survey search in patents and full-text articles helps in understanding the state-of-the-art scientific advancements and uncovers the knowledge needs required for medical and pharmaceutical decision support as well as improved public health. Systematic information mining in patents can promote the understanding of secondary uses of current inventions or the influence of old ideas on current technological developments. Patent space denote an important source of scientific information for the applied science and the ability to retrieve relevant patents (or to automatically mine and extract relevant information from patents) is a challenge of at least the same dimension as MEDLINE mining. Searching in patents can pose more challenges in comparison to searching in MEDLINE or e-health records because of *(a)* Huge length of patents and sheer amount of information covered within a single patent (*e.g.* The patent application US20070224201A for *Compositions And Methods For The Diagnosis And Treatment Of Tumor* has 7,154 pages) *(b)* Information within the patents are heterogeneously distributed amongst the *Description* and *Claims* sections *(c)* Plenty of information are embedded as non-free-text (*i.e.* images, tables, etc.). Challenges associated with mining in patents and full-text articles are discussed by Müller et al. (2010).

Technology survey search can be addressed as an information retrieval or information extraction problem. In the biomedical and chemistry domains, this is not a trivial task due to an existence of various denominations of biomedical and chemistry concepts in free-text. Furthermore, the goal to support development or validation of efficient patent search engine is hard to achieve due to the proprietary nature of patents as well as extremely limited availability of annotated patent data. Therefore, the TRECCHEM addresses this challenge in terms of a trier namely Technology Survey (TS) task. The TS task provides a set of expert-defined natural language questions of information needs (also known as TS topics) for retrieving sets of documents from a predefined collection that can best answer those questions. This chapter focusses on the customization of the previously developed semantic search platform (*see* Chapter 7) for searching and information retrieval from biomedical patents using the TRECCHEM patent collection.

8.1 Task Description

The data used for the Technology Survey task contains approximately 1.3 million patents from the European, US, and WIPO patent offices. A subset of approx. 130,000 full-text articles that were a part of TRECCHEM TS task were excluded from the current experimental corpus. 9 topics (*see* Table A.1) that were formulated by human experts as a natural language narratives were collected from 2009 and 2010 TRECCHEM TS topics. The task is to retrieve sets of patents from the collection that can best answer the topic questions. An example of TS topic is as follows:

Topic: *TS-29*
Title: *Inhibitors for acetylcholinesterase*
Narrative:*Acetylcholinesterase inhibitor is a potential target for Alzheimer's disease so identifying potent inhibitors of this human enzyme may lead to new treatments of this devastating disease.*
Chemicals: *Acetylcholinesterase inhibitors*
Conditions: *Alzheimer's disease*

Every TS topic contains a title, a narrative text of the information needed, and a separate indication of chemicals or conditions that the topic is looking for. For a given TS topic, a retrieved patent was manually evaluated as *highly relevant*, *relevant*, *irrelevant*, *unjudged*, or *unsure*.

8.2 Data Preprocessing

The TREC collection was provided in the Extensible Markup Language (XML). As a preliminary step, an analysis of different sections or zones within the patents was performed. Patent documents contain several fields that are presumably not necessary during retrieval and generate substantial noise while processing the documents. Examples of such fields are country, legal-status, non-English abstracts, etc. The aim was to use only those fields that have high text/noise ratio and that encompass rich information content. Therefore, from a retrieval point of view, the following fields were chosen to be used for indexing and further assessments: UCID[1], Publication date, Authors, Citations, IPC[2] class, Title, Abstract, Description, and Claims.

8.3 Concept Identification in TS Topics

For the identification of concepts in TS topics, the MetaMap program was used. MetaMap was strictly applied to title, chemical, and condition sections of all TS topics. Although the UMLS is a comprehensive terminological resource containing

[1]User Reference Identifier
[2]International Patent Classification, `http://www.wipo.int/classifications/ipc/en/`

over 2 million concepts, it has been shown to lack several biomedical concepts. Elements in TS topics that could not be mapped to UMLS such as *DNA-based asymmetric catalysis*, *asymmetric catalysis*, *hydrophobic amino acid*, and *endogenous phospholipid* were used as-is and stored for further processing. Constraints were applied on the MetaMap to restrict the semantic classes of mapped concepts to *chemicals and drugs*, *physiology*, and *disorders*. A threshold of 950 was applied for the confidence score of mapping in order to be accepted as a valid concept. During the concept mapping process, the MetaMap also indicates the source vocabularies from which concepts are derived from. Therefore, if a concept exists in the MeSH hierarchy, its hyponym concepts (also known as *child concepts*) and their synonyms were extracted from MeSH. For example, the concept *Bacterial Infection* that appears in TS-28 co-exists in UMLS and MeSH. Since MeSH is hierarchically organized, it provides different hyponyms of *bacterial infections*. Figure 8.1 shows an illustration of hierarchical structure of MeSH from which the hyponym concepts were extracted.

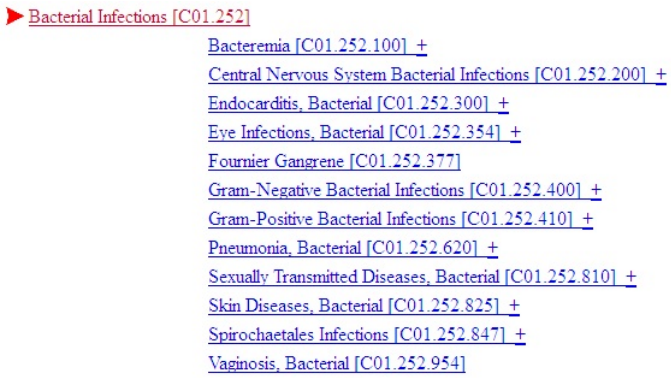

Figure 8.1: An example of hyponyms of a concept *Bacterial Infection* in MeSH.

8.4 Concept Tagging in TREC Collection

Concepts obtained from TS topics and their hyponyms and synonyms were used to generate a dictionary of TS concepts. The dictionary contains 21 concepts obtained from 9 TS topics where 17 concepts were generated from automatic mapping and the remaining four concepts were extracted from topic annotations (*e.g.* the field *chemicals* of TS Topics). ProMiner was applied for tagging the TS concepts in the patent collection. In patents, the *title*, *abstract*, *description*, and *claims* sections were tagged by ProMiner.

8.5 Document Indexing

SCAIView was used for document indexing and retrieval. Free-text in the form of stemmed tokens appearing in *title*, *abstract*, *claims*, and *description* sections of patents

	TA	Claims	Description	Document
No. of concepts	34,999	81,022	898,112	1,014,138
No. of documents	24,477	27,585	198,759	205,772

Table 8.1: Counts of number of concepts occurring in patent sections, and counts of documents containing at least one TS concept. TA indicates *title* and *abstract* combined.

were indexed. Meta-data such as *publication date*, *assignee*, etc. were indexed as-is. Concepts occurring in *title*, and *abstract* of patents were merged and indexed as a separate field (referred to as CONCEPT-TA). Concepts appearing in *description*, and *claims* of patents were separately indexed (referred to as CONCEPT-DESC and CONCEPT-CLM respectively). Concepts appearing in *title*, *abstract*, *claims*, and *description* were merged and indexed as CONCEPT-DOC. Counts of concept occurrences and documents containing at least one TS concept are shown in Table 8.1.

8.6 Query and Retrieval

Various querying strategies such as semantic search in the concept space, and text search were performed in different sections of patents. Lucene BM25F (*see* Section 3.2.2) was applied as a scoring function to measure the similarity between documents and the query. Descriptions of different runs and the underlying query formulation strategies are discussed in the following subsections.

TSRUNS1

TSRUNS1 denote a set of runs where queries were formed by manual extraction of key terms from the TS topics. Queries were formulated in a way to reflect knowledge-based human queries. These runs provide a rationale for the comparison of performances of semantic searches with knowledge-based human searches. Searches were performed in various sections of patents.

TSRUNS2

TSRUNS2 denote a set of runs where the strategy of semantic search was applied for searching in the concept space of patents. Concepts extracted from TS topics were searched against concepts indexed in various sections of patents.

Run Combinations

Documents retrieved during best four runs amongst TSRUNS1 and TSRUNS2 were systematically merged. If a *Document* occurs in more than one run, its final score was

computed using $\sum \frac{BM25F(Document_i)}{Rank(Document_i)}$ where i indicates the run.

Co-citation based Ranking

Documents retrieved during best six runs amongst TSRUNS1 and TSRUNS2 were subjected to co-citation based document ranking (*see* Section 9.6.1). Based on previous experiences from TRECCHEM, exploiting the information about patent citations can drastically improve the retrieval outcome [Gobeill et al. (2009)]. During co-citation analysis, queries were applied as-is whereas only the retrieved documents were subjected to post-processing. Therefore, the outcome of searches were systematically enriched with the patent citation information based on the co-citation ranking scheme defined in Section 9.6.1.

Impact of IPC Classification

Patent documents are assigned with International Patent Classification (IPC) codes that classify them into pre-defined categories according to the subject of claims. Previous experiences have shown that the application of IPC can improve the patent retrieval [Gurulingappa et al. (2009)] substantially. Therefore, a systematic analysis of the impact of IPC on the patent retrieval was performed. IPCCAT[3], a publicly available tool for the prediction of IPC classes for any input arbitrary text was applied to determine the potential IPC classes of all TS topics. Four code IPC classes were predicted using the IPCCAT tool. *Title* and *Narrative* sections of TS topics were used for the IPC prediction. Since all the chosen topics belong to the medical and pharmaceutical domains, most of the predicted IPC classes belonged to A61K (*PREPARATIONS FOR MEDICAL, DENTAL, OR TOILET PURPOSES*) and A61P (*SPECIFIC THERAPEUTIC ACTIVITY OF CHEMICAL COMPOUNDS OR MEDICINAL PREPARATIONS*).

8.7 Results and Discussion

8.7.1 Performance Evaluation

In information retrieval, along with the relevance of the retrieved documents, the order in which they are presented is important. For example, a system that returns maximum relevant documents within top N documents is worthier than the system that returns maximum relevant documents within middle N documents. Therefore, performances of the experimented runs were evaluated using the Binary Preference score (bpref) as a primary metric and R-Precision (R-Prec) as a secondary metric.

[3] https://www3.wipo.int/ipccat/

Run ID	Section(s) searched	bpref	R-Prec
TSRUN1-TA	Title, Abstract	0.1883	0.0763
TSRUN1-TAC	Title, Abstract, Claims	0.2377	0.1087
TSRUN1-CLM	Claims	0.2217	0.0867
TSRUN1-DES	Description	0.2508	0.1404
TSRUN1-DOC	Full-Document	**0.2772**	**0.1449**

Table 8.2: Results of text-based searches (TSRUNS1) across various sections of patents.

Run ID	Section(s) searched	bpref	R-Prec
TSRUN2-TA	Title, Abstract	0.1328	0.1340
TSRUN2-TAC	Title, Abstract, Claims	0.1908	0.1756
TSRUN2-CLM	Claims	0.1844	0.1705
TSRUN2-DES	Description	0.2700	0.2051
TSRUN2-DOC	Full-Document	**0.3208**	**0.2516**

Table 8.3: Results of concept-based searches (TSRUNS2) across various sections of patents.

8.7.2 Results of the TS Task

The reported results of retrieval are based on the bpref, and R-Prec scores. Table 8.2 shows the results of text-based searches (*i.e.* TSRUNS1) across the different sections of patents. Similarly, Table 8.3 shows the results of concept-based searches (*i.e.* TSRUNS2) across the different sections of patents.

Observations from Table 8.2 and Table 8.3 show that full-document searches in patents perform better than searching in individual sections. Document searches showed significantly higher bpref and R-Prec scores during the text search as well as the semantic search. On comparison of performances of text search with the semantic search, it was observed that semantic searches are high precision searches (defined by R-Prec scores). Full-document text search resulted in bpref and R-Prec scores of 0.2772 and 0.1449 respectively whereas the semantic search indicated bpref and R-Prec scores of 0.3208 and 0.2516 respectively. Figure 8.2 shows per-topic bpref scores for full-document search with text (TSRUN1-DOC) and semantic concepts (TSRUN2-DOC). It was observed that bpref scores with semantic search was greater for 5 topics whereas text search indicated better bpref scores for 3 topics. For topic TS-20, there was no significant difference in performance observed between text and semantic searches. Section 8.8 provides a study on behavior of patent retrieval with text and semantic concepts.

Table 8.4 shows performances as a result of systematic merging of retrieved documents from various runs (*only best four preliminary runs in terms of bpref scores were chosen*). Merging the retrieved documents from the full-document text searches and

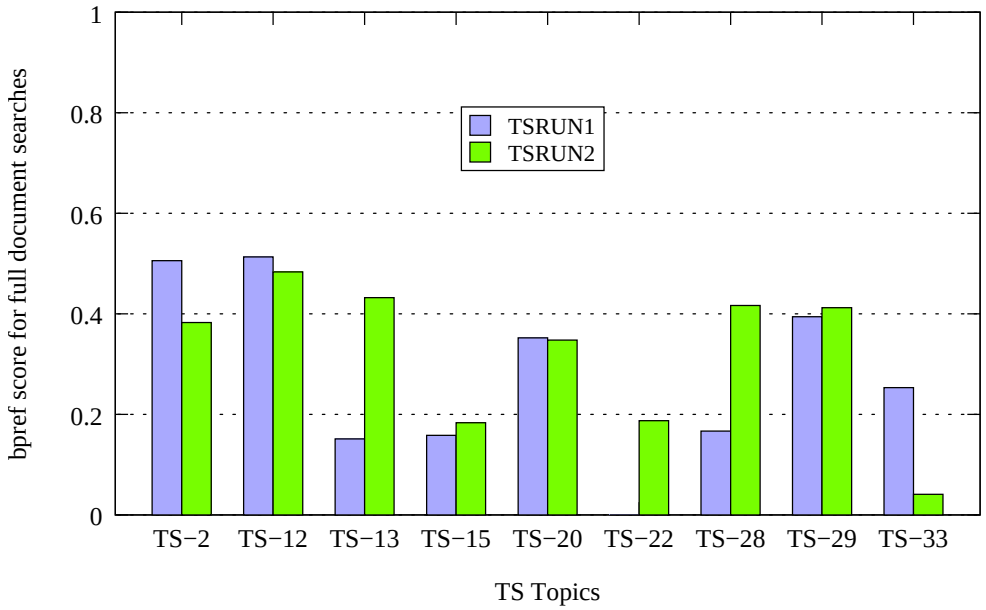

Figure 8.2: Results of full document searches for different TS topics.

Run ID	bpref	R-Prec
TSRUN1-DOC + TSRUN2-DOC	**0.3767**	**0.2356**
TSRUN2-DOC + TSRUN2-DES	0.3181	0.2216
TSRUN2-DOC + TSRUN1-DES	0.3627	0.2155
TSRUN2-DOC + TSRUN1-DOC + TSRUN2-DES	0.3624	0.2127
TSRUN2-DOC + TSRUN1-DOC + TSRUN1-DES	0.3715	0.2346
TSRUN1-DOC + TSRUN2-DOC + TSRUN1-DES + TSRUN2-DES	0.3727	0.2310

Table 8.4: Performance measures of merging retrieved documents from different runs.

semantic searches showed significant improvement in the bpref scores in comparison to individual searches. Merging the documents retrieved during full-document searches with description searches did not contribute to an improvement in the performance. Therefore, summarizing the observations from Table 8.4 shows that coupling full-document text search with semantic search can outperform individual searches or section-based searches, and therefore can yield higher bpref with an improved ordering of the relevant documents.

Impact of Co-Citation based Post-Processing on Retrieval

Table 8.5 shows the results of different runs after co-citation based document ranking. It was observed that systematic enrichment of retrieved documents with citation information thoroughly hampered the performance of retrieval in terms of bpref as well as R-Prec scores. Performances of both text searches as well as semantic searches

Run ID	bpref	R-Prec
TSRUN1-TAC	0.2133	0.0855
TSRUN1-DES	0.2439	0.1157
TSRUN1-DOC	0.2472	0.1152
TSRUN2-TAC	0.1634	0.1235
TSRUN2-DES	0.2538	0.1544
TSRUN2-DOC	0.2828	0.1672

Table 8.5: Performance measures after co-citation based post-processing of different TS runs.

Run ID	bpref	R-Prec
TSRUN1-TAC	0.2377	0.1082
TSRUN1-DES	0.2567	0.1429
TSRUN1-DOC	0.2841	0.1513
TSRUN2-TAC	0.1908	0.1756
TSRUN2-DES	0.2739	0.2141
TSRUN2-DOC	0.3263	0.2584

Table 8.6: Performance measures of the impact of IPC on patent searches. The impact on text searches (TSRUNS1) and semantic searches (TSRUNS2) has been shown.

declined with co-citation based post-processing. For instance, during the run TSRUN2-DOC which indicates sematic searches in full-documents, the bpref of the system dropped from 0.3208 to 0.2828 due to the co-citation based post-processing.

Impact of IPC Classes on Retrieval

Table 8.6 shows the impact of application of IPC classes during text and semantic searches in patents. Impact of IPC on full-document as well as section-based searches was analyzed. Using the IPC during searching improved both bpref and R-Prec scores (*i.e.* ordering of the relevant documents). IPC usage proved to improve the retrieval results during text and semantic searches as well as during full-document and section-based searches. For instance, during the run TSRUN2-DOC, the bpref score improved from 0.3208 to 0.3263 with the application of IPC during the semantic search in full documents. Although this improvement in performance is statistically insignificant, the application of IPC did not hurt the overall retrieval.

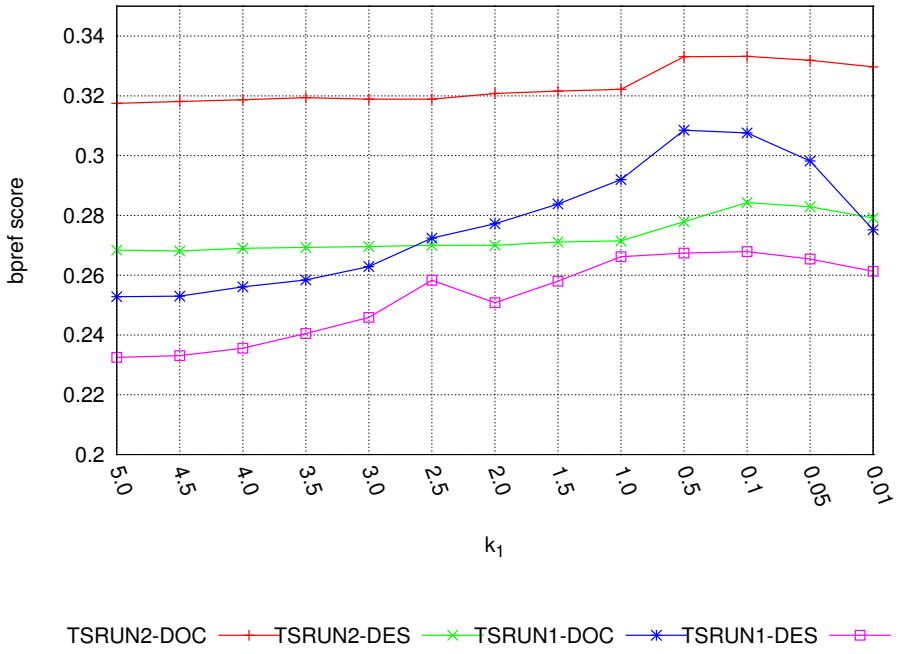

Figure 8.3: Performance of retrieval (bpref scores) for different values of k_1 (for BM25F scoring) for different TS runs.

Parameter Optimization of BM25F and its Influence on Patent Retrieval

The similarity scoring function BM25F can be tuned with two free parameters *i.e.* b and k_1 (*see* Section 3.2.2). Experiments were performed using the four best preliminary runs (*i.e.* TSRUN1-DOC, TSRUN2-DOC, TSRUN1-DESC, and TSRUN2-DESC) with different values of b and k_1. It was observed that altering the parameter b did not have any influence on the performance of retrieval whereas altering k_1 showed changes in the behavior of the retrieval. By default, BM25F uses k_1=2. Different values of k_1 were chosen between the values 0.005 and 5.0, and its impact on retrieval was measured that can be observed in Figure 8.3.

Table 8.7 shows the retrieval performance with the best chosen parameter k_1 for different runs. At the end of parameter optimization, the best result was obtained by TSRUN2-DOC with k_1=0.1 with the bpref score of 0.3332. The best value of parameter k_1 for TSRUN2-DOC increased its bpref score from 0.3208 to 0.3332, and similarly for run TSRUN1-DOC, its bpref score improved from 0.2772 to 0.3085. A paired t-test for the differences in performances after parameter optimization of runs TSRUN2-DOC and TSRUN1-DOC resulted in P-values of 0.1945 and 0.0693 respectively both indicating statistically low significance in observed differences. Performances of different runs varied with changes in the parameter k_1. Parameter optimization improved bpref scores of all the runs whereas the R-Prec scores declined for semantic searches. Although it was not possible to establish one global maximum value of k_1 that suits different

Run-ID	k_1	bpref	R-Prec
TSRUN1-DES	0.1	0.2674	0.1533
TSRUN1-DOC	0.5	0.3085	0.1507
TSRUN2-DES	0.1	0.2843	0.1794
TSRUN2-DOC	0.1	**0.3332**	**0.2291**

Table 8.7: Performance measures with the best chosen parameter k_1 (for BM25F scoring) for different TS runs.

	BM25F		Lucene	
Run-ID	bpref	R-Prec	bpref	R-Prec
RUN1TAC	0.2377	0.1087	0.2033	0.0968
RUN1DES	0.2508	0.1404	0.2928	0.1776
RUN1DOC	0.2772	0.1449	0.2912	0.1769
RUN2TAC	0.1908	0.1756	0.1971	0.1843
RUN2DES	0.2700	0.2051	0.2741	0.1858
RUN2DOC	**0.3208**	**0.2516**	0.3131	0.1882

Table 8.8: Comparison of retrieval performances with Lucene and BM25F scoring for different TS runs.

runs, observations showed that both searching with text and concepts favored lower k_1 values such as 0.1 to 0.5. Additional observations from Figure 8.3 show that tuning the parameter k_1 has higher impact on text search than on semantic search.

Comparison of Lucene Vs BM25F for TS Task

A systematic assessment of the retrieval performance using Lucene and BM25F functions was performed. Queries generated during the best 6 preliminary runs were applied for retrieval using Lucene similarity scoring function[4]. Lucene uses improved version of Cosine similarity to measure the relevance between the query and documents. Table 8.8 shows comparison of performances of retrieval using Lucene and default BM25F.

Observation from Table 8.8 shows that Lucene and BM25F perform competitively in different scenarios. Nevertheless, the default BM25F scoring for semantic full-document search indicated the best result with bpref and R-Prec scores of 0.3208 and 0.2516 respectively. Comparison of results from Tables 8.7 and 8.8 indicates that a systematic optimization of BM25F parameters can outperform the results with Lucene and untuned BM25F in terms of bpref scores, however with a marginal drop in R-Prec scores.

[4] http://lucene.apache.org/java/2_9_0/api/all/org/apache/lucene/search/Similarity.html

8.8 Error Analysis

Documents retrieved during different runs were analyzed in comparison to gold standard judgements by topic evaluators in order to understand common sources of errors. Observation from Tables 8.2 and 8.3 show that both semantic and text searches in title, abstract, and claims sections of patents performed poorer in comparison to searching in descriptions or full-documents. Investigation using TS-2: *Dipeptidyl peptidase-4 inhibitors for the treatment of type-2-diabetes* showed that in many patents the necessary information was distributed across claims and description sections. For example, the patent US-20050147662 was a relevant patent retrieved during description or full-document searches but not during the title, abstract, or claims search. The reason was because the claims of patent US-20050147662 provide only the information that the patented substance is a *Dipeptidyl peptidase-4 inhibitor* exemplified by the claim *The composition of claim 7, further comprising a dipeptidyl peptidase IV (DPP-4) inhibitor*. Upon investigating the description section, it revealed the fact that the given substance can be used for the treatment of *type-2 diabetes* described as *Physiological Effect: Reduction of amyloid deposits and systemic amyloidoisis often (but not always) in connection with Alzheimer's disease, Type II diabetes, and other amyloid-based disease*.

From Figure 8.2, it is evident that although semantic full-document search resulted in higher bpref and R-Prec scores, for few topics the text-based full-document searches performed better. An investigation on the reasons for shortcomings of the semantic search revealed the descriptiveness of the relevant information that could not be covered within the concept space (*i.e.* TS concept dictionary). For example, the topic TS-29 required finding the documents related to *Acetylcholinesterase inhibitors for the treatment of alzheimer's disease*. For TS-29, the patent US-6436937 was a relevant document retrieved during the text search but not during the semantic search. Upon investigating the reason for failure of the semantic search to fetch US-6436937 revealed that this patent addresses the *use of desoxypeganine in the treatment of alzheimer's dementia*. The necessary (relevant) information about pharmacological action of the patented substance was mentioned descriptively like *Deoxypeganine in fact inhibits not only acetylcholinesterase and thus the degradation of acetylcholine, but also monoamine oxidase and thus the degradation of dopamine*. Whereas, the whole document did not contain the relevant pharmacological information in terms of terminologies that is covered by MeSH or UMLS (such as acetylcholinesterase inhibitor, or acetylcholinesterase antagonist). This exemplifies one situation where text searches can be more beneficial.

Investigation of low ranked relevant retrieved documents during various searches revealed judgement errors performed by topic evaluators. For instance with TS-29, the patent WO-1999009999 was retrieved with low relevancy whereas it was judged as *relevant* by the respective topic evaluator. Careful investigation of the patent revealed that the subject of the document addressed *Saccharide composition for the treatment of alzheimer's disease* that has a pharmacological action of *inhibition of amyloid beta proteins*. The patent was wrongly judged as relevant although it does not address the inhibition of acetylcholinesterase. Such evidences indicate either a non-expert evaluation, or extreme hard cases of judgement for experts.

8.9 Discussion

The semantic framework that has been previously developed (*see* Chapter 7) has been customized to the domain of biomedical patent mining, which is one of the most challenging issues. Indexing with pre-selected concepts, their hyponyms and synonyms allows semantic search in the concept space. In addition, the performance of system with different search strategies has been systematically evaluated. It was shown that full-document patent searches perform better than sub-section searches. Semantic search in the concept space indicated good results in comparison to the conventional text search. During preliminary experiments, the result of full-document concept-based search outperformed rest of the runs with best bpref score of 0.32. Combining the text search with semantic search can yield improved results when compared to individual searches. A systematic optimization of retrieval function, and exploiting the IPC classification information can further improve the performance of retrieval.

Currently, the system is indexed with pre-selected concepts that appear in TS topics. Indexing the biomedical and chemistry concepts that appear in complete MeSH or UMLS thesauri makes the system more applicable for general ad-hoc retrieval situations. However, this is not a trivial task since these thesauri contain substantial noise that may hinder the performance of retrieval. Currently, the performance of retrieval has been tested with 9 topics. In the future, it is necessary to evaluate the system using more questions. This minimizes the deviation of results from the standard average and gives a better estimation of actual system's performance. The system with comprehensively indexed biomedical and chemical terminologies is believed to substantially enhance the search performance.

From an application point of view, the adapted semantic search strategy can substantially improve the daunting task of fetching relevant information from patents. However, the next challenge to be addressed is an ability to efficiently perform semantic searches over patents from different national patent offices such as the German Patent Office (DPMA), Canadian Intellectual Patent Office (CIPO), and so forth. The author believes that the employed strategy will strongly benefit patent searchers in biomedical, pharmaceutical, and healthcare domains to address the challenges associated with patent searching as well as drives the idea of faster knowledge-to-innovation.

Alike the importance of technology survey in patents, finding similarity of contents amongst different patents is an interesting issue and this will be addressed in the upcoming chapter.

Chapter 9

Prior Art Search in Patents based on Semantic Concepts

Prior art search is a task of querying and retrieving patents in order to uncover any knowledge existing prior to the inventor's question or invention at hand. In other words, prior art search can be addressed as a task of finding similarity between different patents. An example of prior art search can be demonstrated as *find all patents that are similar to the US patent US7985758 entitled Piperidine derivatives for treatment of Alzheimer's disease*. Prior art search in patents can find potential benefits in biomedical, pharmaceutical, and healthcare domains by interlinking the related knowledge described in different patents published over different times. Beyond the information gain, it can help patent search professionals in industries to carry out patent infringement searches more effectively. A recent example of patent infringement the one where the Merck[1] filed a lawsuit against the Impax Laboratories[2] for selling a copycat version of Vytorin, a controversial cholesterol-lowering medication.[3] Therefore, the prior art search in patents can find benefits in fostering the research and development as well as support the secure way of usage of modern technologies.

TRECCHEM addresses this challenge in terms of a trier namely the prior art search task. This task provides a set of test patents for retrieving sets of patents from the predefined collection that can potentially invalidate the given set of test patents. Based on the success demonstrated by the semantic platform for technology survey in patents (see Chapter 8), this chapter focusses on the application of semantic search strategy for performing prior art search using the TRECCHEM patent collection.

9.1 Task Description

The data provided for the Prior Art (PA) search task contains approximately 1.3 million patents from the European Patent Office (EPO), the US Patent and Trademark Office (USPTO), the World Intellectual Property Organization (WIPO) as well as 1000 test (query) patent applications. The task is to retrieve sets of documents from the patent

[1] http://www.merck.com/index.html
[2] http://www.impaxlabs.com/
[3] http://www.theheart.org/article/1112923.do

corpus that can invalidate each test patent application. An example of such a task is "PA-1: Find all patents in the collection that would potentially be able to invalidate US-6090800-A". For a given test patent application, a retrieved patent was considered to be relevant if it satisfies one of the following three situations:

- It is directly cited by the test patent.
- It is a family member of a patent directly cited by the test patent.
- It is directly cited by a family member of the test patent.

9.2 Data Preprocessing

The TRECCHEM corpus collection was provided in Extensible Markup Language (XML). As a preliminary measure, an analysis of different sections within the patents was performed. Patent documents contain several fields that are presumably not necessary during retrieval and generate substantial noise while processing the documents. Examples of such fields are country, legal-status, or non-English abstracts. The aim was to use only those fields that have high text-to-noise ratio and that encompass rich information content. Therefore, with a retrieval point of view, the following fields were chosen to be used for indexing and further assessments: UCID, publication date, priority date(s), patent citation(s), inventor(s), assignee(s), author(s), IPC[4] class, title, abstract, description, and claim(s).

9.3 Recognition of Biomedical and Chemical Entities

A preliminary analysis of the IPC classes showed that a large portion of the corpus belongs to A61 (Medical and Veterinary Science) and C07 (Organic Chemistry). The hypothesis is that named entity recognition of chemicals and biomedical terms helps to overcome the problems associated with synonyms by automatic query expansion. ProMiner was used for the task of named entity recognition in the title, abstract, claims, and description sections of all the patents. The following classes of entities were used for tagging:

Chemical Names: Chemical names including synonyms, formulae, IUPAC, and brand names of chemical compounds as extracted from DrugBank, KEGG[5] Drug and KEGG Compound databases. Additionally, a machine learning-based system [Klinger et al. (2008)] was applied for tagging the IUPAC-like names. It performs an internal normalization to map different variants to one base form.

[4] International Patent Classification
[5] http://www.genome.jp/kegg/

Informative Noun Phrases	Non-informative Noun Phrases
curable composition	1 2 3 1 2 m 4 R=H
methoxypropynyl group	the claims
biodegradable collagen	about 1800 mg/kg
self-adhesive CODAL tape	A)1>[M M]/(4 [M M] [M M])
tyrosine kinase inhibitor	such difficulties

Table 9.1: Examples of extracted noun phrases classified as either informative or non-informative.

Genes/Proteins: Human genes and protein names as well as their synonyms that are extracted from EntrezGene[6] and UniProt[7].

Diseases: Disease names and their synonyms that are extracted from the Medical Subject Headings (MeSH).

Pharma Terms: Pharmacological terms that are extracted from the Anatomical Therapeutic Chemical (ATC) drug classification system. Since the ATC does not contain synonyms and term variants, this information was gathered from UMLS with the help of the MetaMap program [Aronson (2001)].

Noun Phrases: The OpenNLP-based NP chunker[8] was applied for tagging the noun phrases. Noninformative noun phrases were filtered off in a systematic way [Gurulingappa et al. (2009)]. Examples of informative and non-informative noun phrases can be found in Table 9.1. The remaining noun phrases were normalized using the LVG Norm program [Browne et al. (2003)] provided within the Specialist NLP package by the National Library of Medicine (NLM).

9.4 Indexing

Following the data preprocessing and name entity recognition, the document texts as well as the biomedical entities, chemical entities, and noun phrases occurring within them were indexed with SCAIView. Figure 9.1 shows an overview of the workflow implemented for the PA task. Unlike a conventional index that contains only tokens, the used index additionally contains noun phrases, chemicals, and biomedical entities. Table 9.2 shows the frequency of different entities occurring in the entire corpus as well as the number of documents that contain at least one entity of interest.

[6] http://www.ncbi.nlm.nih.gov/gene
[7] http://www.uniprot.org/
[8] http://opennlp.sourceforge.net/projects.html

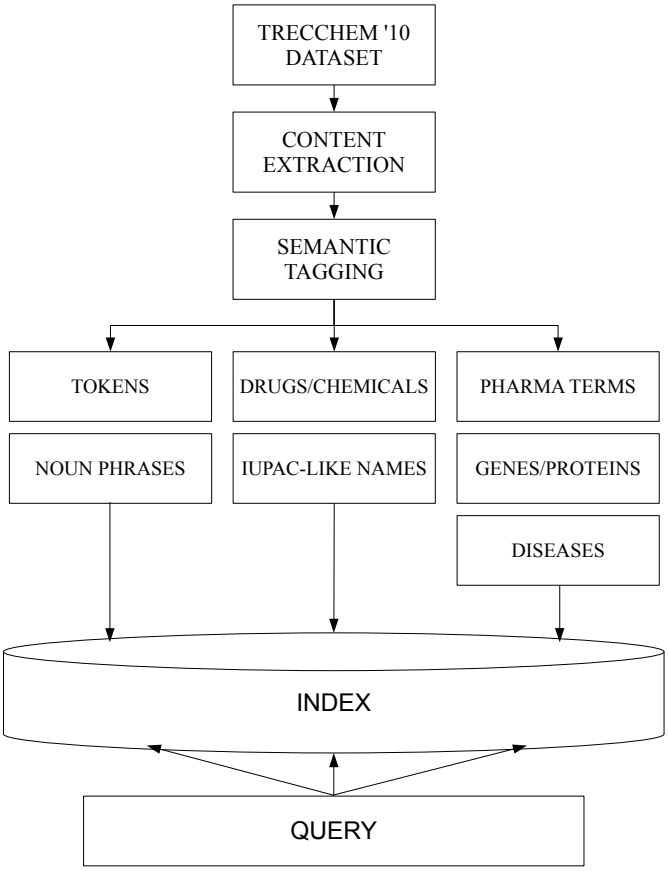

Figure 9.1: Overview of the workflow implemented for prior art search task.

9.5 Querying and Retrieval

Altogether, 7 runs were performed for the prior art search task. The queries were performed using different entity types occurring in the query documents. Based on the experiences from previous TREC task, only the complete document searches were performed and the 4-digit IPC information was utilized. The documents were retrieved and ranked based on the Lucene BM25F function with the default parameters. Different objects that were used for querying are:

Tokens: Search with all tokens that occur in a query patent

Noun Phrases: Search with all noun phrases that occur in a query patent

Entities: Search with all chemical entities (chemical names and IUPAC-like) and biomedical entities (pharma terms, genes/proteins and diseases) that occur in a query patent.

The retrieved documents were filtered based on the following criteria:

	No. of unique entities		No. of documents	
Entity Class	Large Corpus	Query Corpus	Large Corpus	Query Corpus
Chemical Names	12,296	2,467	1,151,477	999
IUPAC-like	2,656,128	18,374	283,677	484
Pharma Terms	479	232	725,325	915
Genes/Proteins	18,641	1,132	883,333	478
Diseases	4,222	833	565,763	336
Noun Phrases	10,158,177	167,851	1,276,229	1000

Table 9.2: Frequencies of dictionary entries occurring within the the large corpus as well as the query corpus and counts of documents containing at least one entity of interest.

Priority date: The earliest priority date of the retrieved document must be lesser than the earliest priority date of the query document.

Family: The retrieved document and the query document must not belong to the same family.

Assignee: The retrieved document and the query document must not have the same assignee and title.

9.6 Results

For the PA task, the reported results are based on the Binary Preference (bpref) and Mean Average Precision (MAP) scores. Table 9.3 shows the results of retrieval using tokens, noun phrases and entities. The run with noun phrase queries outperformed the run with token queries with a boost in MAP score by 0.0379. Since the entities does not occur in all the query documents they were coupled with noun phrases and used for querying. A run with the combination of noun phrase and entity queries performed better than the run with the noun phrase queries alone with an improvement in the MAP score by 0.0114. In order to test the significance of using entities for querying, a paired t-test was performed using the results of noun phrase queries and combined noun phrase and entity queries. A p-value of 0.0001 indicated that using the entities in combination with noun phrases can have a significant impact on the retrieval.

9.6.1 Co-Citation Analysis

The experiences from 2009 TREC task showed that utilizing the citation information for post-processing can boost the results by a large margin [Gobeill et al. (2009)]. Therefore, for each query document, the citations in the retrieved documents were systematically used to generate a ranked document set that can potentially invalidate the respective

Query Type	Run ID	MAP	bpref
Tokens	SCAI10NRMTOK	0.0172	0.1536
NP	SCAI10NRMNP	0.0551	0.3702
NP + Ent	SCAI10NRMENT	0.0665	0.4171

Table 9.3: Results of baseline runs with tokens, noun phrases (NP), and entities (Ent) used as queries.

Query Type	Run ID	MAP	bpref
Tokens	SCAI10CITTOK	0.0947	0.2804
NP	SCAI10CITNP	0.2065	0.5110
NP + Ent	SCAI10CITENT	0.2336	0.5468

Table 9.4: Results of runs with tokens, noun phrases (NP) and entities (Ent) used as queries and co-citation based post-processing.

query document. For a given query document D_i, let D_j be the retrieved document and D_c is cited by D_j. Then, the *co-citation score* of D_c is computed from the top retrieved j number of documents as in equation 9.1.

Table 9.4 shows the results of co-citation based document ranking with tokens, noun phrases and entities used as queries. In comparison to the baseline results, the performance of the system improved by a factor of nearly 4. When the priority date filter was turned off, the co-citation based post-processing with noun phrase and entity queries yielded the MAP score of 0.4121 and Bpref score of 0.7075 (Run ID: SCAI10CIENTP). Nevertheless, using the patents that have priority date later than the query patent makes the model unrealistic.

In addition, the co-citation network based document re-ranking strategy proposed by Gobeill et al. Gobeill et al. (2009) was tested. Querying with noun phrases and entities coupled with the post-processing as proposed by Gobeill et al. resulted in the MAP score of 0.1420 and Bpref score of 0.5700. Therefore, the post-processing strategy implemented within this work resulted in a MAP score better than the proposed state-of-the-art strategy with a slight decrease in the bpref score.

The best result obtained by the run SCAI10CITENT was analysed based on the different IPC classes. Figure 9.2 and Figure 9.3 show the average MAP and bpref scores achieved by the top 20 IPC classes of query patents respectively. Analysis of Figure 9.2 shows that the best MAP scores are achieved by the test patent that belong to the IPC class A61B (DIAGNOSIS; SURGERY;IDENTIFICATION) followed by C25C (PROCESSES FOR THE ELECTROLYTIC PRODUCTION, RECOVERY OR REFINING OF METALS; APPARATUS THEREOF) and A23C (DAIRY PRODUCTS, E.G. MILK, BUTTER, CHEESE; MILK OR CHEESE SUBSTITUTES; MAKING THEREOF). Figure 9.3 shows that the best bpref scores are acheieved by the test patent that belong to the IPC class A61B, C21C (processing of pig-iron, e.g. refining, manufacture of wrought-iron or steel) and A23C. Figure 9.4 shows the average MAP

and bpref scores achieved by the test patents belonging to the different patent offices. Since the citations are used as a gold standard for evaluation and a major portion of TREC dataset is formed by the USPTO patents, this may be one potential reason for achieving the better performance with USPTO patents than EPO or WIPO patents.

$$co\text{-}citation\ score(D_c) = \sum_{j=1}^{1000} \frac{retrieval\ score(D_j)}{rank(D_j)} \tag{9.1}$$

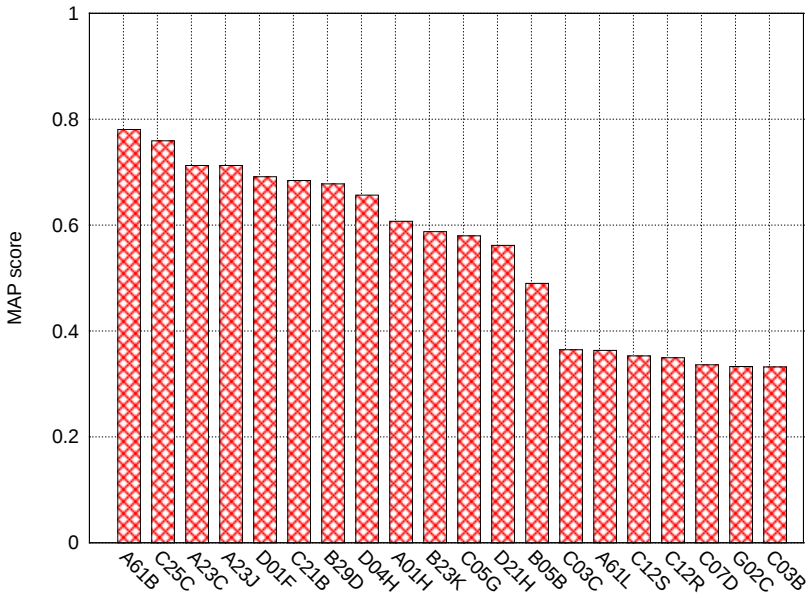

Figure 9.2: Average MAP scores achieved by the top 20 IPC classes of test patents.

Figure 9.5 shows the differences in MAP scores between noun phrase-based querying (Run ID: SCAI10NRMNP) and token-based querying (Run ID: SCAI10NRMTOK). It can be observed that over 60% of the test patents had an observable gain in the MAP score with noun phrase queries. For about 35% of the test patents, using the noun phrases did not show any effect. Whereas for nearly 5% of the test patents, using the noun phrases resulted in a decrease in MAP scores. The test patents that showed an improvement with using the noun phrase queries were analyzed with respect to their IPC classes. It was observed that a large portion of test patents having an improvement in retrieval belongs to the following IPC classes: A61K (PREPARATIONS FOR MEDICAL, DENTAL, OR TOILET PURPOSES), C07D (HETEROCYCLIC COMPOUNDS) and A61P (SPECIFIC THERAPEUTIC ACTIVITY OF CHEMICAL COMPOUNDS OR MEDICINAL PREPARATIONS).

Figure 9.6 shows the differences in MAP scores between a combined entity-noun phrase querying (Run ID: SCAI10NRMENT) and noun phrase-based querying (Run ID: SCAI10NRMNP). It can be observed that nearly 50% of the test patents had an observable gain in the MAP score with a combined entity-noun phrase querying. Nearly

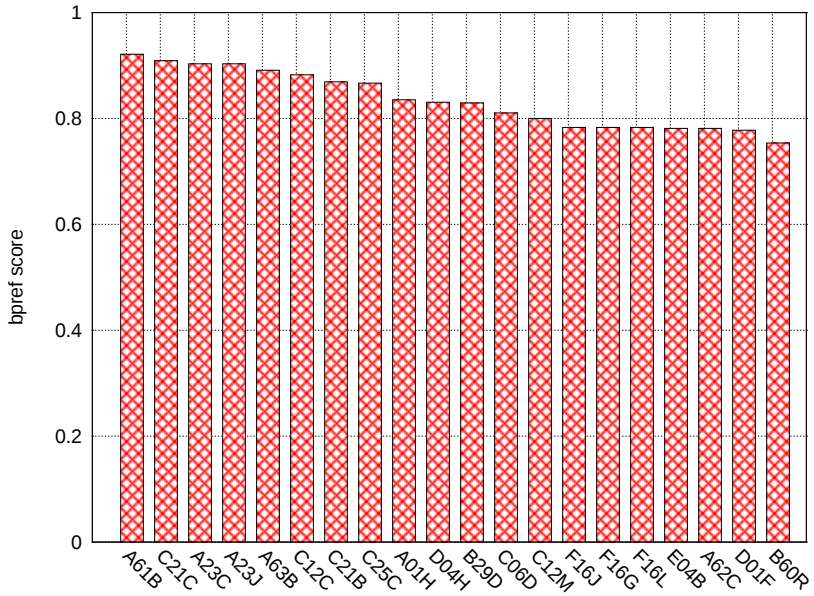

Figure 9.3: Average bpref scores achieved by the top 20 IPC classes of test patents.

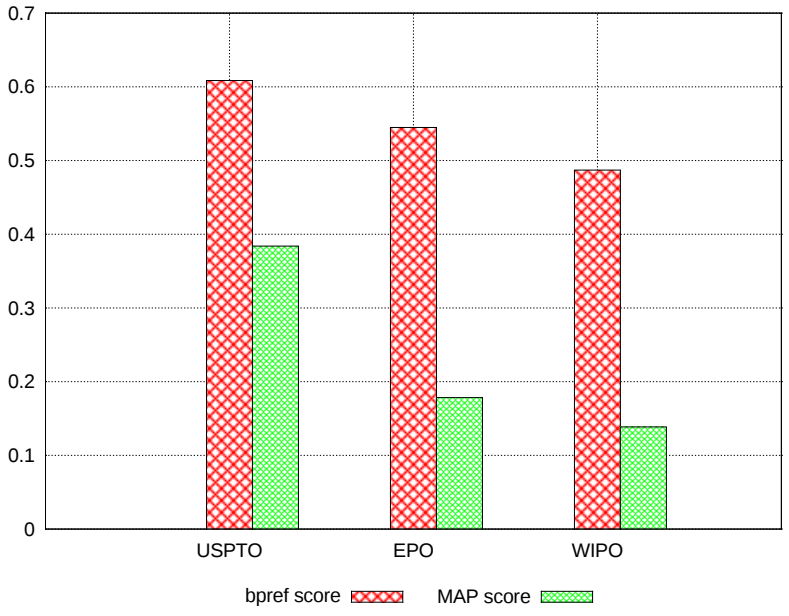

Figure 9.4: Average MAP and bpref scores achieved by the test patents from different patent offices.

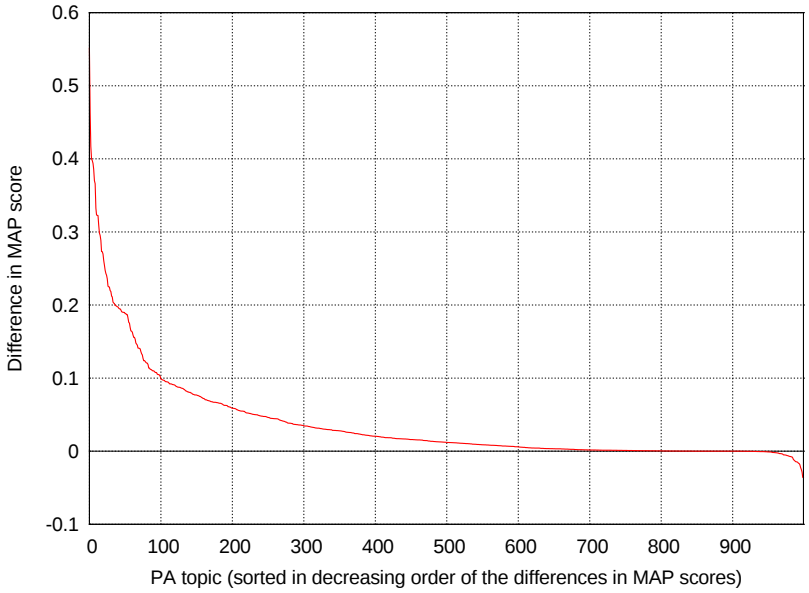

Figure 9.5: Differences in MAP scores between the runs SCAI10NRMNP and SCAI10NRMTOK. PA-topics are sorted in the decreasing order of the differences in MAP scores.

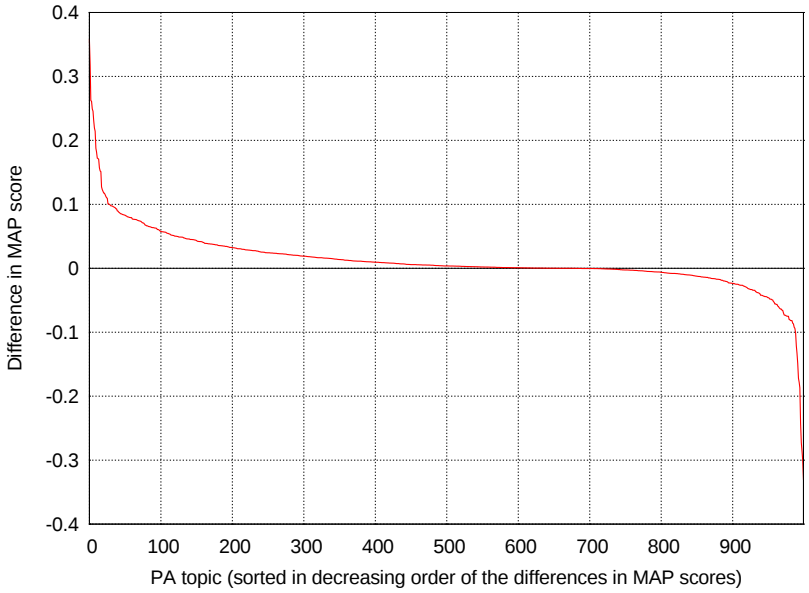

Figure 9.6: Differences in MAP scores between the runs SCAI10NRMENT and SCAI10NRMNP. PA-topics are sorted in the decreasing order of the differences in MAP scores.

30% of the test patents had no impact with entities whereas nearly 20% of the test patents showed decrease in the performance. It was observed that a large portion of test patents having an improvement with entity-noun phrase querying belongs to the following IPC classes: A61K, A61P and C08B (POLYSACCHARIDES; DERIVATIVES THEREOF).

9.7 Discussion

This chapter demonstrates the application of semantic search platform for solving the challenging task of prior art search in patents. Performance of retrieval using tokens, noun phrases, and named entities has been demonstrated. It was shown that using a combination of noun phrases and entities for querying can perform significantly better than using the tokens or noun phrases alone. The ability of co-citation based post-processing strategy for boosting the performance has been successfully shown. In comparison to state-of-the-art, the performance of adopted co-citation based post-processing has been shown to achieve relatively higher MAP score.

There are several ways to improve the performance of the retrieval. Currently, the breadth of knowledge sources that has been used is limited. For example, only the chemicals present in DrugBank and KEGG databases have been used. These databases are specialized to include the compounds that are of biomedical interest and does not focus on the chemicals present in ink formulations, cement or fertilizers. Considering the scope of IPC classes of the documents provided within the TREC data set, less than 50% of the documents belong to the biomedical domain. Therefore, indexing the entities using broader resources that cover terminologies beyond the biomedical domain has to be tested in future. Improving the recognition performance of the entity recognizers and the noun phrase chunker over patents can also contribute to the better retrieval.

The applied semantic search strategy has demonstrated success during both technology survey search (see Chapter 8) and prior art search in patents. It has a potential to support biomedical and chemical patent experts and researchers to perform patent searches more efficiently than ever. This in-turn can positively influence the strategies of patent mining in next days. Finally, the developed framework is believed to find potential applications in patent infringement analysis, portfolio analysis, R&D investment policies, literature-based knowledge discovery, and biomedical and pharmaceutical decision-making that can drive lab-to-product strategies.

Chapter 10

Adverse Drug Event Detection in Medical Case Reports

Adverse effects of drugs is a bothersome issue that confronts pharmaceutical and healthcare domains. Majority of adverse effects appear after drug regulatory approval and marketing [Hauben and Bate (2009)]. Surveillance of fate of drugs after their release into the market is a challenging issue since a lot of information about their adverse effects are discretely reported amongst surveillance systems (such as the FDA's MedWatch[1]), and free-text (such as case reports, blogs, etc.). A recent example include Shetty and Dalal (2011) who investigated by means of statistical document classification that nearly 54% of "detected FDA warnings" about particular drugs existed in the literature before those alerts were officially issued[2]. Therefore, a strategy for automatic identification of adverse effects of drugs from free-text resources can accelerate signal detection and medical decision making with a benefit of limited manual reading requirement. However, the goal is hard to achieve since there is an extremely limited availability of annotated textual corpus that can support the development or validation of literature mining techniques for the adverse effect detection.

This chapter covers a strategy for systematic development of a corpus of medical case reports that can support the development of adverse effect detection systems from text. A lot of medical case reports contain information about patient's treatment, diagnosis, and their outcomes that are unusual or novel in terms of appearance [Vandenbroucke (2001)]. Furthermore, the generated corpus is applied for the development of a machine learning-based system for the automatic identification of adverse drug event assertive sentences in case reports. In addition, the system also employs dictionary-based named entity recognition for identifying the co-occurring drugs and conditions. A study conducted in order to investigate the ability of the system to capture novel or rarely noticed adverse effects of selected drug in the market showed interesting results. The following sections provide details of the workflow implementation and results of the system's evaluation.

[1] http://www.fda.gov/Safety/MedWatch/default.htm

[2] http://www.drugsafetydirections.com/forum

10.1 Corpus Generation

10.1.1 The ADE Corpus Characteristics

During the development of a benchmark corpus, two characteristics have to be considered. They are the domain suitability of the corpus and the target user group. Considering the domain suitability, medical case reports were of the first choice since they provide important and detailed information about symptoms, signs, diagnosis, treatment, and follow-up of the individual patients. More importantly, case reports can serve as an early warning signal for the under-reported or unusual adverse effects of medications [Kidd and Hubbard (2007)]. Since the goal of this work is to generate a corpus for public usability, MEDLINE articles were used due to their nature of free public availability. Therefore, the ADE corpus constitutes a subset of MEDLINE case reports.

10.1.2 Document Sampling

Currently, MEDLINE contains more than 1.5 million medical case reports. In order to restrict the scope of the corpus to drug-related adverse events, a PubMed search with drug therapy and adverse effect as MeSH terms was performed limiting the language to English. The text option was chosen to be abstract in order to eliminate the documents with only title and no abstract text. A precise PubMed query performed on 2010/10/07 is as follows:

"adverse effects"[sh] AND (hasabstract[text] AND Case Reports[ptyp]) AND "drug therapy"[sh] AND English[lang] AND (Case Reports[ptyp] AND ("1" [PDAT] : "2010/10/07" [PDAT]))

This process retrieved nearly 30,000 documents from PubMed out of which 3,000 documents (referred to as ADE corpus) were randomly selected for the annotation and benchmarking purpose. A corpus of 3,000 systematically annotated documents is believed to be substantially large to support the development and validation of information extraction systems. An additional set of 100 non-overlapping documents (referred to as ADE-SEED corpus) were selected in order to be used by the annotators for practicing the annotation task as well as for annotation quality assessment.

10.2 Annotation Guidelines

A critical issue that reflects the quality of an annotated corpus is consistency [Roberts et al. (2009)]. In order to generate an annotated corpus for information extraction modeling or performance benchmarking, consistent and uniform annotation across all the documents is essential. To ensure consistency, a set of draft guidelines were developed and provided to the annotators. The guidelines provide a set of rules which

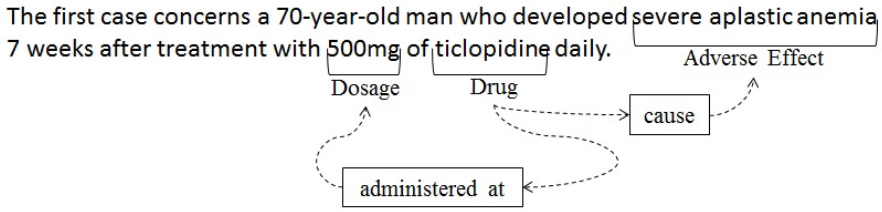

Figure 10.1: Example of a sentence in the ADE corpus annotated with drug, adverse effect, and the relationship between them.

annotators should follow when working on a document. The draft guidelines were periodically revised before beginning the annotation of ADE corpus (see Section 10.3.2 for details). Important components of the annotation guidelines are as follows:

Drug: Names of drugs and chemicals that include brand names, trivial names, abbreviations and systematic names were annotated. The mentions of drug or chemicals should strictly be in a therapeutic context. This category does not include the names of metabolites, reaction byproducts, or hospital chemicals (e.g. surgical equipment disinfectants).

Adverse effect: The mentions of adverse effects include signs, symptoms, diseases, disorders, acquired abnormalities, deficiencies, organ damage or death that strictly occur as a consequence of drug intake.

Dosage: Dosage information that includes the quantitative measurements (e.g. 0.1 mg/kg/day) as well the frequency mentions (e.g. two tablets twice daily) was annotated.

Relationship: The scope of a relationship was defined and restricted to the sentence level. There should be a clear mention of a drug/chemical resulting in an adverse effect defined within the context of a sentence. The mentions of drug, disorders or dosages that do not fit into a relation were not annotated. Relationships were annotated between the drugs and adverse effects as well as between the drugs and the dosages in an implicit manner. This means that the interrelated entities were represented in a systematic way that allows machine adaptation and training but not explicitly marked using the annotation tool. Figure 10.1 shows an illustration of a sentence annotated with the entities and relationship between them.

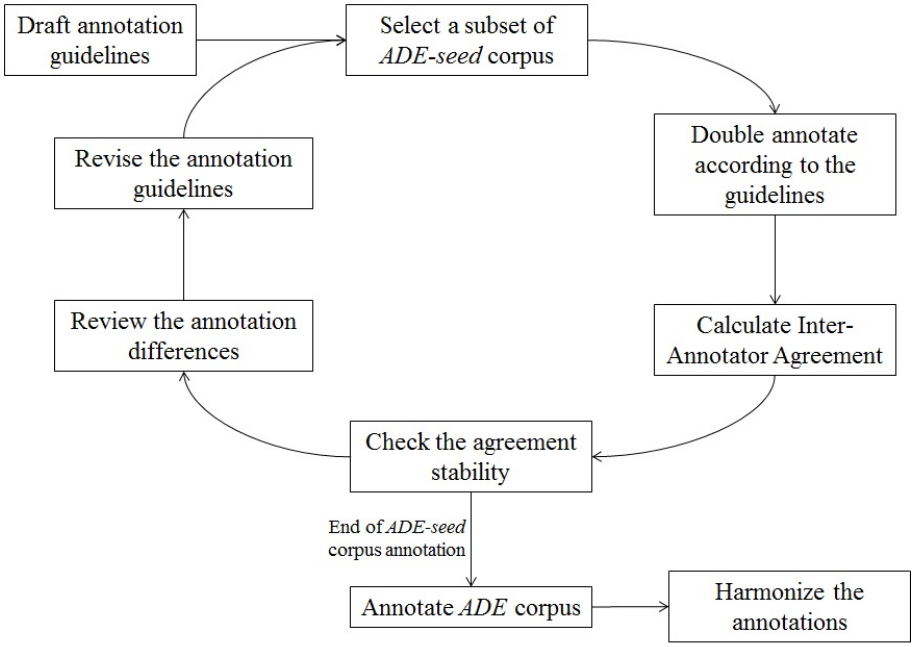

Figure 10.2: The workflow employed for the annotation task.

10.3 Annotation Methodology

10.3.1 Annotation Participants

Altogether, five individuals participated in the generation and revision of the annotation guidelines. Amongst them, three individuals were involved in the annotation task. All the annotators possess a minimum qualification of Master of Science academic degree with the background related to Biomedicine. Two annotators have substantial experience working in the biomedical text mining domain whereas the third annotator has comparatively little practical experience working with text mining-related topics.

10.3.2 Annotation Workflow

The annotation workflow follows the standards established by the CLEF framework [Roberts et al. (2009)]. Knowtator[3] version 1.9 beta 2 was the tool used for annotation. The CLEF framework provides an easily configurable text annotation environment plugged into the knowtator toolkit. Figure 10.2 shows the workflow adapted for the annotation task.

An individually single annotated document can reflect several problems. They include idiosyncratic errors made by the annotators, missing annotations or the consistent under-performance of the individuals. In order to overcome these problems, a strategy

[3] http://knowtator.sourceforge.net/

of double annotation [Wilbur et al. (2006)] was applied. During the process of double annotation, each document is independently annotated by at least two annotators and the sets of annotations are compared thereafter for quality assurance. The annotation task started with applying the draft guidelines for annotating the ADE-SEED corpus. First, the ADE-SEED corpus of 100 documents was divided into ADE-SEED-SET1 and ADE-SEED-SET2 comprising 50 non-overlapping documents each. As indicated in the Figure 10.2, initially the ADE-SEED-SET1 sub-corpus was annotated by all the three annotators by strictly applying the draft guidelines provided. The agreement between the annotators was calculated using the Inter-Annotator Agreement (IAA) score (see Section 10.4.1). The IAA scores were determined for the entities as well as for the relationships (see Section 10.4.2). The stability of agreement was determined for all the documents and the under-performing documents were manually reviewed to check for the disagreeing instances. Depending on the necessity, changes were made to the annotation guidelines that were used. The process was repeated for the ADE-SEED-SET2. Counts of the annotated entities and relationships over the ADE-SEED corpus for two preliminary rounds of annotation are provided in Table 10.1 and Table 10.2. Before starting the annotation of the ADE corpus, an interactive stabilization of the annotation guidelines was performed based on the experiences gained during the annotation of ADE-SEED corpus. The ADE corpus of 3,000 documents was divided into ADE-SET1, ADE-SET2, and ADE-SET3 subsets with each comprising 1,000 non-overlapping documents. Each annotator processed two subsets of corpus. With this strategy, every document was annotated by two annotators and the total number of documents that each annotator has to read was reduced by one-third. Figure 10.3 shows the distribution of the subsets of ADE corpus among the different annotators. Table 10.3 shows the counts of the annotated entities and relationships over the ADE corpus.

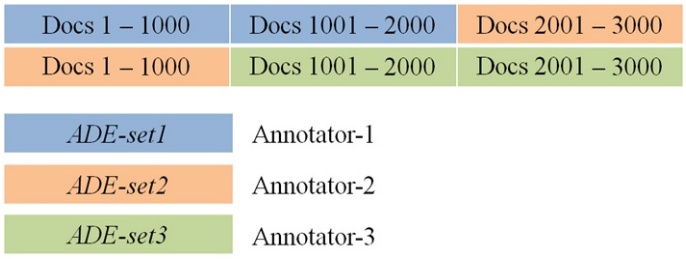

Figure 10.3: Distribution of the subsets of ADE corpus among the different annotators. Each subset contains 1,000 non-overlapping documents.

10.3.3 Annotation Harmonization

During the harmonization process, the double annotated documents were subjected to a review by the respective annotators in order to resolve the conflicting annotations and to improve the overall quality of the annotated corpus. The aim of annotation

	Entity Counts			Relation Counts	
	Drug	Adverse Effect	Dosage	Drug-Adverse Effect	Drug-Dosage
Annotator-1	116	139	0	166 (90)	0 (0)
Annotator-2	120	159	0	177 (84)	0 (0)
Annotator-3	57	132	0	52 (26)	0 (0)

Table 10.1: Counts of the annotated entities and relations in the ADE-SEED-SET1 corpus. Numbers within the brackets indicate the unique number of sentences that contain at least one relation. Enumerations related to dosages are zeroes since no dosage information was annotated during this round.

	Entity Counts			Relation Counts	
	Drug	Adverse Effect	Dosage	Drug-Adverse Effect	Drug-Dosage
Annotator-1	91	83	4	110 (68)	4 (4)
Annotator-2	86	77	3	95 (65)	3 (3)
Annotator-3	54	60	0	59 (46)	0 (0)

Table 10.2: Counts of the annotated entities and relations in the ADE-SEED-SET2 corpus. Numbers within the brackets indicate the unique number of sentences that contain at least one relation.

harmonization is to focus on the differences in the annotator's interpretation of the guidelines and the differences in their interpretation of the documents. However, the process of harmonization does not attempt to find the actual ground truth mentioned in the documents. Documents that do not contain any annotation from both the annotators or the documents where both the annotators agree completely were not reviewed. Documents that contain at least one conflicting annotation were subjected to the review process by the respective annotators. The following precautions were taken during the harmonization process.

	Entity Counts			Relation Counts	
	Drug	Adverse Effect	Dosage	Drug-Adverse Effect	Drug-Dosage
Annotator-1	2391	3330	129	3995 (2490)	140 (111)
Annotator-2	3097	3464	69	4028 (2681)	71 (60)
Annotator-3	3999	4604	77	5489 (3404)	83 (77)

Table 10.3: Counts of the annotated entities and relations in the ADE corpus. Numbers within the brackets indicate the unique number of sentences that contain at least one relation. Each annotator handles only 2000 documents that are distributed according to Figure 10.3.

Entity Counts		Relation Counts	
Drug	5063	Drug-Adverse Effect	6821 (4272)
Adverse Effect	5776	Drug-Dosage	279 (213)
Dosage	231		

Table 10.4: Counts of the annotated entities and relations in the ADE corpus after harmonization. Numbers within the brackets indicate the unique number of sentences that contain at least one relation.

1. No entirely new annotations were added if they were not marked earlier by either of the annotators.
2. No annotations were removed if they were marked earlier by both the annotators.
3. Annotations were added or removed if they were marked by any one of the annotators and provided they both agree on the decision thereafter.
4. In case of partially overlapping annotations, only the conflicting parts were resolved. For instance, Annotator-1 marks *acute lymphoblastic leukemia* whereas the Annotator-2 marks *lymphoblastic leukemia*, then the decision will be made to resolve the annotation of the word *acute*.

The harmonization was performed over the complete ADE corpus in the presence of annotators for both the entities as well as the relationships. Table 10.4 shows the counts of the annotated entities and relationships over the ADE corpus after the harmonization procedure. 28 documents were removed from the ADE corpus due to errors induced by the annotation software as well as manual handling errors (such as missing annotations, annotation offset shifts, etc.). After the end of harmonization, the ADE corpus contains 2,972 documents having 4272 sentences annotated with names and relationships between drugs, adverse effects and dosages. The sentences with drug-dosage relationships (*i.e.* 213 sentences) constitute a subset of 4272 sentences that contain drug-adverse effect relationships.

10.4 Assessment of Inter-Annotator Agreement

10.4.1 Inter-Annotator Agreement Metrics

Over the ADE-SEED as well as the ADE corpora, the double annotated documents were used for the determination of Inter-Annotator Agreement (IAA) scores. The IAA scores were calculated using the F_1 score as a criterion. The F_1 score measures the harmonic mean of precision and recall between the annotators using one annotator as a standard and the other as a reference. The IAA scores were determined for both the entities as well as for the relationships. GATE [Cunningham et al. (2002)] framework was used

for the determination of IAA scores. For the entities, IAA scores were determined using the *exact match* and *partial match* as criteria. *Exact match* is a situation where both the annotations should completely overlap whereas *partial match* is a situation where the annotations may partially or completely overlap. For the relationships, two types of evaluations were applied. They are the *exact entity match with exact relation* and *partial entity match with exact relation*. The *exact entity match with exact relation* requires that the annotations of the entities overlap completely and the relationship is correctly annotated. In case of *partial entity match with exact relation*, a relationship that links two partially or completely matching entity spans is considered to be correct. For the entities, in addition to F_1 score for the IAA calculation, *kappa*[4] values have been provided in-order to allow cross-group IAA comparisons.

10.4.2 Inter-Annotator Agreement Calculation

The IAA scores between the annotators were determined over the ADE-SEED corpus during two preliminary rounds of annotation. Whereas, the IAA scores over the ADES corpus were determined before the final harmonization was performed. The agreement levels were determined for the entities as well as for the relationships. Table 10.5 and Table 10.7 show the IAA F_1 scores over the ADE-SEED-SET1 corpus for entities and relationships. Table 10.8 and Table 10.10 show the IAA F_1 scores over the ADE-SEED-SET2 corpus for entities and relationships respectively. The ADE-SEED-SET1 corpus did not contain any mentions of dosages that fit into a pre-defined relationship with drugs. Therefore, the IAA scores for dosages were enumerated as zero for the entity mentions as well as for the relationships with drugs. During the preliminary annotation rounds, the level of agreement between Annotator-1 and Annotator-2 remained consistent for the drug names. A potential reason is that the drug names often occur as one word entities (*e.g. minocycline*) and they hardly suffer from boundary mismatch problems. However, the agreement level for the exact name matches of adverse effects and dosages was unsatisfactory. The names of adverse effects often occur as descriptive multi-word terms and deciding the correct term boundaries was a major problem. For instance, Annotator-1 marked *non-metastatic gestational trophoblastic tumor* whereas the Annotator-2 marked the same instance as *gestational trophoblastic tumor*. Nevertheless, the partial name matches for adverse effects had substantial level of agreement. Dosage information faced severe annotation problems. Mentions such as low-dose were often misinterpreted or overseen by the annotators and were not annotated. Such instances represent the contemporary errors induced during the annotation process that were improved later on. Typical examples of the relationship annotation errors include the distantly related entities. For instance, in the sentence *the patient developed monoarthritis 2 weeks after initiation of IFN-beta, which persisted during 14 months of therapy and resolved with discontinuation of IFN-beta*, there exist two relationships between *monoarthritis* and two mentions of *IFN-beta*. The relationship between the nearest co-occurring entities was correctly annotated whereas the second relationship was overseen by one of

[4] http://en.wikipedia.org/wiki/Cohen's_kappa

the annotators. Such instances were exemplified in the annotation guidelines and thoroughly discussed before the annotation of main corpus was performed. Annotator-3 having minimum experience with text annotation exercises often achieved lower agreement scores with rest of the annotators.

	Entity (*Exact Match*)			Entity (*Partial Match*)		
Annotators	Drug	Adverse Effect	Dosage	Drug	Adverse Effect	Dosage
1 & 2	0.76	0.66	0.00	0.82	0.86	0.00
1 & 3	0.28	0.43	0.00	0.38	0.55	0.00
2 & 3	0.29	0.40	0.00	0.38	0.51	0.00

Table 10.5: IAA F_1 scores over entities between the annotators on the ADE-SEED-SET1 corpus containing 50 documents. Enumerations related to dosages are zeroes since no dosage information was annotated during this round.

Annotators	Drug	Adverse Effect	Dosage
1 & 2	0.81	0.74	0.00
1 & 3	0.27	0.36	0.00
2 & 3	0.28	0.32	0.00

Table 10.6: IAA *kappa* scores over entities between the annotators on the ADE-SEED-SET1 corpus containing 50 documents.

	Relation (*Exact Entity & Exact Relation*)		Relation (*Patrial Entity & Exact Relation*)	
Annot.	Drug-Adverse Effect	Drug-Dosage	Drug-Adverse Effect	Drug-Dosage
1 & 2	0.64	0.00	0.79	0.00
1 & 3	0.14	0.00	0.37	0.00
2 & 3	0.10	0.00	0.37	0.00

Table 10.7: IAA F_1 scores over relations between the annotators on the ADE-SEED-SET1 corpus containing 50 documents.

Table 10.11 and Table 10.13 show the IAA F_1 scores between the annotators over the large ADE corpus that contains 3,000 documents. The ADE corpus was strategically divided and annotated by three annotators. Therefore, the IAA scores were determined over the sets of 1,000 documents that were commonly annotated by two annotators. Based on the experiences gained during the preliminary annotation rounds, all the three annotators were able to consistently annotate the drug names. Although, the names of adverse effects underwent frequent boundary problems, the results of partial name matches were consistent amongst all the three annotators.

	Entity (*Exact Match*)			Entity (*Partial Match*)		
Annotators	Drug	Adverse Effect	Dosage	Drug	Adverse Effect	Dosage
1 & 2	0.73	0.88	0.29	0.90	0.88	0.86
1 & 3	0.63	0.65	0.00	0.77	0.66	0.00
2 & 3	0.57	0.66	0.00	0.76	0.67	0.00

Table 10.8: IAA F_1 scores over entities between the annotators on the ADE-SEED-SET2 corpus containing 50 documents.

Annotators	Drug	Adverse Effect	Dosage
1 & 2	0.87	0.76	0.50
1 & 3	0.63	0.59	0.00
2 & 3	0.65	0.54	0.00

Table 10.9: IAA *kappa* scores over entities between the annotators on the ADE-SEED-SET2 corpus containing 50 documents.

	Relation (*Exact Entity & Exact Relation*)		Relation (*Patrial Entity & Exact Relation*)	
Annot.	Drug-Adverse Effect	Drug-Dosage	Drug-Adverse Effect	Drug-Dosage
1 & 2	0.69	0.28	0.87	0.85
1 & 3	0.51	0.00	0.65	0.00
2 & 3	0.46	0.00	0.66	0.00

Table 10.10: IAA F_1 scores over relations between the annotators on the ADE-SEED-SET2 corpus containing 50 documents.

	Entity (*Exact Match*)			Entity (*Partial Match*)		
Annotators	Drug	Adverse Effect	Dosage	Drug	Adverse Effect	Dosage
1 & 2	0.80	0.72	0.26	0.82	0.80	0.43
1 & 3	0.75	0.68	0.05	0.77	0.77	0.37
2 & 3	0.76	0.63	0.03	0.78	0.77	0.09

Table 10.11: IAA F_1 scores over entities between the annotators on the ADE corpus containing 3,000 documents. IAA scores are calculated over the sets of 1,000 documents that are commonly annotated by two annotators.

Annotators	Drug	Adverse Effect	Dosage
1 & 2	0.87	0.73	0.27
1 & 3	0.83	0.67	0.06
2 & 3	0.82	0.65	0.03

Table 10.12: IAA *kappa* scores over entities between the annotators on the ADE corpus containing 3,000 documents. IAA scores are calculated over the sets of 1,000 documents that are commonly annotated by two annotators.

	Relation (*Exact Entity & Exact Relation*)		Relation (*Patrial Entity & Exact Relation*)	
Annot.	Drug-Adverse Effect	Drug-Dosage	Drug-Adverse Effect	Drug-Dosage
1 & 2	0.68	0.17	0.78	0.26
1 & 3	0.63	0.14	0.74	0.18
2 & 3	0.60	0.12	0.75	0.15

Table 10.13: IAA F_1 scores over relations between the annotators on the ADE corpus containing 3,000 documents. IAA scores are calculated over the sets of 1,000 documents that are commonly annotated by two annotators.

In spite of two preliminary rounds of annotation, the IAA scores over the dosage did not improve over the ADE corpus. The primary reason is the missing annotations that confirm with the finding of low IAA scores for partial matches. Since annotators were highly focussed over drugs and their adverse effects, perhaps a lot of dosage annotations were overseen by the annotators. The second reason is the low stability of annotation guidelines for dosages. The dosage annotations of annotator-1 were strictly adhering to quantitative measures of drug administration, whereas the annotator-2 included route of administration, duration, and so-forth. For example, in the sentence *PMID: 3365032 A patient is presented with typical hyperthyroidism, who developed a severe proximal muscle weakness and a raised creatine phosphokinase after treatment for hyperthyroidism with propylthiouracil (100 mg orally, three times a day)*, annotator-1 tagged *100 mg* as the dosage of *propylthiouracil*, whereas the annotator-2 tagged the dosage as *100 mg orally, three times a day*. The dosage information being the poorest annotated entity class was strictly resolved during the harmonization process. All the annotated entities and relationships were subjected to the harmonization procedure after the complete annotation of ADE corpus in the presence of respective annotators in order to achieve a consistent final annotation.

10.4.3 Semantic Corpus Analysis

After the harmonization procedure, in order to analyze the semantic distribution of entities in the ADE corpus, the annotated names of drugs and adverse effects were mapped to standard ontologies using the ProMiner system. The drug names were

ATC Class	% of drugs
Antineoplastic agents	22
Ophthalmologicals	11
Antibacterial agents	11
Immunosuppressants	9
Antiepileptics	8

Table 10.14: Top 5 ATC classes to which the frequently occurring drugs belong.

MedDRA Class	% of AE
Cardiac arrhythmias	12
General system disorders	11
Epidermal and dermal conditions	9
Allergic conditions	9
Hepatic and hepatobiliary disorders	8

Table 10.15: Top 5 MedDRA classes to which the frequently occurring adverse effects (AE) belong.

mapped to the Anatomical Therapeutic Chemical (ATC) classification system using the DrugBank dictionary. The ATC hierarchically classifies several drugs according to their pharmacotherapeutic properties. Since ATC is hierarchical, its level two classes were used for the analysis. The names of adverse effects were mapped to the MedDRA classification system. MedDRA contains a hierarchically organized medical terminology and it been widely applied for pharmacovigilance and drug regulatory affairs. Similar to ATC, the level two MedDRA classes were used for analysis. Out of 5,063 annotated drug names, 4,205 could be normalized to the ATC (*i.e.* 82%) whereas for the adverse effects, 4,356 out of 5776 names (*i.e.* 73%) could be mapped to the MedDRA. Table 10.14 shows top 5 ATC classes to which frequently occurring drugs belong to. Table 10.15 shows top 5 MedDRA classes to which frequently occurring adverse effects belong to.

10.5 Corpus Preparation for Sentence Classification

Any supervised learning problem requires independent training and test sets. For the purpose of training and validating the sentence classifier, the ADE corpus (see Section 10.1) was applied.

The ADE corpus was randomly split into a training set (referred to as ADE-TRAIN) and a test set (referred to as ADE-TEST) containing 2378 and 594 documents respectively (after removal of duplicate sentence). Later on, sentences in the training and test sets were extracted and labeled as either **POSITIVE** or **NEGATIVE**. A sentence was labeled as **POSITIVE** if it contains at least one annotation of drug associated with at

least one adverse effect. The remaining sentences were labeled as **NEGATIVE**. Table 10.16 shows the distribution of sentences and named entities occurring in the working corpora. In addition to ADE corpus, an additional set of 27,000 case reports (referred to as ADE-EXAM) were randomly extracted through a predefined PubMed query[5]. The ADE-EXAM serves as a reference corpus in order to study real use-case scenarios of the drug-related adverse effect sentence classification framework.

Corpus	Training set	Test set
No. of **POSITIVE** sentences	3443	829
No. of **NEGATIVE** sentences	13355	3340
No. of Drug annotations	4085	978
No. of Adverse Effect annotations	4658	1119

Table 10.16: Distribution of sentences and named entities in training and test sets.

10.6 Sentence Classification Framework

The principle behind the adverse effect sentence classification framework is that it utilizes morphosyntactic textual features derived from sentences to classify them as either **POSITIVE** or **NEGATIVE**. A **POSITIVE** labeled sentence is one that contains a clear definition of drug causing or worsening a medical condition. Various sets of features and different supervised classifiers were tested during the training and validation phase. Classifiers that were tested include Naive Bayes, Decision Tree, Maximum Entropy and Support Vector Machines. Finally, the best suited feature set and the classifier were applied to classify sentences in the ADE-EXAM corpus. The sentence classification phase is followed by the named entity detection phase where ProMiner (a dictionary-based named entity recognition system) is employed for identifying the co-occurring drug and condition names in the **POSITIVE** labeled sentences. Quality controlled (referred to as *curated*) versions of DrugBank and MedDRA dictionaries were used for the identification of drugs and conditions respectively.

10.6.1 Feature Generation

During the evaluation of sentence classifier over the training and test sets, various textual feature were employed. Table 10.17 shows an illustration of different feature sets applied for the classification task. Feature sets generated are as follows:

All-Words: Indicates all the words occurring in a sentence expect special characters.

[5] "adverse effects"[sh] AND (hasabstract[text] AND Case Reports[ptyp]) AND "drug therapy"[sh] AND English[lang] AND (Case Reports[ptyp] AND ("1"[PDAT] : "2010/10/07"[PDAT]))

Lemmatized-Tokens: The sentences were tokenized using the Genia[6] tokenizer. All non-special character tokens were lemmatized to their base forms using MorphAdorner[7] English lemmatizer. Lemmatization normalizes lexical token variation to its base form. For instance, tokens *drugs* and *inducing* will be normalized to *drug* and *induce* respectively.

Lexicon-Token-Match: Manually curated single word lexicons generated from DrugBank and MedDRA databases were used as reference sources for drug and condition names. Lemmatized tokens in a sentence were checked for their presence in DrugBank and MedDRA lexicons. Two special features were included that counts the number of *drug-match* and *condition-match* tokens in every sentence.

Lemmatized-Token-Bigrams: Indicates all the pairs of adjoining and lemmatized tokens excluding the special characters strictly occurring in the forward order as their occurrence in a sentence. Tokens that match drug or condition lexicon entries were bound with a common arbitrary string (*see* Table 10.17).

Lemmatized-Token-Trigrams: Indicates all the adjoining and lemmatized three token tuples excluding the special characters strictly occurring in the forward order as their occurrence in the sentence. Tokens that match drug or condition lexicon entries were bound with a common arbitrary string (*see* Table 10.17).

Noun-Character-Affixes: Raw sentences were subjected to parts-of-speech (POS) tagging by the Genia POS tagger. Two, three, and four character suffixes and prefixes of all the noun forms occurring in a sentence were extracted.

Lemmatized-Verbs: For all the *drug-match* and *condition-match* tokens in the sentence, their immediate preceding and succeeding lemmatized verbs were extracted.

Lemmatized-Tokens-In-Window: For all *drug-match* and *condition-match* tokens in a sentence, their immediate preceding and succeeding lemmatized non-special character tokens occurring in a window of size 5 were extracted. Tokens that match drug or condition lexicon entries were bound with a common arbitrary string (*see* Table 10.17).

Stanford-Token-Dependencies: Stanford parser was applied over all the raw sentences and the token dependency pairs generated by the parser were extracted

[6] `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA`
[7] `http://morphadorner.northwestern.edu/`

[Klein and Manning (2003)]. Figure 10.4 shows an example of a sentence parsed by the Stanford parser. Stanford dependency (*collapsed dependency-type*) tokens were lemmatized and strictly used in the order according to the direction of dependencies between the tokens (*see directed edges in* Figure 10.4). Tokens that match drug or condition lexicon entries were bound with a common arbitrary string (*see* Table 10.17).

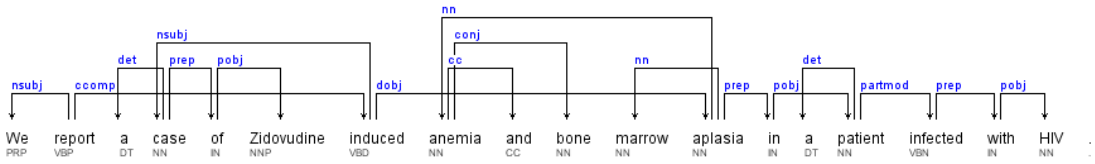

Figure 10.4: Example of Stanford parser token dependencies in a sentence.

Sentence	We report a case of Zidovudine induced anaemia and bone marrow aplasia in a patient infected with HIV.
Label	**POSITIVE**
All-Words	we, report, a, case, of, zidovudine, induced, anaemia, and, bone, etc.
Lemmatized-Tokens	we, report, a, case, of, zidovudine, induce, anaemia, and, bone, etc.
Lexicon-Token-Match	drug-match=1 (i.e. *zidovudine*), condition-match=2 (i.e. *anaemia*, *aplasia*)
Lemmatized-Token-Bigrams	we-report, report-a, a-case, case-of, of-DRUG, DRUG-induce, etc.
Lemmatized-Token-Trigrams	we-report-a, report-a-case, a-case-of, case-of-DRUG, of-DRUG-induce, etc.
Noun-Character-Affixes	affixes of *case* (ca, cas, case, se, ase, case), affixes of *zidovudine* (zi, zid, zido, ne, ine, dine), etc.
Lemmatized-Verbs	zidovudine (Pre-verb=*report*, Post-verb=*induce*), anaemia (Pre-verb=*induce*, Post-verb=*infect*), etc.
Lemmatized-Tokens-In-Window	zidovudine (Pre-words =*we, report, a, case, of*, Post-words =*induce,* CONDITION, *and, bone, marrow*), etc.
Stanford-Token-Dependencies	report-we, report-induce, induce-case, case-of, induce-CONDITION, etc.

Table 10.17: Example of features generated for a sentence in the working corpus. Tokens that match the drug or condition lexicon entries were bound with keywords DRUG or CONDITION respectively.

10.7 Results of Sentence Classification

10.7.1 Performance Evaluation Criteria

The performance of sentence classification was evaluated by 10-fold cross-validation of the training data using different feature sets as well as different classifier. During cross-validation experiments, the classifier performances were assessed using F-score over the class **POSITIVE**. The class **POSITIVE** is of primary interest because positively labeled sentences are indicative of presence of adverse drug effects. For the evaluation of final performance of the classifier over independent test data, various metrics such as Accuracy, class-specific Precision, class-specific Recall, class-specific F_1 score, micro-averaged F_1 score, and macro-averaged F_1 score are reported.

10.7.2 Assessment of Sentence Classification

During the initial experiments, performances of four different classifiers i.e. Naive Bayes (NB), Decision Tree (DT), Maximum Entropy (ME) and Support Vector Machines (SVM) were evaluated by 10-fold cross-validation of labeled sentences in ADE-TRAIN corpus. Class distribution and counts of instances in the training data are available in Table 10.16. Baseline experiments began with applying simple *All-Words* as features and evaluating the performances of different classifiers. Later, different feature sets were incrementally added to the initial feature set and simultaneously the performances of different classifiers were monitored. Table 10.18 shows the results of classifier performances using different feature sets.

All the feature sets extracted from sentences resulted in varying improvement of the classifier performances expect for the *Lemmatized-Token-Trigrams*. Token trigram-based features resulted in a decrease in performances of Naive Bayes and Maximum Entropy classifiers, and therefore were eliminated during further experiments. The Maximum Entropy classifier and SVM demonstrated competitive results with different combinations of feature sets. However, the Maximum Entropy classifier outperformed rest of the classifiers with the best F_1 score of 0.77.

Based on the results obtained during cross-validation experiments, all the feature sets except *Lemmatized-Token-Trigrams* were applied for the classification of sentences in the test set. In addition to an experiment with morphosyntactic features, a baseline test with simple *All-Words* was also performed. The test set comprises of labeled sentences in ADE-TEST corpus. The Maximum Entropy classifier that showed best results during the cross-validation experiments was chosen for the classification of instances in the test set. The results of performance of sentence classification with *All-Words* and morphosyntactic features are provided in Table 10.19. In comparison to baseline results, the F_1 score over class **POSITIVE** for the Maximum Entropy classifier improved by 16% with morphosyntactic textual features. This exemplifies the advantages of applying complex textual features in comparison to trivial word-like features for the sentence classification task.

Feature sets	NB	DT	ME	SVM
All-Words	0.68	0.60	0.70	0.69
Lemmatized-Tokens (A)	0.70	0.64	0.72	0.72
A+ Lexicon-Token-Match (B)	0.70	0.65	0.72	0.73
A+B+ Lemmatized-Token-Bigrams (C)	0.72	0.70	0.74	0.74
A+B+C+ Lemmatized-Token-Trigrams (D)	0.71	0.70	0.73	0.74
A+B+C+ Noun-Character-Affixes (E)	0.72	0.70	0.75	0.75
A+B+C+E+ Lemmatized-Verbs (F)	0.73	0.72	0.76	0.76
A+B+C+E+F+ Lemmatized-Tokens-In-Window (G)	0.73	0.73	0.76	0.76
A+B+C+E+F+G+ Stanford-Token-Dependencies (H)	0.74	0.73	0.77	0.76

Table 10.18: Performance evaluation of different classifiers in combination with different feature sets evaluated by 10-fold cross-validation. F_1 scores over the class **POSITIVE** are reported.

Performance Measure	All-Words	Morphosyntactic feat.
Overall Accuracy	0.86	0.91
Precision over class **POSITIVE**	0.69	**0.82**
Recall over class **POSITIVE**	0.53	**0.70**
F_1 score over class **POSITIVE**	0.60	**0.76**
Precision over class **NEGATIVE**	0.94	0.93
Recall over class **NEGATIVE**	0.89	0.96
F_1 score over class **NEGATIVE**	0.91	0.95
Macro-averaged F_1 score	0.76	0.85
Micro-averaged F_1 score	0.85	0.91

Table 10.19: Performance evaluation of sentence classification with Maximum Entropy classifier over the ADE-TEST set.

Sentence Classification with Entity-Binding

Previous experiments with the sentence classification (Tables 10.18 and 10.19) applied token-binding strategy for identifying the potential occurrences of drug and condition inferring tokens in the training and test sentences. Experiments were performed by replacing the system of token-binding with the entity-binding. ProMiner was applied with the DrugBank and MedDRA dictionaries for tagging the drug and condition names in the ADE-TRAIN and ADE-TEST sentences. Features were generated as described in Section 10.6.1 and the datasets were subjected to the performance evaluation with the Maximum Entropy classifier by cross validation over the training set and a final assessment over the test set. Table 10.20 provides a comparison of performances of evaluation with token-binding and entity-binding.

From Table 10.20 it can be observed that the performance of classification does not significantly differ between the token-binding and entity-binding. However, the token-binding strategy offers an advantage of skipping the named entity recognition to

	Token-binding			Entity-binding		
Dataset	Precision	Recall	F_1 score	Precision	Recall	F_1 score
ADE-TRAIN	0.83	0.72	0.77	0.83	0.72	0.77
ADE-TEST	0.82	0.70	0.76	0.82	0.72	0.77

Table 10.20: Comparison of classification performances with token-binding and entity-binding over the ADE-TRAIN (cross-validation) and ADE-TEST datasets. Performances over the class **POSITIVE** are indicated.

be performed before the sentence classification.

Upon examination of **POSITIVE** sentences classified in the ADE-TEST corpus with the token-binding strategy, 584 out of 829 sentences were correctly classified. Amongst the 584 true positive classified sentences, ProMiner was able to identify both classes of entities (*i.e.* drugs and conditions) in 502 sentences only. The remaining 82 sentences (constituting 14% of true positives) were correctly classified where NER failed to capture both classes of entities. This exemplifies the advantages of sentence classifier and its independence from the named entity recognition.

Sentence Classification with Ensemble Classifier

Experiences from the past have shown that using an ensemble of classifiers can perform better than applying a single classifier [Thomas et al. (2011)]. An ensemble classifier approach uses a voting scheme from the predictions of different classifiers to judge the final label of an instance. Individual classifiers (*i.e.* DT, NB, ME, and SVM) trained over the ADE-TRAIN sentences were applied over the ADE-TEST sentences and the final predictions were merged using a voting system. Three types of voting were tested *i.e. AtLeast One*, *AtLeast Two*, *AtLeast Three*. For instance the *AtLeast One* voting system judges a sentence in ADE-TEST as **POSITIVE** if any one of the applied four classifiers predicts the given sentence as **POSITIVE**. Table 10.21 shows the results of performance of classification over ADE-TEST dataset using an ensemble of classifiers with different voting schemes. It can be observed that the recall of classification can be improved with *Atleast One* voting scheme whereas the precision can be greatly improved with *AtLeast Three* voting scheme. Using *AtLeast Two* voting system generated similar results to applying the ME classifier alone. Experiments with *AtLeast Four* voting system generated zero F_1 score since no sentence was classified as **POSITIVE** by all four classifiers.

10.7.3 Recall Optimization by Instance Selection

The performance of system indicated good precision of 82% and moderate recall of 70%. Considering the requirements of final users (such as drug safety experts), experiments were performed in order to improve the recall of the sentence classification system. Considering the skewness of the training data, an instance selection approach was applied.

Voting	Precision	Recall	F_1 score
AtLeast One	0.56	0.88	0.68
AtLeast Two	0.80	0.74	0.77
AtLeast Three	0.90	0.49	0.63

Table 10.21: Performance evaluation over the ADE-TEST dataset using an ensemble of classifiers. Performances over the class **POSITIVE** are indicated.

Several approaches have been defined in the past to deal with unbalanced datasets such as the undersampling, oversampling, and many more [Kotsiantis et al. (2006)]. However, the approach proposed here applies the principles of undersampling and active learning to eliminate non-informative and noisy majority class instances (*i.e.* sentences labeled as **NEGATIVE**) from the training set. First, a random undersampling was performed to generate a balanced dataset of **POSITIVE** and **NEGATIVE** sentences. Later on, applying the principles of active learning, additional **NEGATIVE** sentences were systematically selected by iterative training and evaluation until a satisfactory criterion is attained. The process of undersampling and active learning for instance selection is shown in Algorithm 2. In the ADE-TRAIN, 3,443 **POSITIVE** sentences were used as-is and 3,443 **NEGATIVE** sentences were picked randomly to from a ADE-TRAIN-SEED training set. The remaining 9,912 unpicked **NEGATIVE** sentences formed a ADE-TRAIN-NEG set. Applying Algorithm 2, informative **NEGATIVE** sentences were systematically selected from the ADE-TRAIN-NEG and added to the ADE-TRAIN-SEED until a satisfactory performance of the system was observed.

Algorithm 2: Instance selection by undersampling and active learning

Require: Balanced training set ADE-TRAIN-SEED
Require: A set of **NEGATIVE** labeled sentences ADE-TRAIN-NEG
Require: The model M trained on ADE-TRAIN-SEED
Repeat
1. Apply the trained model classifier M on ADE-TRAIN-NEG
2. Rank the sentences of ADE-TRAIN-NEG that are misclassified as **POSITIVE** in decreasing value of P(**POSITIVE**)
3. Pick top ranked 100 sentences and add them to ADE-TRAIN-SEED
4. Train the model on the extended ADE-TRAIN-SEED and evaluate on ADE-TEST
Until the stopping criterion is reached

Figure 10.5 shows the performance of the system at different rounds of active learning. During each round of active learning, the preliminary feature set described in Section 10.6.1 was used. Altogether, 8 rounds of active learning were performed in order to observe the convergence in F_1 of the system. The process of active learning resulted in a reformed training corpus (referred to as ADE-TRAIN-AL) containing 7,386 sentences. Finally, it turned out that training over the reformed corpus (with 5 rounds of active learning) substantially increased the recall of the system upto 80% with an

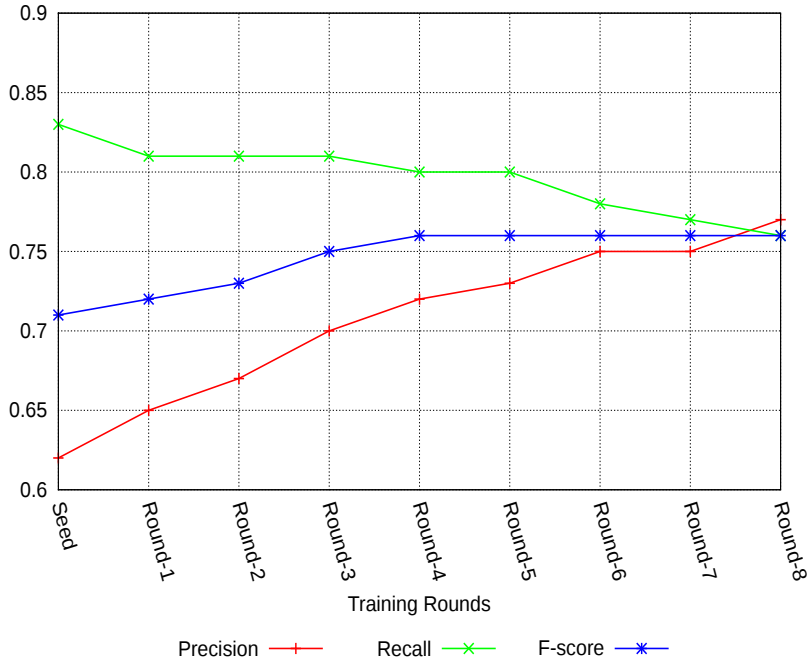

Figure 10.5: Results of the performance of the system attained during different rounds of instance selection by undersampling and active learning.

acceptable decline in the precision.

10.7.4 Error Analysis of Sentence Classification

An analysis was performed to determine the primary sources of errors generated by the sentence classification system. For this task, a subset of misclassified instances (both false positives and false negatives) were manually investigated to understand classification uncertainties. Few **NEGATIVE** labeled sentences that were classified as **POSITIVE** contained adverse effects in relation to certain classes of drugs. An instance of such a sentence is *PMID:16344532 The authors report two cases of catechol-O-methyltransferse (COMT) inhibitor-induced asymptomatic hepatic dysfunction in women with Parkinson disease*. Few misclassified sentences had co-reference to the adverse effect related to the drug without an explicit mentioning of the observed medical problem. An example of sentence belonging to this error class include *PMID:12523465 Until then, clinicians need to be aware of this possible complication associated with zonisamide*. Few sentences contained adverse effects associated with different forms of medical treatments. Examples include *PMID:16633932 We conclude that immunological dysfunction resulting from the thymectomy contributed significantly to the subsequent development of PRCA, SLE, IPH in this patient*.

Examples of instances misclassified as **NEGATIVE** include the sentences containing descriptive adverse effects often not contained in the lexicon. *PMID: 14607011 The case history and toxicological findings of an infant fatality involving pseudoephedrine, brompheni-*

ramine, and dextromethorphan are presented. Long sentences were also one source of errors. An example of misclassified lengthy sentence is *PMID: 7351000 Seven of the eight cases of acute leukemia occurred in a series of 553 patients treated with Treosulfan for ovarian cancer in the period from 1970 to 1977 and followed closely for a total of 1159 patient-years up to February 1978*. A couple of misclassified instances had sparsely defined association between drug and a medical problem. For instance, the sentence *"PMID:8979664 METHODS/RESULTS: This paper presents a new case of rifabutin uveitis and a review of the various published reports to date"* mentions uveitis occurring as a result of rifabutin administration described within a sparse context.

10.7.5 Retrospective Assessment of NER

ProMiner, a dictionary-based NER system, was employed for identifying the named entities. Curated DrugBank and MedDRA dictionaries were integrated into ProMiner and applied for the recognition of drugs and conditions respectively. An experiment was performed to examine if the co-occurrence of drug and condition would serve as a basis for adverse effect sentence classification that resulted in precision and recall of 0.60 and 0.62 respectively. Therefore, the sentence classification using co-occurrence criteria is not sufficient to differentiate between positive and negative sentence.

A recall of 0.82 and 0.73 for the partial recognition of drug and adverse effect entities respectively over the entire ADE corpus has been shown previously (see Section 10.4.3). Main reason for false negative entities is that several drugs as well as conditions are missing in the respective dictionaries. Examples of missing drug names include abbreviations (e.g. *5-FU*, *ARA-C*, etc.) and trivial names (such as *Suxamethonium*, *sulprostone*). Examples of unrecognized condition names include abbreviations and descriptive enumerations such as *decrease in peripheral blood leucocytes*, *t-AML*, etc. With an acceptable recall, the NER can be used to support automatic assignment of sentences to drug classes and medical conditions.

Additional experiments were performed to assess the performances of Disorder-Recognizer (*see* Chapter 5) and MeSH disease dictionary for the recognition of adverse effects in the ADE corpus. Table 10.22 provides an overview on the recall of different NER methods for the drug and adverse effect identification with complete match and partial match criteria. Although the CRF-based Disorder-Recognizer showed good results for adverse effect entity recognition, MedDRA dictionary-based approach has been applied for case-studies since MedDRA is used as a global standard for adverse effect classification and therefore using MedDRA ensures compliance with best practices applied in the drug safety research.

10.7.6 Use-Case Study of Adverse Effect Classification

The purpose of use-case study is to determine the usefulness of the developed framework for its ability to track undocumented or rare adverse effects. For this purpose, SIDER and Drug Information Online databases were used as a reference for documented

Entity Recognizer	Class	Complete	Partial
ProMiner (MedDRA)	AE	0.54	0.73
ProMiner (MeSH)	AE	0.39	0.60
Disorder-Recognizer	AE	0.59	0.80
DrugBank	Drug	0.79	0.82

Table 10.22: Results of assessments of different NER methods for drugs and adverse effects (AE) identification. Only the recall values have been reported.

No. of sentences in ADE-EXAM corpus	199,633
No. of sentences classified as **POSITIVE**	35,159
No. of sentences containing at least one drug & one condition	24,178
No. of unique drug names recognized (after normalization)	882
No. of unique condition names recognized (after normalization)	2,076

Table 10.23: Results of sentence classification and named entity recognition over the ADE-EXAM corpus.

drug-related adverse effects. The pre-trained Maximum Entropy-based sentence classifier (trained over ADE corpus) ensembled with morphosyntactic feature generation machinery was applied for classifying the sentences in ADE-EXAM corpus. The **POSITIVE** classified sentences were subjected to named entity detection module for recognizing the co-occurring drug and condition names. Table 10.23 shows the results of sentence classification and named entity recognition over the ADE-EXAM corpus.

Amongst the most frequently occurring drugs in the ADE-EXAM corpus, three drugs were manually selected based on the differences in their pharmacology and therapeutic application. Table 10.24 shows the analysis of adverse effect profiles of pre-selected drugs in the ADE-EXAM corpus. For three analyzed drugs, the sentence classification

Drug	Occurrences	Sentences	Conditions	Examples of Novel Associations
Methotrexate	837	349	181	PMID:1450620 A **malignant teratoma** was diagnosed in a 65-year-old asthmatic man 16 months after initiation of **methotrexate** therapy (15 mg per week).
Infliximab	362	185	122	PMID:17534091 **Coccidioidomycosis pneumonia** in a nonendemic area associated with **infliximab**.
Clozapine	360	210	102	PMID:16342008 **Guillain-Barré syndrome** after septicemia following **clozapine**-induced agranulocytosis. A case report.

Table 10.24: Adverse effect profile analysis of pre-selected drugs in the ADE-EXAM corpus. Columns two, three, and four indicate frequency of drug occurrence, no. of unique sentences containing the drug, and no. of normalized unique co-occurring conditions respectively.

framework helped in identifying novel drug-adverse effect associations that are not documented in standard databases. This indicates a potential application of the established framework in identifying the drug safety signals from unstructured textual data to support pharmacovigilance.

10.8 Discussion

This chapter outlines two major accomplishments for supporting automated pharmacovigilance that has been one of the most challenging issues associated with drug safety surveillance in the market. First, a systematically annotated and substantially large corpus of medical case reports was generated that can be used by medical NLP community for development of literature mining systems to support pharmacovigilance. Secondly, a dual-phased application has been developed and its potential has been demonstrated for the identification of adverse drug effect assertive sentences in medical case reports. The system also identifies co-occurring drug and condition names in positively implicated sentences. An evaluation of the system showed convincing results. An additional use-case study indicated a potential application of the system in detecting under-reported and under-documented adverse drug effects.

In future, the author plans to improve various components of the system like expansion of the medical lexicons and dictionaries used, feature space optimization to enhance the system's performance, application of the system over different text sources including full-text articles, consumer generated media (blogs and forum), e-health records, and benchmarking the system's capabilities against commercial adverse effect detection technologies (e.g. Luxid Skill Cartridges[8]).

Considering the real world application of sentence classifier, the system would be of high value when it can identify completely new cases of adverse effects or change the existing statistics of drug-adverse effect relationships. The system's ability to detect adverse effect assertions on drugs does not replace the manual task of drug safety monitoring but can greatly help in reducing the final reports that a safety expert needs to investigate. This can revolutionize the way in which the drug safety surveillance is performed by regulatory authorities such as the FDA, or EMEA.[9] The developed system is believed to accelerate the drug safety monitoring through automated knowledge extraction from text and therefore facilitate higher adherence to the daunting task of automated pharmacovigilance, faster response time and better service to the patients from the viewpoints of both pharmaceutical industries and health regulatory agencies.

[8] http://www.temis.com/

[9] http://www.fda.gov/MedicalDevices/ScienceandResearch/ucm243158.htm

Chapter 11

Conclusion and Perspectives

11.1 Conclusion

Free-text, especially in the medical domain, is a rich and important resource that encompasses information about novel or unseen scientific findings. They also act on protection of intellectual property in the form of patents. In general, the study of literature is inevitable for the generation of new knowledge, hypothesis development, and to update the most recent scientific developments. From an industrial perspective, the study of literature fosters the product development strategies whereas in the clinical settings it helps in quick monitoring of patient profiles and therefore improve the overall effectiveness of clinical care and safety. However, considering the sheer amount of free-text data that is available and generated in various forms, automatic strategies for efficient processing these data have become immensely crucial. Therefore, this work focuses on the development of efficient strategies for mining the medical and patent literature with an aim of supporting the healthcare and pharmacovigilance research.

Firstly, this thesis focused on the development of techniques for the recognition of disease and adverse effect (collectively called as medical disorders) named entities in free-text and their systematic evaluation. In this context, several state-of-the-art medical terminologies (such as MeSH, MedDRA, etc.) were tested for their ability to support dictionary-based named entity recognition of medical disorders in scientific abstracts. An outcome of this assessment indicated MedDRA to be the best suited resource for the disorder recognition. Unlike the MeSH, UMLS, or ICD, MedDRA is less frequently used resource for biomedical named entity recognition and this survey spotlighted the competencies of MedDRA as a valuable resource. Later on, a machine learning technique based on CRF was adapted and evaluated for the recognition of medical disorders that indicated good results. A comparative assessment of performance of the implemented CRF-based Disorder-Recognizer against several other state-of-the-art entity recognition systems showed highly competitive results. Furthermore, the ability of the entity recognition system was expanded to test its adaptability to work on patient health records. The system was trained for the recognition of medical disorders, treatments, and tests in e-health records. An open evaluation during the public assessment (*i.e.* I2B2 challenge 2010) indicated superior results. The system was ranked fourth in comparison to various other competing systems [Uzuner et al. (2011)]. This demonstrates the capabilities and competence of the so-developed strategies for the

recognition of medical entities and concepts in different sources of free-text documents. Furthermore, the identification of assertions made over medical problems in free-text is essential for the development of precise medical IR or IE systems. In the context of I2B2 challenge 2010, a SVM-based system was implemented and systematically optimized for the classification of assertions in e-health records. An evaluation of the system indicated good performance with the F_1 score of 0.90. Comparative evaluation of the implemented assertion classification approach during the I2B2 2010 showed highly competitive results.

Based on the successful scenarios demonstrated for the identification of concepts in various document sources, their capabilities were systematically exploited for developing a semantic information retrieval platform. Therefore, in the context of TREC challenges, this thesis focused on implementing a scalable retrieval platform that was systematically customized for information retrieval from e-health records, and patents. The named entity or concept recognition techniques were applied for tagging the important concepts in document collections that were used as a backbone for performing semantic searches. During the TRECMED, the system was optimized for the retrieval of e-health records. Medical concepts, relations, and assertions were systematically indexed. An open evaluation of the adapted semantic search strategy showed results that were highly competitive to state-of-the-art technologies. Evaluations during the TREC-CHEM 2010 and 2011, showed the superiority of the applied system in efficient patent retrieval based on semantic concepts. The system achieved top results in comparison to other competent retrieval systems during both evaluations. This demonstrates that the integration of so-developed concept recognition techniques, and in-house retrieval engine SCAIVIEW can deliver highly efficient and precise environment for information retrieval from patents, and e-health records that can outperform conventional document search paradigm.

For the identification of adverse effects of medicinal drugs, this thesis focused on development of a strategy to support automated pharmacovigilance wherein a sentence classification system was developed. The system is based on the Maximum Entropy classifier and it utilizes automatically extracted lexico-syntactic features from text for sentence classification. A quantitative and qualitative evaluation of the system indicated robust results. Competencies of the concept recognition approaches where applied for spotting the occurrences of drugs and disorders in sentences asserting an adverse event. Studies on real use-case scenarios demonstrated its ability to identify novel side effect associations. An in-house as well as expert evaluation demonstrated the system's ability to efficiently promote the pharmacovigilance research.

Finally, the development of any efficient IE or IR system needs systematically annotated corpora. Different sets of corpora for development and evaluation of disorder recognition techniques and adverse event identification have been generated and made publicly available. The generated corpora are believed to promote the research in the direction of disorder recognition and drug safety detection by promoting the development, optimization, and evaluation of automatic approaches.

Summarizing the accomplishments within this thesis, following enablements are desirable:

> Ability to identify medical concepts and assertions in different document sources with high sensitivity and specificity can assist quick document look-up, spotlight informative snippets (*e.g.* sentence with drug and disease co-occurrence). It can apparently support the development of semantic search engine, relation extraction, and knowledge discovery more efficiently.
>
> Semantic platform that enables searching within medical records, document repositories, and patents can support medical information retrieval, evidence-based practices, clinical decision making, prior art search, and infringement searches in patents.
>
> Pharmacovigilance supporting system that can identify sentences asserting adverse drug events can promote the development of alerting systems, and support the generation of drug safety warning and signals (unseen adverse effects).
>
> Corpora for disorder entity recognition, and adverse effect detection can drive the research and development of new approaches or optimize the existing solutions. Ideally, it can promote corpus re-annotation with new information.

Strategies implemented during this thesis have been evaluated at various stages and have demonstrated success during open assessments such as TREC and I2B2. This work is believed to enhance the biomedical and clinical text mining scenarios. This can ultimately promote effective patient healthcare and safety, and improve compliance with best clinical practices, and accelerate promising services to the patients.

11.2 Future Perspectives

Apart from various success stories demonstrated during this thesis, several challenging aspects and issues paves way for the future research. The investigation of the disorder named entity recognition by ProMiner shows that the existing medical terminologies are insufficient to cover the entire space of medical disorder mentions in text. Although, the machine learning-based technique (CRF) demonstrated success in this task, this technique suffers from entity normalization problems. Although solutions are available for the normalization of CRF to standard terminological entries, this approach still is strongly dependent on the dictionary coverage. In the current work, performances of publicly available disorder-centric terminological resources have been tested. In future, a systematic evaluation of commercial resources needs to be performed. A strategic combination of terminologies from different sources such as MedDRA, or MeSH may also contribute to the betterment of the terminological coverage. However, this is not a

trivial task since most of the terminologies are hierarchically organized and merging the knowledge from these resources needs certain degree of medical expert intervention as well as dedicated techniques such as ontology alignment or hierarchy alignment.

The document indexing and retrieval platform (SCAIVIEW) has been successfully demonstrated for its ability to efficiently retrieve patents and e-health records. Although, evaluations have been performed under standard test conditions provided by TREC conference, it is important to demonstrate few real use-case scenarios to help medical professionals or patent searchers. This can be made possible by scientific collaboration with patent user groups or medical professionals who could use the retrieval system for real use-case question answering. Currently, the document collections over which SCAIVIEW system works are scientific corpus subsets and it is essential to operate the retrieval system with larger collections of documents. Evaluation of the performance of retrieval over different text collections such as abstracts, blogs, or forums is another interesting issue. From the technical point of view, although standard retrieval functions have been systematically implemented and successfully applied for the retrieval task, various parallel approaches such as document clustering, topic modeling, and latent semantic indexing can be tested.

The author proposes several improvements that can be tested in future for automatic adverse effect detection. The current implemented system recognizes sentences that contain information about adverse drug events. In future, it is desirable to extend the capabilities of the system to detect exact related pairs (drug-adverse effect pairs). Current system assumes that drugs and adverse effects co-occur in the same sentence and that is true for majority of cases. Nevertheless, using anaphora resolution to detect non-sentence level existing drug-adverse events would add improvement to the overall recall of the system. Evaluating the performance of the developed system in comparison to commercial tools such as LUXID skill cartridges would improve the overall applicability of the existing model. Development of an alerting system that can generate instant and timely alerts about adverse drug events published in different resources can find direct applications for drug safety monitoring.

Finally, the author believes that the outcome of this work can find potential applications in the biomedical and clinical settings to support efficient literature searches across various sources such as scientific articles, medical text, and patents. It has an ability to revolutionize the literature search strategies followed within the biomedical, clinical, and pharmaceutical settings. The developed strategies for information retrieval and information extraction can support evidence-based medical practices that can lead to an advancement of medical research, improve the quality of healthcare, and enhance the patient safety. The author desires to bring the applied methodologies into deliverable solutions that can be applied for daily usage to support biomedical and clinical professionals. Improving various functionalities of the developed methods and providing user-friendly literature search environments to support real use cases scenarios is the future motto.

Appendix A

TREC Topics

A.1 Topics used for technology survey search in patents.

Topic	Title
TS-2	Dipeptidyl peptidase-IV inhibitors
TS-12	Diazepam or RN: 439-14-5
TS-13	Tetrahydrocannabinol as an anti-tumor agent
TS-15	Betaines for peripheral artery disease
TS-20	Tests for HCG hormones
TS-22	Uses of hormones in detection of menopause
TS-28	D-ala-D-ala ligase inhibitors
TS-29	Inhibitors of acetylcholinesterase
TS-33	Respiratory tract disorders using inhalation of porous particles containing amino acid and endogenous phospholipid

Table A.1: Topic IDs and their titles used for the TRECCHEM technology survey task.

A.2 Topics used for searching in e-health records.

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

Table A.2: Topic IDs and their titles used for the TRECMED task.

Bibliography

[Ahmad 2003] Ahmad, Syed R.: Adverse drug event monitoring at the Food and Drug Administration. In: *J Gen Intern Med* 18 (2003), Jan, No. 1, pp. 57–60.

[Ahmed et al. 2011] Ahmed, Jessica; Meinel, Thomas; Dunkel, Mathias; Murgueitio, Manuela S.; Adams, Robert; Blasse, Corinna; Eckert, Andreas; Preissner, Saskia; Preissner, Robert: CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. In: *Nucleic Acids Res* 39 (2011), Jan, No. Database issue, pp. D960–D967. – URL http://dx.doi.org/10.1093/nar/gkq910.

[Arita 2011] Arita, Isao: A personal recollection of smallpox eradication with the benefit of hindsight: in commemoration of 30th anniversary. In: *Jpn J Infect Dis* 64 (2011), No. 1, pp. 1–6.

[Aronson 1999] Aronson, A. R.: Filtering the UMLS Metathesaurus for MetaMap. / National Library of Medicine. URL http://skr.nlm.nih.gov/papers/references/filtering99.pdf, 1999. – Technical Report.

[Aronson 2001] Aronson, A. R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: *Proc AMIA Symp* (2001), pp. 17–21.

[Arrow et al. 2009] Arrow, Kenneth; Auerbach, Alan; Bertko, John; Brownlee, Shannon; Casalino, Lawrence P.; Cooper, Jim; Crosson, Francis J.; Enthoven, Alain; Falcone, Elizabeth; Feldman, Robert C.; Fuchs, Victor R.; Garber, Alan M.; Gold, Marthe R.; Goldman, Dana; Hadfield, Gillian K.; Hall, Mark A.; Horwitz, Ralph I.; Hooven, Michael; Jacobson, Peter D.; Jost, Timothy S.; Kotlikoff, Lawrence J.; Levin, Jonathan; Levine, Sharon; Levy, Richard; Linscott, Karen; Luft, Harold S.; Mashal, Robert; McFadden, Daniel; Mechanic, David; Meltzer, David; Newhouse, Joseph P.; Noll, Roger G.; Pietzsch, Jan B.; Pizzo, Philip; Reischauer, Robert D.; Rosenbaum, Sara; Sage, William; Schaeffer, Leonard D.; Sheen, Edward; Silber, B M.; Skinner, Jonathan; Shortell, Stephen M.; Thier, Samuel O.; Tunis, Sean; Wulsin, Lucien; Yock, Paul; Nun, Gabi B.; Bryan, Stirling; Luxenburg, Osnat; van de Ven, Wynand P M M.: Toward a 21st-century health care system: recommendations for health care reform. In: *Ann Intern Med* 150 (2009), Apr, No. 7, pp. 493–495.

[Barakat et al. 2010] Barakat, Khaled; Mane, Jonathan; Friesen, Douglas; Tuszynski, Jack: Ensemble-based virtual screening reveals dual-inhibitors for the p53-MDM2/MDMX interactions. In: *J Mol Graph Model* 28 (2010), Feb, No. 6, pp. 555–568. – URL http://dx.doi.org/10.1016/j.jmgm.2009.12.003.

[Barakat and Tuszynski 2011] Barakat, Khaled; Tuszynski, Jack: Relaxed complex scheme suggests novel inhibitors for the lyase activity of DNA polymerase beta. In: *J Mol Graph Model* 29 (2011), Feb, No. 5, pp. 702–716. – URL http://dx.doi.org/10.1016/j.jmgm.2010.12.003.

[Barakat et al. 2009] Barakat, Khaled H.; Torin Huzil, J.; Luchko, Tyler; Jordheim, Lars; Dumontet, Charles; Tuszynski, Jack: Characterization of an inhibitory dynamic pharmacophore for the ERCC1-XPA interaction using a combined molecular dynamics and virtual screening approach. In: *J Mol Graph Model* 28 (2009), Sep, No. 2, pp. 113–130. – URL http://dx.doi.org/10.1016/j.jmgm.2009.04.009.

[Bates et al. 2003] Bates, David W.; Evans, R S.; Murff, Harvey; Stetson, Peter D.; Pizziferri, Lisa; Hripcsak, George: Detecting adverse events using information technology. In: *J Am Med Inform Assoc* 10 (2003), No. 2, pp. 115–128.

[Ben Abacha and Zweigenbaum 2011] Ben Abacha, Asma; Zweigenbaum, Pierre: Automatic extraction of semantic relations between medical entities: a rule based approach. In: *J Biomed Semantics* 2 Suppl 5 (2011), pp. S4. – URL http://dx.doi.org/10.1186/2041-1480-2-S5-S4.

[Boehning 1992] Boehning, Dankmar: Multinomial logistic regression algorithm. In: *Annals of the Institute of Statistical Mathematics* 44 (1992), pp. 197–200.

[Breiman et al. 1984] Breiman, Leo; Friedman, Jerome; Olshen, R. A.; Stone, Charles: *Classification and regression trees*. Wadsworth and Brooks/Cole Advanced Books and Software, 1984.

[Brill 1992] Brill, Eric: A simple rule-based part of speech tagger. In: *Proceedings of the third conference on applied natural language processing*, 1992, pp. 152–155.

[Brown et al. 1999] Brown, E. G.; Wood, L.; Wood, S.: The medical dictionary for regulatory activities (MedDRA). In: *Drug Saf* 20 (1999), Feb, No. 2, pp. 109–117.

[Brown et al. 2009] Brown, M.; Dunn, W. B.; Dobson, P.; Patel, Y.; Winder, C. L.; Francis-McIntyre, S.; Begley, P.; Carroll, K.; Broadhurst, D.; Tseng, A.; Swainston, N.; Spasic, I.; Goodacre, R.; Kell, D. B.: Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. In: *Analyst* 134 (2009), Jul, No. 7, pp. 1322–1332. – URL http://dx.doi.org/10.1039/b901179j.

[Browne et al. 2003] Browne, Allen C.; Divita, Guy; Aronson, Alan R.; McCray, Alexa T.: UMLS language and vocabulary tools. In: *AMIA Annu Symp Proc* (2003), pp. 798.

[Bruijn et al. 2010] Bruijn, Berry d.; Cherry, Colin; Kiritchenko, Svetlana; Martin, Joel; Zhu, Xiaodan: NRC at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. In: *Proceedings*

of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data, 2010.

[Buckley and Voorhees 2004] Buckley, Chris; Voorhees, Ellen: Retrieval Evaluation with Incomplete Information. In: *Proceedings of Special Interest Group on Information Retrieval*, 2004.

[Bundschus et al. 2008] Bundschus, Markus; Dejori, Mathaeus; Stetter, Martin; Tresp, Volker; Kriegel, Hans-Peter: Extraction of semantic biomedical relations from text using conditional random fields. In: *BMC Bioinformatics* 9 (2008), pp. 207. – URL http://dx.doi.org/10.1186/1471-2105-9-207.

[Burns 2006] Burns, Phillip: *MorphAdorner: Morphological Adorner for English Text*. 2006. – URL http://morphadorner.northwestern.edu/.

[Campillos et al. 2008] Campillos, Monica; Kuhn, Michael; Gavin, Anne-Claude; Jensen, Lars J.; Bork, Peer: Drug target identification using side-effect similarity. In: *Science* 321 (2008), Jul, No. 5886, pp. 263–266. – URL http://dx.doi.org/10.1126/science.1158140.

[Can and Baykal 2007] Can, Aysu B.; Baykal, Nazife: MedicoPort: a medical search engine for all. In: *Comput Methods Programs Biomed* 86 (2007), Apr, No. 1, pp. 73–86. – URL http://dx.doi.org/10.1016/j.cmpb.2007.01.007.

[Ceusters et al. 2008] Ceusters, Werner; Capolupo, Maria; De Moor, Georges; Devlies, Jos: *Various Views on Adverse Events: a collection of definitions.* 2008. – URL http://org.buffalo.edu/RTU/papers/AdverseEventDefs.pdf.

[Chang et al. 2010] Chang, E; Xu, Y; Hong, K: A hybrid approach to extract structured information from narrative clinical discharge summaries. In: *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, 2010.

[Chapman et al. 2001] Chapman, W. W.; Bridewell, W.; Hanbury, P.; Cooper, G. F.; Buchanan, B. G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. In: *J Biomed Inform* 34 (2001), Oct, No. 5, pp. 301–310. – URL http://dx.doi.org/10.1006/jbin.2001.1029.

[Chen et al. 2007] Chen, Xin; Fang, Ying; Yao, Lixia; Chen, Yuzong; Xu, Huan: Does drug-target have a likeness? In: *Methods Inf Med* 46 (2007), No. 3, pp. 360–366. – URL http://dx.doi.org/10.1160/ME0425.

[Chiang et al. 2010] Chiang, Jung-Hsien; Lin, Jou-Wei; Yang, Chen-Wei: Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE). In: *J Am Med Inform Assoc* 17 (2010), No. 3, pp. 245–252. – URL http://dx.doi.org/10.1136/jamia.2009.000182.

[Choi et al. 2010] Choi, Jooyoung; Davis, Melissa J.; Newman, Andrew F.; Ragan, Mark A.: A semantic web ontology for small molecules and their biological targets. In: *J Chem Inf Model* 50 (2010), May, No. 5, pp. 732–741. – URL http://dx.doi.org/10.1021/ci900461j.

[Clark et al. 2010] Clark, C; Aberdeen, J; Coarr, M: Determining assertion status for medical problems in clinical records. In: *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, 2010.

[Cohen and Hersh 2005] Cohen, Aaron M.; Hersh, William R.: A survey of current work in biomedical text mining. In: *Brief Bioinform* 6 (2005), Mar, No. 1, pp. 57–71.

[Cohen and Hunter 2004] Cohen, Bretonnel; Hunter, Lawrence: Natural Language Processing and Systems Biology. In: *Artificial Intelligence Methods and Tools for Systems Biology*, 2004, pp. 147–173.

[Cohen and Hunter 2008] Cohen, K B.; Hunter, Lawrence: Getting started in text mining. In: *PLoS Comput Biol* 4 (2008), Jan, No. 1, pp. e20. – URL http://dx.doi.org/10.1371/journal.pcbi.0040020.

[Cohn et al. 1994] Cohn, David; Ladner, Richard; Waibel, Alex: Improving generalization with active learning. In: *Machine Learning* 15 (1994), pp. 201–221.

[Conde-Pueyo et al. 2009] Conde-Pueyo, Nuria; Munteanu, Andreea; Solé, Ricard V.; Rodríguez-Caso, Carlos: Human synthetic lethal inference as potential anti-cancer target gene detection. In: *BMC Syst Biol* 3 (2009), pp. 116. – URL http://dx.doi.org/10.1186/1752-0509-3-116.

[Corney et al. 2004] Corney, David P A.; Buxton, Bernard F.; Langdon, William B.; Jones, David T.: BioRAT: extracting biological information from full-length papers. In: *Bioinformatics* 20 (2004), Nov, No. 17, pp. 3206–3213. – URL http://dx.doi.org/10.1093/bioinformatics/bth386.

[Cover and Hart 1967] Cover, Thomas; Hart, Peter: Nearest Neighbor Pattern Classification. In: *IEEE Transactions on Information Theory* Vol. 13, 1967.

[Crammer et al. 2006] Crammer, Koby; Dekel, Ofer; Keshet, Joseph; Shalev-Shwartz, Shai; Singer, Yoram: Online Passive-Aggressive Algorithms. In: *Journal of Machine Learning Research* 7 (2006), pp. 551–585.

[Cunningham et al. 2002] Cunningham, H.; Maynard, D.; Bontcheva, K.; Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: *Proceedings of the 40th Annual Meeting of the ACL*, 2002.

[Côté and Robboy 1980] Côté, R. A.; Robboy, S.: Progress in medical information management. Systematized nomenclature of medicine (SNOMED). In: *JAMA* 243 (1980), No. 8, pp. 756–762.

[Demner-Fushman et al. 2011] Demner-Fushman, Dina; Abhyankar, Swapna; Jimeno-Yepes, Antonio; Loane, Russell; Rance, Bastien; Lang, Francois; Ide, Nicholas; Apostolova, Emilia; Aronson, Alan: A knowledge-based approach to medical records retrieval. In: *Notebook Proceedings of Twenteith Text Retrieval Conference*, 2011.

[Demner-Fushman et al. 2009] Demner-Fushman, Dina; Chapman, Wendy W.; McDonald, Clement J.: What can natural language processing do for clinical decision support? In: *J Biomed Inform* 42 (2009), Oct, No. 5, pp. 760–772. – URL http://dx.doi.org/10.1016/j.jbi.2009.08.007.

[Denecke 2008] Denecke, K.: Semantic structuring of and information extraction from medical documents using the UMLS. In: *Methods Inf Med* 47 (2008), No. 5, pp. 425–434.

[Denny et al. 2003] Denny, Joshua C.; Smithers, Jeffrey D.; Miller, Randolph A.; Spickard, Anderson: "Understanding" medical school curriculum content using KnowledgeMap. In: *J Am Med Inform Assoc* 10 (2003), No. 4, pp. 351–362. – URL http://dx.doi.org/10.1197/jamia.M1176.

[Dietmann et al. 2009] Dietmann, Sabine; Georgii, Elisabeth; Antonov, Alexey; Tsuda, Koji; Mewes, Hans-Werner: The DICS repository: module-assisted analysis of disease-related gene lists. In: *Bioinformatics* 25 (2009), Mar, No. 6, pp. 830–831. – URL http://dx.doi.org/10.1093/bioinformatics/btp055.

[Divita et al. 2006] Divita, Guy; Browne, Allen C.; Loane, Russell: dTagger: a POS tagger. In: *AMIA Annu Symp Proc* (2006), pp. 200–203.

[Doukas et al. 2010] Doukas, Charalampos; Goudas, Theodosis; Fischer, Simon; Mierswa, Ingo; Chatziioannou, Aristotle; Maglogiannis, Ilias: An open data mining framework for the analysis of medical images: application on obstructive nephropathy microscopy images. In: *Conf Proc IEEE Eng Med Biol Soc* 2010 (2010), pp. 4108–4111. – URL http://dx.doi.org/10.1109/IEMBS.2010.5627332.

[Efron 1969] Efron, Bradley: Student's t-test under symmetry conditions. In: *Journal of the American Statistical Association* 64 (1969), pp. 1278–1302.

[Efron 1979] Efron, Bradley: Bootstrap Methods: Another Look at the Jackknife. In: *The Annals of Statistics* 7 (1979), pp. 1–26.

[Ellner and Joyner 2012] Ellner, Scott J.; Joyner, Paul W.: Information technologies and patient safety. In: *Surg Clin North Am* 92 (2012), Feb, No. 1, pp. 79–87. – URL http://dx.doi.org/10.1016/j.suc.2011.11.002.

[Embi et al. 2005] Embi, Peter J.; Jain, Anil; Clark, Jeffrey; Harris, C M.: Development of an electronic health record-based Clinical Trial Alert system to enhance recruitment at the point of care. In: *AMIA Annu Symp Proc* (2005), pp. 231–235.

[Engelson and Dagan 1996] Engelson, Sean; Dagan, Ido: Minimizing manual annotation cost in supervised training from corpora. In: *Proceeding ACL '96 Proceedings of the 34th annual meeting on Association for Computational Linguistics*, 1996.

[Facult et al. 1992] Facult, Hans P.; Paulussen, Hans; Martin, Willy: Dilemma-2: A Lemmatizer-Tagger For Medical Abstracts . In: *Third Conference on Applied Language Processing*, 1992.

[Farkas and Szarvas 2008] Farkas, Richárd; Szarvas, György: Automatic construction of rule-based ICD-9-CM coding systems. In: *BMC Bioinformatics* 9 Suppl 3 (2008), pp. S10. – URL http://dx.doi.org/10.1186/1471-2105-9-S3-S10.

[Forney 1973] Forney, David: The viterbi algorithm. In: *Proceedings of the IEEE*, 1973.

[Forster et al. 2005] Forster, Alan J.; Andrade, Jason; van Walraven, Carl: Validation of a discharge summary term search method to detect adverse events. In: *J Am Med Inform Assoc* 12 (2005), No. 2, pp. 200–206. – URL http://dx.doi.org/10.1197/jamia.M1653.

[Freund et al. 1997] Freund, Yoav; Seung, Sebastian; Shamir, Eli; Tishby, Naftali: Selective Sampling Using the Query by Committee Algorithm. In: *Machine Learning* 28 (1997), pp. 133–168.

[Friedman et al. 2004] Friedman, Carol; Shagina, Lyudmila; Lussier, Yves; Hripcsak, George: Automated encoding of clinical documents based on natural language processing. In: *J Am Med Inform Assoc* 11 (2004), No. 5, pp. 392–402. – URL http://dx.doi.org/10.1197/jamia.M1552.

[Friedman et al. 1996] Friedman, Carol; Shagina, Lyudmila; Socratous, Socrates; Zeng, Xiao: A WEB-Based Version of MedLEE: A Medical Language Extraction and Encoding System. In: *Proc AMIA Annu Fall Symp*, 1996, pp. 938.

[Friedrich et al. 2008] Friedrich, Christoph M.; Dach, Holger; Gattermayer, Tobias; Engelbrecht, Gerhard; Benkner, Siegfried; Hofmann-Apitius, Martin: @neuLink: a service-oriented application for biomedical knowledge discovery. In: *Stud Health Technol Inform* 138 (2008), pp. 165–172.

[Frolkis et al. 2010] Frolkis, Alex; Knox, Craig; Lim, Emilia; Jewison, Timothy; Law, Vivian; Hau, David D.; Liu, Phillip; Gautam, Bijaya; Ly, Son; Guo, An C.; Xia, Jianguo; Liang, Yongjie; Shrivastava, Savita; Wishart, David S.: SMPDB: The Small Molecule Pathway Database. In: *Nucleic Acids Res* 38 (2010), Jan, No. Database issue, pp. D480–D487. – URL http://dx.doi.org/10.1093/nar/gkp1002.

[Fundel et al. 2007] Fundel, Katrin; Küffner, Robert; Zimmer, Ralf: RelEx–relation extraction using dependency parse trees. In: *Bioinformatics* 23 (2007), Feb, No. 3, pp. 365–371. – URL http://dx.doi.org/10.1093/bioinformatics/btl616.

[Gao et al. 2008] Gao, Zhenting; Li, Honglin; Zhang, Hailei; Liu, Xiaofeng; Kang, Ling; Luo, Xiaomin; Zhu, Weiliang; Chen, Kaixian; Wang, Xicheng; Jiang, Hualiang: PDTD: a web-accessible protein database for drug target identification. In: *BMC Bioinformatics* 9 (2008), pp. 104. – URL http://dx.doi.org/10.1186/1471-2105-9-104.

[Garcia 2006] Garcia, E.: An information retrieval tutorial on cosine similarity measures, dot products and term weight calculations. 2006. – Technical Report.

[Giuliano et al. 2007] Giuliano, Claudio; Lavelli, Alberto; Pighin, Daniele; Romano, Lorenza: FBK-IRST: Kernel Methods for Semantic Relation Extraction. In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 2007.

[Gobeill et al. 2009] Gobeill, Julien; Teodoro, Douglas; Patsche, E; Ruch, Patrick: Report on the TREC 2009 Experiments: Chemical IR Track. In: *The Proceedings of Eighteenth Text RETrieval Conference (TREC 2009)*, 2009.

[Gold et al. 2008] Gold, Sigfried; Elhadad, Noémie; Zhu, Xinxin; Cimino, James J.; Hripcsak, George: Extracting structured medication event information from discharge summaries. In: *AMIA Annu Symp Proc* (2008), pp. 237–241.

[Goodwin et al. 2011] Goodwin, Travis; Rink, Bryan; Roberts, Kirk; Harabagiu, Sanda: Cohort Shepherd: Discovering Cohort Traits from Hospital Visits. In: *Notebook Proceedings of Twenteith Text Retrieval Conference*, 2011.

[Grinberg et al. 1995] Grinberg, Dennis; Lafferty, John; Sleator, Daniel: A robust parsing algorithm for link grammars. In: *Proceedings of the Fourth International Workshop on Parsing Technologies*, 1995.

[Gurulingappa et al. 2010] Gurulingappa, Harsha; Klinger, Roman; Hofmann-Apitius, Martin; Fluck, Juliane: An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature. In: *2nd Workshop on Building and evaluating resources for biomedical text mining (7th edition of the Language Resources and Evaluation Conference)*, 2010.

[Gurulingappa et al. 2009] Gurulingappa, Harsha; Müller, Bernd; Klinger, Roman; Mevissen, Heinz-Theo; Hofmann-Apitius, Martin; Fluck, Juliane; Friedrich, Christoph M.: Patent Retrieval in Chemistry based on semantically tagged Named Entities. In: *The Eighteenth Text RETrieval Conference (TREC 2009) Proceedings*, 2009.

[Hale 2005] Hale, Roger: Text mining: getting more value from literature resources. In: *Drug Discov Today* 10 (2005), Mar, No. 6, pp. 377–379. – URL http://dx.doi.org/10.1016/S1359-6446(05)03409-4.

[Hamon and Grabar 2010] Hamon, Thierry; Grabar, Natalia: Linguistic approach for identification of medication names and related information in clinical narratives. In: *J Am Med Inform Assoc* 17 (2010), No. 5, pp. 549–554. – URL http://dx.doi.org/10.1136/jamia.2010.004036.

[Hanauer 2006] Hanauer, David A.: EMERSE: The Electronic Medical Record Search Engine. In: *AMIA Annu Symp Proc* (2006), pp. 941.

[Hanisch et al. 2005] Hanisch, Daniel; Fundel, Katrin; Mevissen, Heinz-Theodor; Zimmer, Ralf; Fluck, Juliane: ProMiner: rule-based protein and gene entity recognition. In: *BMC Bioinformatics* 6 Suppl 1 (2005), pp. S14. – URL http://dx.doi.org/10.1186/1471-2105-6-S1-S14.

[Harkema et al. 2009] Harkema, Henk; Dowling, John N.; Thornblade, Tyler; Chapman, Wendy W.: ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. In: *J Biomed Inform* 42 (2009), Oct, No. 5, pp. 839–851. – URL http://dx.doi.org/10.1016/j.jbi.2009.05.002.

[Harmston et al. 2010] Harmston, Nathan; Filsell, Wendy; Stumpf, Michael P H.: What the papers say: text mining for genomics and systems biology. In: *Hum Genomics* 5 (2010), Oct, No. 1, pp. 17–29.

[Hauben and Bate 2009] Hauben, M.; Bate, A.: Decision support methods for the detection of adverse events in post-marketing data. In: *Drug Discov Today* 14 (2009), Apr, No. 7-8, pp. 343–357. – URL http://dx.doi.org/10.1016/j.drudis.2008.12.012.

[Hawizy et al. 2011] Hawizy, Lezan; Jessop, David M.; Adams, Nico; Murray-Rust, Peter: ChemicalTagger: A tool for semantic text-mining in chemistry. In: *J Cheminform* 3 (2011), No. 1, pp. 17. – URL http://dx.doi.org/10.1186/1758-2946-3-17.

[Haynes et al. 2010] Haynes, R B.; Wilczynski, Nancy L.; , Computerized Clinical Decision Support System (C. C. D. S. S) Systematic Review T. : Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: methods of a decision-maker-researcher partnership systematic review. In: *Implement Sci* 5 (2010), pp. 12. – URL http://dx.doi.org/10.1186/1748-5908-5-12.

[Hersh and Bhupatiraju 2003] Hersh, William; Bhupatiraju, Ravi T.: TREC Genomics Track Overview. In: *Proceedings of the Text REtrieval Conference*, 2003.

[Hettne et al. 2009] Hettne, Kristina M.; Stierum, Rob H.; Schuemie, Martijn J.; Hendriksen, Peter J M.; Schijvenaars, Bob J A.; Mulligen, Erik M v.; Kleinjans, Jos; Kors, Jan A.: A dictionary to identify small molecules and drugs in free text. In: *Bioinformatics* 25 (2009), Nov, No. 22, pp. 2983–2991. – URL http://dx.doi.org/10.1093/bioinformatics/btp535.

[Hirschman et al. 2005] Hirschman, Lynette; Yeh, Alexander; Blaschke, Christian; Valencia, Alfonso: Overview of BioCreAtIvE: critical assessment of information extraction for biology. In: *BMC Bioinformatics* 6 Suppl 1 (2005), pp. S1. – URL http://dx.doi.org/10.1186/1471-2105-6-S1-S1.

[Hobbs 2002] Hobbs, Jerry R.: Information extraction from biomedical text. In: *J Biomed Inform* 35 (2002), Aug, No. 4, pp. 260–264.

[Hofmann et al. 2008] Hofmann, Thomas; Scholkopf, Bernhard; Smola, Alexander: Kernel Methods in Machine Learning. In: *The Annals of Statistics* 36 (2008), pp. 1171–1220.

[Hofmann-Apitius et al. 2008] Hofmann-Apitius, Martin; Fluck, Juliane; Furlong, Laura; Fornes, Oriol; Kolárik, Corinna; Hanser, Susanne; Boeker, Martin; Schulz, Stefan; Sanz, Ferran; Klinger, Roman; Mevissen, Theo; Gattermayer, Tobias; Oliva, Baldo; Friedrich, Christoph M.: Knowledge environments representing molecular entities for the virtual physiological human. In: *Philos Transact A Math Phys Eng Sci* 366 (2008), Sep, No. 1878, pp. 3091–3110. – URL `http://dx.doi.org/10.1098/rsta.2008.0099`.

[Hsu et al. 2010] Hsu, Chih-Wei; Chang, Chih-Chung; Lin, Chih-Jen: *A Practical Guide to Support Vector Classification*. April 2010.

[Huang et al. 2003] Huang, Jin; Lu, Jingjing; Ling, Charles: Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy. In: *Third IEEE International Conference on Data Mining*, 2003, pp. 553 – 556.

[Härmark and van Grootheest 2008] Härmark, L.; van Grootheest, A. C.: Pharmacovigilance: methods, recent developments and future perspectives. In: *Eur J Clin Pharmacol* 64 (2008), Aug, No. 8, pp. 743–752. – URL `http://dx.doi.org/10.1007/s00228-008-0475-9`.

[Ingsriswang and Pacharawongsakda 2007] Ingsriswang, Supawadee; Pacharawongsakda, Eakasit: sMOL Explorer: an open source, web-enabled database and exploration tool for Small MOLecules datasets. In: *Bioinformatics* 23 (2007), Sep, No. 18, pp. 2498–2500. – URL `http://dx.doi.org/10.1093/bioinformatics/btm363`.

[Jakob et al. 2007] Jakob, R.; Ustün, B.; Madden, R.; Sykes, C.: The WHO Family of International Classifications. In: *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 50 (2007), Jul, No. 7, pp. 924–931. – URL `http://dx.doi.org/10.1007/s00103-007-0281-z`.

[Jiang and Zhai 2007] Jiang, Jing; Zhai, Cheng X.: An empirical study of tokenization strategies for biomedical information retrieval. In: *Information Retrieval* 10 (2007), pp. 341–363.

[Jiang et al. 2010] Jiang, M; Chen, Y; Liu, M: Hybrid approaches to concept extraction and assertion classification - vanderbilt's systems for 2010 I2B2 NLP Challenge. In: *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data.*, 2010.

[Jiang et al. 2011] Jiang, Min; Chen, Yukun; Liu, Mei; Rosenbloom, S T.; Mani, Subramani; Denny, Joshua C.; Xu, Hua: A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. In: *J Am Med Inform Assoc* 18 (2011), No. 5, pp. 601–606. – URL http://dx.doi.org/10.1136/amiajnl-2011-000163.

[Jimeno et al. 2008] Jimeno, Antonio; Jimenez-Ruiz, Ernesto; Lee, Vivian; Gaudan, Sylvain; Berlanga, Rafael; Rebholz-Schuhmann, Dietrich: Assessment of disease named entity recognition on a corpus of annotated sentences. In: *BMC Bioinformatics* 9 Suppl 3 (2008), pp. S3. – URL http://dx.doi.org/10.1186/1471-2105-9-S3-S3.

[Jimeno-Yepes et al. 2011] Jimeno-Yepes, Antonio J.; McInnes, Bridget T.; Aronson, Alan R.: Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. In: *BMC Bioinformatics* 12 (2011), pp. 223. – URL http://dx.doi.org/10.1186/1471-2105-12-223.

[Joachims 1998] Joachims, Thorsten: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: *Proceedings of the European Conference on Machine Learning*, 1998.

[John and Langley 1995] John, George; Langley, Pat: Estimating Continuous Distributions in Bayesian Classifiers. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995, pp. 338–345.

[Jonnalagadda and Gonzalez 2010] Jonnalagadda, S; Gonzalez, G: Can distributional statistics aid clinical concept extraction? In: *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, 2010.

[Kaelbling et al. 1996] Kaelbling, L. P.; Littman, M. L.; Moore, A. W.: Reinforcement Learning: A Survey. In: *Journal of Artificial Intelligence Research* 4 (1996), pp. 237–285.

[Kamel Boulos et al. 2010] Kamel Boulos, Maged N.; Sanfilippo, Antonio P.; Corley, Courtney D.; Wheeler, Steve: Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance. In: *Comput Methods Programs Biomed* 100 (2010), Oct, No. 1, pp. 16–23. – URL http://dx.doi.org/10.1016/j.cmpb.2010.02.007.

[Kang et al. 2010] Kang, N; Barendse, RJ; Afzal, Z: Erasmus MC approaches to the i2b2 Challenge. In: *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, 2010.

[Karimi et al. 2011] Karimi, Sarnaz; Martinez, David; Ghodke, Sumukh; Zhang, Lumin; Suominen, Hanna: Search for Medical Records: NICTA at TREC 2011 Medical Track. In: *Notebook Proceedings of Twenteith Text Retrieval Conference*, 2011.

[Karopka et al. 2006] Karopka, Thomas; Fluck, Juliane; Mevissen, Heinz-Theodor; Glass, Anne: The Autoimmune Disease Database: a dynamically compiled literature-derived database. In: *BMC Bioinformatics* 7 (2006), pp. 325. – URL http://dx.doi.org/10.1186/1471-2105-7-325.

[Kashyap 2003] Kashyap, Vipul: The UMLS Semantic Network and the Semantic Web. In: *AMIA Annu Symp Proc* (2003), pp. 351–355.

[Khoury et al. 2009] Khoury, Muin J.; Rich, Eugene C.; Randhawa, Gurvaneet; Teutsch, Steven M.; Niederhuber, John: Comparative effectiveness research and genomic medicine: an evolving partnership for 21st century medicine. In: *Genet Med* 11 (2009), Oct, No. 10, pp. 707–711. – URL http://dx.doi.org/10.1097/GIM.0b013e3181b99b90.

[Kidd and Hubbard 2007] Kidd, Michael; Hubbard, Charlotte: Introducing journal of medical case reports. In: *J Med Case Reports* 1 (2007), pp. 1. – URL http://dx.doi.org/10.1186/1752-1947-1-1.

[Kim et al. 2003] Kim, J-D.; Ohta, T.; Tateisi, Y.; Tsujii, J.: GENIA corpus–semantically annotated corpus for bio-textmining. In: *Bioinformatics* 19 Suppl 1 (2003), pp. i180–i182.

[Kind and Fiehn 2007] Kind, Tobias; Fiehn, Oliver: Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. In: *BMC Bioinformatics* 8 (2007), pp. 105. – URL http://dx.doi.org/10.1186/1471-2105-8-105.

[King et al. 2011] King, Benjamin; Wang, Lijun; Provalov, Ivan: Cengage Learning at TREC 2011 Medical Track. In: *Notebook Proceedings of Twenteith Text Retrieval Conference*, 2011.

[Klein and Manning 2003] Klein, Dan; Manning, Christopher: Accurate Unlexicalized Parsing. In: *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 2003, pp. 423–430.

[Klinger et al. 2007] Klinger, Roman; Friedrich, Christoph M.; Mevissen, Heinz T.; Fluck, Juliane; Hofmann-Apitius, Martin; Furlong, Laura I.; Sanz, Ferran: Identifying gene-specific variations in biomedical text. In: *J Bioinform Comput Biol* 5 (2007), Dec, No. 6, pp. 1277–1296.

[Klinger et al. 2008] Klinger, Roman; Kolárik, Corinna; Fluck, Juliane; Hofmann-Apitius, Martin; Friedrich, Christoph M.: Detection of IUPAC and IUPAC-like chemical names. In: *Bioinformatics* 24 (2008), Jul, No. 13, pp. i268–i276. – URL http://dx.doi.org/10.1093/bioinformatics/btn181.

[Klinger and Tomanek 2007] Klinger, Roman; Tomanek, Katrin: Classical Probabilistic Models and Conditional Random Fields. 2007. – Technical Report.

[Knox et al. 2011] Knox, Craig; Law, Vivian; Jewison, Timothy; Liu, Philip; Ly, Son; Frolkis, Alex; Pon, Allison; Banco, Kelly; Mak, Christine; Neveu, Vanessa; Djoumbou, Yannick; Eisner, Roman; Guo, An C.; Wishart, David S.: DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. In: *Nucleic Acids Res* 39 (2011), Jan, No. Database issue, pp. D1035–D1041. – URL http://dx.doi.org/10.1093/nar/gkq1126.

[Kohavi 1995] Kohavi, Ron: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1995, pp. 1137–1143.

[Kolárik et al. 2007] Kolárik, Corinna; Hofmann-Apitius, Martin; Zimmermann, Marc; Fluck, Juliane: Identification of new drug classification terms in textual resources. In: *Bioinformatics* 23 (2007), Jul, No. 13, pp. i264–i272. – URL http://dx.doi.org/10.1093/bioinformatics/btm196.

[Kotsiantis 2007] Kotsiantis, Sotiris: Supervised Machine Learning: A Review of Classification Techniques. In: *Informatica* 31 (2007), pp. 249–268.

[Kotsiantis et al. 2006] Kotsiantis, Sotiris; Kanellopoulos, Dimitris; Pintelas, Panayiotis: Handling imbalanced datasets: A review. In: *GESTS International Transactions on Computer Science and Engineering* Vol. 30, 2006.

[Krallinger et al. 2005] Krallinger, Martin; Erhardt, Ramon Alonso-Allende; Valencia, Alfonso: Text-mining approaches in molecular biology and biomedicine. In: *Drug Discov Today* 10 (2005), Mar, No. 6, pp. 439–445. – URL http://dx.doi.org/10.1016/S1359-6446(05)03376-3.

[Krauthammer and Nenadic 2004] Krauthammer, Michael; Nenadic, Goran: Term identification in the biomedical literature. In: *J Biomed Inform* 37 (2004), Dec, No. 6, pp. 512–526. – URL http://dx.doi.org/10.1016/j.jbi.2004.08.004.

[Kuhn et al. 2010] Kuhn, Michael; Campillos, Monica; Letunic, Ivica; Jensen, Lars J.; Bork, Peer: A side effect resource to capture phenotypic effects of drugs. In: *Mol Syst Biol* 6 (2010), pp. 343. – URL http://dx.doi.org/10.1038/msb.2009.98.

[Kutalik et al. 2008] Kutalik, Zoltán; Beckmann, Jacques S.; Bergmann, Sven: A modular approach for integrative analysis of large-scale gene-expression and drug-response data. In: *Nat Biotechnol* 26 (2008), May, No. 5, pp. 531–539. – URL http://dx.doi.org/10.1038/nbt1397.

[Lafferty et al. 2001] Lafferty, John; McCallum, Andrew; Pereira, Fernando: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[Larrañaga et al. 2006] Larrañaga, Pedro; Calvo, Borja; Santana, Roberto; Bielza, Concha; Galdiano, Josu; Inza, Iñaki; Lozano, José A.; Armañanzas, Rubén; Santafé, Guzmán; Pérez, Aritz; Robles, Victor: Machine learning in bioinformatics. In: *Brief Bioinform* 7 (2006), Mar, No. 1, pp. 86–112.

[Lauss et al. 2007] Lauss, Martin; Kriegner, Albert; Vierlinger, Klemens; Noehammer, Christa: Characterization of the drugged human genome. In: *Pharmacogenomics* 8 (2007), Aug, No. 8, pp. 1063–1073. – URL http://dx.doi.org/10.2217/14622416.8.8.1063.

[Leaman and Gonzalez 2008] Leaman, Robert; Gonzalez, Graciela: BANNER: an executable survey of advances in biomedical named entity recognition. In: *Pac Symp Biocomput* (2008), pp. 652–663.

[Leitner et al. 2010] Leitner, Florian; Mardis, Scott A.; Krallinger, Martin; Cesareni, Gianni; Hirschman, Lynette A.; Valencia, Alfonso: An Overview of BioCreative II.5. In: *IEEE/ACM Trans Comput Biol Bioinform* 7 (2010), No. 3, pp. 385–399.

[Li et al. 2008] Li, Dingcheng; Kipper-Schuler, Karin; Savova, Guergana: Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* , 2008.

[Li et al. 2006] Li, Yvonne Y.; An, Jianghong; Jones, Steven J M.: A large-scale computational approach to drug repositioning. In: *Genome Inform* 17 (2006), No. 2, pp. 239–247.

[Lim et al. 2010] Lim, Emilia; Pon, Allison; Djoumbou, Yannick; Knox, Craig; Shrivastava, Savita; Guo, An C.; Neveu, Vanessa; Wishart, David S.: T3DB: a comprehensively annotated database of common toxins and their targets. In: *Nucleic Acids Res* 38 (2010), Jan, No. Database issue, pp. D781–D786. – URL http://dx.doi.org/10.1093/nar/gkp934.

[Lin and Tseng 2011] Lin, Fang-Yu; Tseng, Yufeng J.: Structure-based fragment hopping for lead optimization using predocked fragment database. In: *J Chem Inf Model* 51 (2011), Jul, No. 7, pp. 1703–1715. – URL http://dx.doi.org/10.1021/ci200136j.

[Lin 2009] Lin, Jimmy: Is searching full text more effective than searching abstracts? In: *BMC Bioinformatics* 10 (2009), pp. 46. – URL http://dx.doi.org/10.1186/1471-2105-10-46.

[Lindberg et al. 1993] Lindberg, D. A.; Humphreys, B. L.; McCray, A. T.: The Unified Medical Language System. In: *Methods Inf Med* 32 (1993), Aug, No. 4, pp. 281–291.

[Liu et al. 2010] Liu, Xiaofeng; Ouyang, Sisheng; Yu, Biao; Liu, Yabo; Huang, Kai; Gong, Jiayu; Zheng, Siyuan; Li, Zhihua; Li, Honglin; Jiang, Hualiang: PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. In: *Nucleic Acids Res* 38 (2010), Jul, No. Web Server issue, pp. W609–W614. – URL http://dx.doi.org/10.1093/nar/gkq300.

[Lowe and Barnett 1994] Lowe, H. J.; Barnett, G. O.: Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. In: *JAMA* 271 (1994), Apr, No. 14, pp. 1103–1108.

[Lu et al. 2009] Lu, Zhiyong; Kim, Won; Wilbur, W J.: Evaluation of Query Expansion Using MeSH in PubMed. In: *Inf Retr Boston* 12 (2009), No. 1, pp. 69–80. – URL http://dx.doi.org/10.1007/s10791-008-9074-8.

[Luo et al. 2008] Luo, Gang; Tang, Chunqiang; Yang, Hao; Wei, Xing: MedSearch: a specialized search engine for medical information retrieval. In: *In CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge mining*, 2008, pp. 143–152.

[Lupu et al. 2011] Lupu, Mihai; Gurulingappa, Harsha; Filippov, Igor; Jiashu, Zhao; Fluck, Juliane; Zimmermann, Marc; Huang, Jimmy; Tait, John: Overview of the TREC 2011 Chemical IR Track. In: *Notebook Proceedings of the Tweentieth Text REtrieval Conference*, 2011.

[Lupu et al. 2009] Lupu, Mihai; Piroi, Florina; Huang, Xiangji; Zhu, Jianhan; Tait, John: Overview of the TREC 2009 Chemical IR Track. In: *Proceedings of the Eighteenth Text REtrieval Conference*, 2009.

[Ma'ayan et al. 2007] Ma'ayan, Avi; Jenkins, Sherry L.; Goldfarb, Joseph; Iyengar, Ravi: Network analysis of FDA approved drugs and their targets. In: *Mt Sinai J Med* 74 (2007), Apr, No. 1, pp. 27–32. – URL http://dx.doi.org/10.1002/msj.20002.

[Mack et al. 2004] Mack, R.; Mukherjea, S.; Soffer, A.; Uramoto, N.; Brown, E.; Coden, A.; Cooper, J.; Inokuchi, A.; Iyer, B.; Mass, Y.; Matsuzawa, H.; Subramaniam, L. V.: Text analytics for life science using the unstructured information management architecture. In: *IBM Syst. J.* 43 (2004), pp. 490–515.

[Mahbub Chowdhury and Lavelli 2010] Mahbub Chowdhury, Faisal; Lavelli, Alberto: Disease Mention Recognition with Specific Features. In: *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (BioNLP)*, 2010, pp. 91–98.

[Mahbub Chowdhury and Lavelli 2011] Mahbub Chowdhury, Faisal; Lavelli, Alberto: Drug-drug Interaction Extraction Using Composite Kernels. In: *Proceedings of Workshop on First Challenge Task: Drug-Drug Interaction Extraction*, 2011, pp. 27–33.

[Manning et al. 2009] Manning, Christopher; Raghavan, Prabhakar; Schutze, Hinrich: *Introduction to Information Retrieval*. Cambridge University Press, 2009.

[Maqungo et al. 2011] Maqungo, Monique; Kaur, Mandeep; Kwofie, Samuel K.; Radovanovic, Aleksandar; Schaefer, Ulf; Schmeier, Sebastian; Oppon, Ekow; Christoffels, Alan; Bajic, Vladimir B.: DDPC: Dragon Database of Genes associated with Prostate Cancer. In: *Nucleic Acids Res* 39 (2011), Jan, No. Database issue, pp. D980–D985. – URL http://dx.doi.org/10.1093/nar/gkq849.

[Marcus et al. 1993] Marcus, Mitchell; Santorini, Beatrice; Marcinkiewicz, Mary A.: Building a Large Annotated Corpus of English: The Penn Treebank. In: *COMPUTATIONAL LINGUISTICS* 19 (1993), pp. 313–330.

[Marken and Pies 2006] Marken, Patricia A.; Pies, Ronald W.: Emerging treatments for bipolar disorder: safety and adverse effect profiles. In: *Ann Pharmacother* 40 (2006), Feb, No. 2, pp. 276–285. – URL http://dx.doi.org/10.1345/aph.1G112.

[McCallum and Nigam 1998] McCallum, Andrew; Nigam, Kamal: A Comparison of Event Models for Naive Bayes Text Classification. In: *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[McCallum 2002] McCallum, Andrew K.: *MALLET: A Machine Learning for Language Toolkit*. 2002. – http://www.cs.umass.edu/ mccallum/mallet.

[Mcclosky et al. 2006] Mcclosky, David; Charniak, Eugene; Johnson, Mark: Effective self-training for parsing. In: *Proceedings of North American Chapter of the Association for Computational Linguistics*, 2006, pp. 152–159.

[Meystre et al. 2008] Meystre, S. M.; Savova, G. K.; Kipper-Schuler, K. C.; Hurdle, J. F.: Extracting information from textual documents in the electronic health record: a review of recent research. In: *Yearb Med Inform* (2008), pp. 128–144.

[Meystre et al. 2010] Meystre, Stéphane M.; Thibault, Julien; Shen, Shuying; Hurdle, John F.; South, Brett R.: Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. In: *J Am Med Inform Assoc* 17 (2010), No. 5, pp. 559–562. – URL http://dx.doi.org/10.1136/jamia.2010.004028.

[Mikut et al. 2006] Mikut, Ralf; Reischl, Markus; Burmeister, Ole; Loose, Tobias: Data mining in medical time series. In: *Biomed Tech (Berl)* 51 (2006), Dec, No. 5-6, pp. 288–293. – URL http://dx.doi.org/10.1515/BMT.2006.059.

[Miller and Britt 1995] Miller, G. C.; Britt, H.: A new drug classification for computer systems: the ATC extension code. In: *Int J Biomed Comput* 40 (1995), Oct, No. 2, pp. 121–124.

[Mingers 1989] Mingers, John: An Empirical Comparison of Pruning Methods for Decision Tree Induction. In: *Machine Learning* 4 (1989), pp. 227–243.

[Morante and Daelemans 2009] Morante, R.; Daelemans, W.: Learning the scope of hedge cues in biomedical texts. In: *Workshop on BioNLP*, 2009, pp. 28–36.

[Muench 1971] Muench, E. V.: A computerized English-Spanish correlation index to five biomedical library classification schemes based on MeSH. In: *Bull Med Libr Assoc* 59 (1971), Jul, No. 3, pp. 404–419.

[Müller et al. 2010] Müller, Bernd; Klinger, Roman; Gurulingappa, Harsha; Mevissen, Heinz-Theodor; Hofmann-Apitius, Martin; Fluck, Juliane; Friedrich, Christoph M.: Abstracts versus full texts and patents: a quantitative analysis of biomedical entities. In: *Advances in multidisciplinary retrieval - 1st Information Retrieval Facility Conference*, 2010, pp. 152–165.

[Mullins et al. 2006] Mullins, Irene M.; Siadaty, Mir S.; Lyman, Jason; Scully, Ken; Garrett, Carleton T.; Miller, W G.; Muller, Rudy; Robson, Barry; Apte, Chid; Weiss, Sholom; Rigoutsos, Isidore; Platt, Daniel; Cohen, Simona; Knaus, William A.: Data mining and clinical data repositories: Insights from a 667,000 patient data set. In: *Comput Biol Med* 36 (2006), Dec, No. 12, pp. 1351–1377. – URL http://dx.doi.org/10.1016/j.compbiomed.2005.08.003.

[Narayanaswamy et al. 2003] Narayanaswamy, Meenakshi; Ravikumar, K. E.; Vijay-Shanker, K.: A biological named entity recognizer. In: *Pac Symp Biocomput* (2003), pp. 427–438.

[Nigam et al. 1999] Nigam, Kamal; Lafferty, John; Mccallum, Andrew: Using Maximum Entropy for Text Classification. In: *In IJCAI-99 Workshop on Machine Learning for Information Filtering*, 1999, pp. 61–67.

[O'Connor et al. 2011] O'Connor, Patrick J.; Sperl-Hillen, Joann M.; Rush, William A.; Johnson, Paul E.; Amundson, Gerald H.; Asche, Stephen E.; Ekstrom, Heidi L.; Gilmer, Todd P.: Impact of electronic health record clinical decision support on diabetes care: a randomized trial. In: *Ann Fam Med* 9 (2011), No. 1, pp. 12–21. – URL http://dx.doi.org/10.1370/afm.1196.

[Pafilis et al. 2009] Pafilis, Evangelos; O'Donoghue, Seán I.; Jensen, Lars J.; Horn, Heiko; Kuhn, Michael; Brown, Nigel P.; Schneider, Reinhard: Reflect: augmented browsing for the life scientist. In: *Nat Biotechnol* 27 (2009), Jun, No. 6, pp. 508–510. – URL http://dx.doi.org/10.1038/nbt0609-508.

[Pauwels et al. 2011] Pauwels, Edouard; Stoven, Véronique; Yamanishi, Yoshihiro: Predicting drug side-effect profiles: a chemical fragment-based approach. In: *BMC Bioinformatics* 12 (2011), pp. 169. – URL http://dx.doi.org/10.1186/1471-2105-12-169.

[Penna-Coutinho et al. 2011] Penna-Coutinho, Julia; Cortopassi, Wilian A.; Oliveira, Aline A.; França, Tanos Celmar C.; Krettli, Antoniana U.: Antimalarial Activity of Potential Inhibitors of Plasmodium falciparum Lactate Dehydrogenase Enzyme Selected by Docking Studies. In: *PLoS One* 6 (2011), No. 7, pp. e21237. – URL http://dx.doi.org/10.1371/journal.pone.0021237.

[Pestian et al. 2007] Pestian, John; Brew, Chris; Matykiewicz, Pawel; Hovermale, D. J.; Johnson, Neil; Cohen, K. B.; Duch, Wlodzislaw: A shared task involving multi-label classification of clinical free text. In: *Biological, translational, and clinical language processing*, 2007, pp. 97–104.

[Plewczynski and Rychlewski 2009] Plewczynski, Dariusz; Rychlewski, Leszek: Meta-basic estimates the size of druggable human genome. In: *J Mol Model* 15 (2009), Jun, No. 6, pp. 695–699. – URL `http://dx.doi.org/10.1007/s00894-008-0353-5`.

[Plisson et al. 2004] Plisson, Joel; Lavrac, Nada; Mladenić, Dunja: A rule based approach to word lemmatization. In: *Proceedings of IS2004* Vol. 3, 2004, pp. 83–86.

[Poppenga 2001] Poppenga, R. H.: Risks associated with the use of herbs and other dietary supplements. In: *Vet Clin North Am Equine Pract* 17 (2001), Dec, No. 3, pp. 455–77, vi–vii.

[Porter 1980] Porter, Martin: *An algorithm for suffix stripping*. Program. 1980.

[Porter 2001] Porter, Martin: *Snowball: A language for stemming algorithms*. Published online. October 2001.

[Qiao et al. 2011] Qiao, Wen-Juan; Cheng, Hai-Yan; Li, Chun-Quan; Jin, Hong; Yang, Shan-Shan; Li, Xia; Zhang, Yun-Yan: Identification of pathways involved in Paclitaxel activity in cervical cancer. In: *Asian Pac J Cancer Prev* 12 (2011), No. 1, pp. 99–102.

[Quint 2000] Quint, Julien: A formalism for universal segmentation of text. In: *Proceeding COLING 'oo Proceedings of the 18th conference on Computational linguistics*, 2000.

[Rajkumar et al. 1982] Rajkumar, S.; Worku, M.; Muhammad, N. D.; Narayanaswamy, G.; Hassan, R.; Laude, T. A.; Cook, C. D.: Prescribing in pediatric ambulatory care. In: *J Ambul Care Manage* 5 (1982), Aug, No. 3, pp. 26–30.

[Rao et al. 1982] Rao, T. V.; Narayanaswamy, K. S.; Shankar, S. K.; Deshpande, D. H.: "Primary" spinal epidural lymphomas. A clinico-pathological study. In: *Acta Neurochir (Wien)* 62 (1982), No. 3-4, pp. 307–317.

[Rask-Andersen et al. 2011] Rask-Andersen, Mathias; Almén, Markus S.; Schiöth, Helgi B.: Trends in the exploitation of novel drug targets. In: *Nat Rev Drug Discov* 10 (2011), No. 8, pp. 579–590. – URL `http://dx.doi.org/10.1038/nrd3478`.

[Ratsch 2004] Ratsch, Gunnar: A Brief Introduction into Machine Learning. In: *22nd Chaos Communication Congress*, 2004.

[Rebholz-Schuhmann et al. 2007] Rebholz-Schuhmann, Dietrich; Kirsch, Harald; Arregui, Miguel; Gaudan, Sylvain; Riethoven, Mark; Stoehr, Peter: EBIMed–text crunching to gather facts for proteins from Medline. In: *Bioinformatics* 23 (2007), Jan,

No. 2, pp. e237–e244. – URL http://dx.doi.org/10.1093/bioinformatics/btl302.

[Reynolds et al. 1981] Reynolds, R. D.; Anson, N.; Narayanaswamy, T. R.; Howells, L. K.; Hafermann, D. R.; Reeves, J. D.: Chemotherapy of metastatic carcinoma of the breast. In: *Mil Med* 146 (1981), Nov, No. 11, pp. 767–770.

[Rink et al. 2011] Rink, Bryan; Harabagiu, Sanda; Roberts, Kirk: Automatic extraction of relations between medical concepts in clinical texts. In: *J Am Med Inform Assoc* 18 (2011), Sep, No. 5, pp. 594–600. – URL http://dx.doi.org/10.1136/amiajnl-2011-000153.

[Roberts et al. 2009] Roberts, Angus; Gaizauskas, Robert; Hepple, Mark; Demetriou, George; Guo, Yikun; Roberts, Ian; Setzer, Andrea: Building a semantically annotated corpus of clinical texts. In: *J Biomed Inform* 42 (2009), Oct, No. 5, pp. 950–966. – URL http://dx.doi.org/10.1016/j.jbi.2008.12.013.

[Roberts et al. 2008] Roberts, Angus; Gaizauskas, Robert; Hepple, Mark; Guo, Yikun: Mining clinical relationships from patient narratives. In: *BMC Bioinformatics* 9 Suppl 11 (2008), pp. S3. – URL http://dx.doi.org/10.1186/1471-2105-9-S11-S3.

[Roberts et al. 2010] Roberts, Kirk; Harabagiu, Sanda; Rink, Bryan: Extraction of medical concepts, assertions, and relations from discharge summaries for the fourth i2b2/VA shared task. In: *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data.*, 2010.

[Robertson et al. 1994] Robertson, Stephen; Walker, Steve; Jones, Susan; Hancock-Beaulieu, Micheline; Gatford, Mike: Okapi at TREC-3. In: *Proceedings of the Third Text REtrieval Conference (TREC 1994)*, 1994.

[Ruch et al. 2008] Ruch, Patrick; Gobeill, Julien; Lovis, Christian; Geissbühler, Antoine: Automatic medical encoding with SNOMED categories. In: *BMC Med Inform Decis Mak* 8 Suppl 1 (2008), pp. S6. – URL http://dx.doi.org/10.1186/1472-6947-8-S1-S6.

[Safavian and Landgrebe 1991] Safavian, Rasoul; Landgrebe, David: A Survey of Decision Tree Classifier Methodology. In: *IEEE transactions on systems, man, AND cybernetics* Vol. 3, 1991.

[Salton et al. 1997] Salton, Gerard; Wong, Andrew; Yang, CS: A vector space model for automatic indexing. In: *Readings in information retrieval*, Morgan Kaufmann Publishers Inc., 1997, pp. 273–280.

[Savova et al. 2008] Savova, Guergana K.; Coden, Anni R.; Sominsky, Igor L.; Johnson, Rie; Ogren, Philip V.; de Groen, Piet C.; Chute, Christopher G.: Word sense disambiguation across two domains: biomedical literature and clinical notes. In: *J Biomed Inform* 41 (2008), Dec, No. 6, pp. 1088–1100. – URL http://dx.doi.org/10.1016/j.jbi.2008.02.003.

[Savova et al. 2010] Savova, Guergana K.; Masanz, James J.; Ogren, Philip V.; Zheng, Jiaping; Sohn, Sunghwan; Kipper-Schuler, Karin C.; Chute, Christopher G.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. In: *J Am Med Inform Assoc* 17 (2010), No. 5, pp. 507–513. – URL http://dx.doi.org/10.1136/jamia.2009.001560.

[Schlich 2007] Schlich, Thomas: Contemporary history of medicine: issues and approaches. In: *Medizinhist J* 42 (2007), No. 3-4, pp. 269–298.

[Schuemie et al. 2005] Schuemie, Martijn J.; Kors, Jan A.; Mons, Barend: Word sense disambiguation in the biomedical domain: an overview. In: *J Comput Biol* 12 (2005), Jun, No. 5, pp. 554–565. – URL http://dx.doi.org/10.1089/cmb.2005.12.554.

[Schultheisz 1981] Schultheisz, R. J.: TOXLINE: evolution of an online interactive bibliographic database. In: *J Am Soc Inf Sci* 32 (1981), Nov, No. 6, pp. 421–429.

[Segota et al. 2008] Segota, Igor; Bartonicek, Nenad; Vlahovicek, Kristian: MAD-Net: microarray database network web server. In: *Nucleic Acids Res* 36 (2008), Jul, No. Web Server issue, pp. W332–W335. – URL http://dx.doi.org/10.1093/nar/gkn289.

[Segura-Bedmar et al. 2011] Segura-Bedmar, Isabel; Martinez, Paloma; Sanchez-Cisneros, Daniel: The 1st DDIExtraction-2011 Challenge Task: Extraction of Drug-Drug Interactions from Biomedical Texts. In: *Proceedings of workshop on First Challenge Task: Drug-Drug Interaction Extraction*, 2011, pp. 1–9.

[Sevenster et al. 2011] Sevenster, Merlijn; van Ommering, Rob; Qian, Yuechen: Automatically Correlating Clinical Findings and Body Locations in Radiology Reports Using MedLEE. In: *J Digit Imaging* (2011), Jul. – URL http://dx.doi.org/10.1007/s10278-011-9411-0.

[Sewell 1964] Sewell, W.: MEDICAL SUBJECT HEADINGS IN MEDLARS. In: *Bull Med Libr Assoc* 52 (1964), Jan, pp. 164–170.

[Sha and Pereira 2003] Sha, Fei; Pereira, Fernando: Shallow parsing with conditional random fields. In: *Proceeding of NAACL '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003.

[Shetty and Dalal 2011] Shetty, Kanaka D.; Dalal, Siddhartha R.: Using information mining of the medical literature to improve drug safety. In: *J Am Med Inform Assoc* 18 (2011), Sep, No. 5, pp. 668–674. – URL http://dx.doi.org/10.1136/amiajnl-2011-000096.

[Siadaty et al. 2007] Siadaty, Mir S.; Shu, Jianfen; Knaus, William A.: Relemed: sentence-level search engine with relevance score for the MEDLINE database of

biomedical articles. In: *BMC Med Inform Decis Mak* 7 (2007), pp. 1. – URL http://dx.doi.org/10.1186/1472-6947-7-1.

[Smith et al. 2004] Smith, L.; Rindflesch, T.; Wilbur, W. J.: MedPost: a part-of-speech tagger for bioMedical text. In: *Bioinformatics* 20 (2004), Sep, No. 14, pp. 2320–2321. – URL http://dx.doi.org/10.1093/bioinformatics/bth227.

[Southan et al. 2009] Southan, Christopher; Várkonyi, Péter; Muresan, Sorel: Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. In: *J Cheminform* 1 (2009), No. 1, pp. 10. – URL http://dx.doi.org/10.1186/1758-2946-1-10.

[Sperandio et al. 2009] Sperandio, O.; Petitjean, M.; Tuffery, P.: wwLigCSRre: a 3D ligand-based server for hit identification and optimization. In: *Nucleic Acids Res* 37 (2009), Jul, No. Web Server issue, pp. W504–W509. – URL http://dx.doi.org/10.1093/nar/gkp324.

[Spärck Jones 1972] Spärck Jones, Karenal: A statistical interpretation of term specificity and its application in retriev. In: *Journal of Documentation* 28 (1972), pp. 11–21.

[Stevenson et al. 2011] Stevenson, Mark; Agirre, Eneko; Soroa, Aitor: Exploiting domain information for Word Sense Disambiguation of medical documents. In: *J Am Med Inform Assoc* (2011), Sep. – URL http://dx.doi.org/10.1136/amiajnl-2011-000415.

[Strömbergsson and Kleywegt 2009] Strömbergsson, Helena; Kleywegt, Gerard J.: A chemogenomics view on protein-ligand spaces. In: *BMC Bioinformatics* 10 Suppl 6 (2009), pp. S13. – URL http://dx.doi.org/10.1186/1471-2105-10-S6-S13.

[Sweeney et al. 2005] Sweeney, James P.; Portell, Keith S.; Houck, James A.; Smith, Reginald D.; Mentel, John J.: Patient note deidentification using a find-and-replace iterative process. In: *J Healthc Inf Manag* 19 (2005), No. 3, pp. 65–70.

[Tari et al. 2010] Tari, Luis; Anwar, Saadat; Liang, Shanshan; Cai, James; Baral, Chitta: Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. In: *Bioinformatics* 26 (2010), Sep, No. 18, pp. i547–i553. – URL http://dx.doi.org/10.1093/bioinformatics/btq382.

[Thomas et al. 2011] Thomas, Philippe; Neves, Mariana; Solt, Ill'es; Tikk, Domonkos; Leser, Ulf: Relation Extraction for Drug-Drug Interactions using Ensemble Learning. In: *Proceedings of the First Challenge Task: Drug-Drug Interaction Extraction*, 2011.

[Thorn et al. 2005] Thorn, Caroline F.; Klein, Teri E.; Altman, Russ B.: PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base. In: *Methods Mol Biol* 311 (2005), pp. 179–191. – URL http://dx.doi.org/10.1385/1-59259-957-5:179.

[Tikk and Solt 2010] Tikk, Domonkos; Solt, Illés: Improving textual medication extraction using combined conditional random fields and rule-based systems. In: *J Am Med Inform Assoc* 17 (2010), No. 5, pp. 540–544. – URL http://dx.doi.org/10.1136/jamia.2010.004119.

[Tomanek et al. 2007a] Tomanek, Katrin; Wermter, Joachim; Hahn, Udo: An Approach to Text Corpus Construction which Cuts Annotation Costs and Maintains Reusability of Annotated Data. In: *Proceedings of the Joint Meeting of the Conference on Empirical Methods on Natural Language Processing and the Conference on Natural Language Learning*, 2007.

[Tomanek et al. 2007b] Tomanek, Katrin; Wermter, Joachim; Hahn, Udo: Sentence and token splitting based on conditional random fields. In: *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, 2007, pp. 19–21.

[Torii and Liu 2010] Torii, M; Liu, H: BioTagger-GM for detecting clinical concepts in electronic medical reports. . In: *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data.*, 2010.

[Trieschnigg et al. 2009] Trieschnigg, Dolf; Pezik, Piotr; Lee, Vivian; de Jong, Franciska; Kraaij, Wessel; Rebholz-Schuhmann, Dietrich: MeSH Up: effective MeSH text classification for improved document retrieval. In: *Bioinformatics* 25 (2009), Jun, No. 11, pp. 1412–1418. – URL http://dx.doi.org/10.1093/bioinformatics/btp249.

[Tsai 2006] Tsai, Richard Tzong-Han: A Hybrid Approach to Biomedical Named Entity Recognition and Semantic Role Labeling. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, 2006, pp. 243–246.

[Tsuruoka et al. 2005] Tsuruoka, Yoshimasa; Tateishi, Yuka; Kim, Jin-Dong; Ohta, Tomoko; McNaught, John; Ananiadou, Sophia; Tsujii, Junichi: Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: *Lecture Notes in Computer Science* 3746 (2005), pp. 382–392.

[Tsuruoka et al. 2008] Tsuruoka, Yoshimasa; Tsujii, Jun'ichi; Ananiadou, Sophia: FACTA: a text search engine for finding associated biomedical concepts. In: *Bioinformatics* 24 (2008), Nov, No. 21, pp. 2559–2560. – URL http://dx.doi.org/10.1093/bioinformatics/btn469.

[Uzuner 2008] Uzuner, Ozlem: Second i2b2 workshop on natural language processing challenges for clinical records. In: *AMIA Annu Symp Proc* (2008), pp. 1252–1253.

[Uzuner et al. 2010] Uzuner, Ozlem; Solti, Imre; Cadag, Eithon: Extracting medication information from clinical text. In: *J Am Med Inform Assoc* 17 (2010), No. 5, pp. 514–518. – URL http://dx.doi.org/10.1136/jamia.2010.003947.

[Uzuner et al. 2011] Uzuner, Ozlem; South, Brett R.; Shen, Shuying; Duvall, Scott L.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. In: *J Am Med Inform Assoc* 18 (2011), Sep, No. 5, pp. 552–556. – URL http://dx.doi.org/10.1136/amiajnl-2011-000203.

[Uzuner et al. 2007] Uzuner, Ozlem; Szolovits, Peter; Kohane, Isaac: *i2b2 Workshop on Natural Language Processing Challenges for Clinical Records*. 2007.

[Van De Belt et al. 2010] Van De Belt, Tom H.; Engelen, Lucien J L P G.; Berben, Sivera A A.; Schoonhoven, Lisette: Definition of Health 2.0 and Medicine 2.0: a systematic review. In: *J Med Internet Res* 12 (2010), No. 2, pp. e18. – URL http://dx.doi.org/10.2196/jmir.1350.

[van Rijsbergen 1975] van Rijsbergen, Keith: Information Retrieval. In: *Butterworths* (1975).

[Vandenbroucke 2001] Vandenbroucke, J. P.: In defense of case reports and case series. In: *Ann Intern Med* 134 (2001), Feb, No. 4, pp. 330–334.

[Vapnik 1995] Vapnik, Vladimir: The Nature of Statistical Learning Theory. In: *Springer* (1995).

[Verspoora et al. 2009] Verspoora, K.; Roeder, C.; Johnson, H.; Cohen, K.; Baumgartner, W.; Hunter, L.: Information Extraction of Normalized Protein Interaction Pairs Utilizing Linguistic and Semantic Cues. In: *Proceedings of the BioCreative II. 5 Workshop 2009 on Digital Annotations*, 2009.

[Vervloet and Durham 1998] Vervloet, D.; Durham, S.: Adverse reactions to drugs. In: *BMJ* 316 (1998), May, No. 7143, pp. 1511–1514.

[Vincent et al. 2001] Vincent, C.; Neale, G.; Woloshynowych, M.: Adverse events in British hospitals: preliminary retrospective record review. In: *BMJ* 322 (2001), Mar, No. 7285, pp. 517–519.

[Voorhees 2000] Voorhees, Ellen: Variations in relevance judgments and the measurement of retrieval effectiveness. In: *Information Processing and Management* 36 (2000), pp. 697–716.

[Voorhees and Tong 2011] Voorhees, Ellen; Tong, Richard: Overview of the TREC 2011 Medical Records Track. In: *Notebook Proceedings of the Tweentieth Text REtrieval Conference*, 2011.

[Wang and Hauskrecht 2008] Wang, Shuguang; Hauskrecht, Milos: Improving biomedical document retrieval using domain knowledge. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008.

[Warnekar et al. 2007] Warnekar, Pradnya P.; Bouhaddou, Omar; Parrish, Fola; Do, Nhan; Kilbourne, John; Brown, Steven H.; Lincoln, Michael J.: Use of RxNorm to exchange codified drug allergy information between Department of Veterans Affairs (VA) and Department of Defense (DoD). In: *AMIA Annu Symp Proc* (2007), pp. 781–785.

[Wilbur et al. 2006] Wilbur, W J.; Rzhetsky, Andrey; Shatkay, Hagit: New directions in biomedical text annotation: definitions, guidelines and corpus construction. In: *BMC Bioinformatics* 7 (2006), pp. 356. – URL http://dx.doi.org/10.1186/1471-2105-7-356.

[Wishart 2007] Wishart, David S.: In silico drug exploration and discovery using DrugBank. In: *Curr Protoc Bioinformatics* Chapter 14 (2007), Jun, pp. Unit 14.4. – URL http://dx.doi.org/10.1002/0471250953.bi1404s18.

[Wishart 2008a] Wishart, David S.: DrugBank and its relevance to pharmacogenomics. In: *Pharmacogenomics* 9 (2008), Aug, No. 8, pp. 1155–1162. – URL http://dx.doi.org/10.2217/14622416.9.8.1155.

[Wishart 2008b] Wishart, David S.: Identifying putative drug targets and potential drug leads: starting points for virtual screening and docking. In: *Methods Mol Biol* 443 (2008), pp. 333–351. – URL http://dx.doi.org/10.1007/978-1-59745-177-2_17.

[Wishart et al. 2008] Wishart, David S.; Knox, Craig; Guo, An C.; Cheng, Dean; Shrivastava, Savita; Tzur, Dan; Gautam, Bijaya; Hassanali, Murtaza: DrugBank: a knowledgebase for drugs, drug actions and drug targets. In: *Nucleic Acids Res* 36 (2008), Jan, No. Database issue, pp. D901–D906. – URL http://dx.doi.org/10.1093/nar/gkm958.

[Wishart et al. 2006] Wishart, David S.; Knox, Craig; Guo, An C.; Shrivastava, Savita; Hassanali, Murtaza; Stothard, Paul; Chang, Zhan; Woolsey, Jennifer: DrugBank: a comprehensive resource for in silico drug discovery and exploration. In: *Nucleic Acids Res* 34 (2006), Jan, No. Database issue, pp. D668–D672. – URL http://dx.doi.org/10.1093/nar/gkj067.

[Wu et al. 2011] Wu, Stephen; Wagholikar, Kavishwar; Sohn, Sunghwan; Kaggal, Vinod; Liu, Hongfang: Empirical Ontologies for Cohort Identification. In: *Notebook Proceedings of Twenteith Text Retrieval Conference*, 2011.

[Xu et al. 2010] Xu, Hua; Stenner, Shane P.; Doan, Son; Johnson, Kevin B.; Waitman, Lemuel R.; Denny, Joshua C.: MedEx: a medication information extraction system for clinical narratives. In: *J Am Med Inform Assoc* 17 (2010), No. 1, pp. 19–24. – URL http://dx.doi.org/10.1197/jamia.M3378.

[Yang and Liu 1999] Yang, Yiming; Liu, Xin: A re-examination of text categorization methods. In: *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 42–49.

[Yang et al. 2010] Yang, Zhihao; Lin, Hongfei; Li, Yanpeng: BioPPISVMExtractor: a protein-protein interaction extractor for biomedical literature using SVM and rich feature sets. In: *J Biomed Inform* 43 (2010), Feb, No. 1, pp. 88–96. – URL http://dx.doi.org/10.1016/j.jbi.2009.08.013.

[Ye et al. 2011] Ye, Hao; Ye, Li; Kang, Hong; Zhang, Duanfeng; Tao, Lin; Tang, Kailin; Liu, Xueping; Zhu, Ruixin; Liu, Qi; Chen, Y. Z.; Li, Yixue; Cao, Zhiwei: HIT: linking herbal active ingredients to targets. In: *Nucleic Acids Res* 39 (2011), Jan, No. Database issue, pp. D1055–D1059. – URL http://dx.doi.org/10.1093/nar/gkq1165.

[Yeh et al. 2005] Yeh, Alexander; Morgan, Alexander; Colosimo, Marc; Hirschman, Lynette: BioCreAtIvE task 1A: gene mention finding evaluation. In: *BMC Bioinformatics* 6 Suppl 1 (2005), pp. S2. – URL http://dx.doi.org/10.1186/1471-2105-6-S1-S2.

[Yeo and Burge 2004] Yeo, Gene; Burge, Christopher B.: Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. In: *J Comput Biol* 11 (2004), No. 2-3, pp. 377–394. – URL http://dx.doi.org/10.1089/1066527041410418.

[Zhang et al. 2011] Zhang, Zengming; Li, Yu; Lin, Biaoyang; Schroeder, Michael; Huang, Bingding: Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. In: *Bioinformatics* 27 (2011), Aug, No. 15, pp. 2083–2088. – URL http://dx.doi.org/10.1093/bioinformatics/btr331.

[Zheng et al. 2011] Zheng, Nan; Tsai, Hobart N.; Zhang, Xinyuan; Shedden, Kerby; Rosania, Gus R.: The Subcellular Distribution of Small Molecules: A Meta-Analysis. In: *Mol Pharm* (2011), Aug. – URL http://dx.doi.org/10.1021/mp200093z.

[Zhou et al. 2004] Zhou, GuoDong; Zhang, Jie; Su, Jian; Shen, Dan; Tan, ChewLim: Recognizing names in biomedical texts: a machine learning approach. In: *Bioinformatics* 20 (2004), May, No. 7, pp. 1178–1190. – URL http://dx.doi.org/10.1093/bioinformatics/bth060.

[Zhu et al. 2003] Zhu, Lingyun; Wu, Baoming; Cao, Changxiu: [Introduction to medical data mining]. In: *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi* 20 (2003), Sep, No. 3, pp. 559–562.

[Zhu et al. 2009a] Zhu, MingZhu; Gao, Lei; Li, Xia; Liu, ZhiCheng: Identifying drug-target proteins based on network features. In: *Sci China C Life Sci* 52 (2009), Apr, No. 4, pp. 398–404. – URL http://dx.doi.org/10.1007/s11427-009-0055-y.

[Zhu et al. 2009b] Zhu, Mingzhu; Gao, Lei; Li, Xia; Liu, Zhicheng; Xu, Chun; Yan, Yuqing; Walker, Erin; Jiang, Wei; Su, Bin; Chen, Xiujie; Lin, Hui: The analysis of the drug-targets based on the topological properties in the human protein-protein interaction network. In: *J Drug Target* 17 (2009), Aug, No. 7, pp. 524–532. – URL `http://dx.doi.org/10.1080/10611860903046610`.

[Zobel 1998] Zobel, Justin: How reliable are the results of large-scale information retrieval experiments? In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998.

[Zobel and Moffat 1998] Zobel, Justin; Moffat, Alistair: Exploring the similarity space. In: *SIGIR Forum* 32 (1998), pp. 18–34.