Personalized medicine through automatic extraction of information from medical texts

by

Oana Magdalena Frunză

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the Ph.D. degree in
Computer Science

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Abstract

The wealth of medical-related information available today gives rise to a multidimensional source of knowledge. Research discoveries published in prestigious venues, electronic-health records data, discharge summaries, clinical notes, etc., all represent important medical information that can assist in the medical decision-making process. The challenge that comes with accessing and using such vast and diverse sources of data stands in the ability to distil and extract reliable and relevant information. Computer-based tools that use natural language processing and machine learning techniques have proven to help address such challenges.

This current work proposes automatic reliable solutions for solving tasks that can help achieve a personalized-medicine, a medical practice that brings together general medical knowledge and case-specific medical information. Phenotypic medical observations, along with data coming from test results, are not enough when assessing and treating a medical case. Genetic, life-style, background and environmental data also need to be taken into account in the medical decision process. This thesis's goal is to prove that natural language processing and machine learning techniques represent reliable solutions for solving important medical-related problems.

From the numerous research problems that need to be answered when implementing personalized medicine, the scope of this thesis is restricted to four, as follows:

1. Automatic identification of obesity-related diseases by using only textual clinical data;

2. Automatic identification of relevant abstracts of published research to be used for building systematic reviews;

3. Automatic identification of gene functions based on textual data of published medical abstracts;

4. Automatic identification and classification of important medical relations between medical concepts in clinical and technical data.

This thesis investigation on finding automatic solutions for achieving a personalized medicine through information identification and extraction focused on individual specific problems that can be later linked in a puzzle-building manner. A diverse representation technique that follows a divide-and-conquer methodological approach shows to be the most reliable solution for building automatic models that solve the above mentioned tasks. The methodologies that I propose are supported by in-depth research experiments and thorough discussions and conclusions.

Acknowledgements

To my entire family and to all the wonderful people that offered their support and encouraged me throughout this dream come true journey.

Dedic această teză bunici mele, Magdalina Frunză care la vârsta de 46 de ani a decedat datorită unei erori medicale.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The hypothesis that drives this thesis's research is that natural language processing (NLP), along with suitable machine learning (ML) techniques represent reliable techniques for solving building-block applications required in a personalized medical practice. According to the National Cancer Institute[1] personalized medicine is "a form of medicine that uses information about a person's genes, proteins, and environment to prevent, diagnose, and treat diseases."

Personalized medicine (PM) is a medical practice that uses electronic health record data, genetic information, medical-test results, patient's lifestyle insights, background information, and new scientific discoveries as sources of knowledge for the medical-decision process. With such complex information to be taken into account, human cognition is pushed far beyond its reach as humans can use only up to seven facts per decision[2]. In this sea of information and with limited resources to navigate it, the question that arises is how to achieve the goals of medical practice: a better prevention,

[1] http://www.cancer.gov/

[2] http://www.iom.edu/Reports/2008/Evidence-Based-Medicine-and-the-Changing-Nature-of-Healthcare.aspx

a better diagnosis, and a shift from a one-size-fits-all medical practice to a personalized one. Medical records and systems that assist medical decisions have shown a positive effect for the medical care, as described by Cebul et al. (2011). But the next challenge that surfaces is: how can a better prevention be achieved and a better estimation of a therapy outcome be given for a patient, taking into account background information, biomarkers, and current diagnosis? How can "the tailoring of medical treatment to the individual characteristics of each patient [...] the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease or their response to a particular treatment", President's Council of Advisors on Science and Technology (2008) be achieved when humans are grappling by themselves with tremendous amounts of data?

The latest research in the fields of medicine, biology, computer science, engineering, and the devotion of the amazing people that stand behind these important fields of study raise hope that the personalized medicine is not a myth, but a reality. To get to the bottom of this new and exciting way of practicing medicine, the traditional way of doing it should undergo some transformations. These transformations can result in the birth of a medical practice that moves from a treatment-based to a preventive-based care. According to Abrahamas (2009), the drawbacks or challenges that this transformation process might require are out-weighted by the positive outcomes that this medical practice brings, e.g., disease detection at early stages, selection of optimal therapy, reduced adverse-drug reactions, better-designed clinical trials, etc.

The research presented in this thesis is guided by the fact that an important aspect of achieving a personalized medicine stands in the ability to surface reliable information from various sources of data.

In the next paragraphs, I will further argue why automatic tools that deal with large amounts of data are of great importance to a personalized medical practice.

In Groopman (2007), language is viewed as "the bedrock of clinical practice" on which other technologies rely on. The patient is the one that communicates the symptoms

that they have to help the healthcare provider with the diagnosis. Penz et al. (2007) showed that clinical text better identified catheter placement in surgical operation reports compared to the corresponding administrative data, e.g., ICD codes. Hripcsak et al. (1995) concluded that a natural language processor tool obtained comparable results with the physicians performance when identifying six clinical conditions in chest radiology reports.

Doctors are required to make split-second decisions and try to combine their expertise with new published information. Groopman (2007) gives an overview of the doctors' thinking process when performing a medical diagnosis. According to this work, the majority of doctors apply a strict pattern recognition technique that relies only on the information written in the person's medical file. Others try to think outside the box and listen to the patient to identify additional information to complement the release notes and discharge summaries written by previous doctors. This later type of thinking, at times, makes life-saving differences in the medical practice, suggesting that clinical data needs to be taken into account by healthcare providers. Emotions, feelings, body language, background culture information, even first impressions are key factors that dictate our relationship with the healthcare provider and, ultimately, the diagnosis.

The book of Groopman (2007) also states that, if only statistical discoveries are used, the medical practice will become one that embodies averages, not individuals. Actively using and being aware of the latest discoveries is in itself a new medical practice called evidence-based medicine. If 50 years ago, a student that finished medicine was ready to practice medicine for life, today this is not the case. A doctor that is not up-to-date with the latest discoveries is considered to be like a doctor that does not know how to use a stethoscope — a pretty bold conclusion stated by Glasziou et al. (2008). Probably this finding should come with no surprise since we live in an information-explosion era where the rate of adding scientific articles to well-known repositories, e.g., *PubMed* grows exponentially Hunter and Cohen (2006).

Since its early beginnings, artificial intelligence research directed its interest towards

the medical field. Expert systems (Ignizio, 1991) are among the first applications that merged the two fields. Since the early 70's, when expert systems like MYCIN[3] were built, a lot has changed. Living in information driven era, a big emphasis is currently put on automatic ways to extract relevant information from huge amounts of various data types (data mining). Even though things have dramatically changed since then, the complexity of the systems, their capacities, and their increased robustness, the main way of assessing and interpreting the information entered by the users is pretty much as rigid as it has ever been. The user interface dictates the data that will be processed and interpreted, giving little weight to free-form input information, such as observations made by medical professionals, patients, etc. Textual data is one of the most common ways of representing medical information, and with the recent trend of using digital medical records, this type of data becomes a valuable resource for both the medical and the computational linguistics field.

Life-science textual data has been of interest to the computational linguistics community recently. Knowledge repositories like Medline[4] helped new research fields like BioNLP (natural language processing for biology texts) emerge and develop. The majority of computational linguistics research has been focused on technical texts (what I refer to in this thesis as being articles published in journals, conferences, and other prestigious venues) and on biology specific tasks, e.g., acronym and term extraction, bio-entity tagging, protein relations, function, annotation and localization of genes, gene normalization, bio-molecular events, etc. Only recently, the focus shifted to clinical data, and this mostly due to the fact that new data sets have been made available to the research community at large.

Recent discoveries in the field of genetics reveal that using DNA information for performing diagnosis opens doors that were not accessible before. An article in the International Herald Tribune called "Searching for similar diagnosis through DNA"[5]

[3]http://www.aaai.org/AITopics/pmwiki/pmwiki.php/AITopics/RuleBasedExpertSystems

[4]http://www.nlm.nih.gov/databases/databases_medline.html

[5]http://www.iht.com/articles/2007/12/28/healthscience/28dna.php

reveals how new diagnoses procedures are now used for diseases that have their origin in mutations that appear at the chromosome level in one person's DNA. Symptoms and behaviors are now pinned down due to DNA research.

The identification of meaningful relations between medical concepts can help in the development of medical ontologies, of question-answering systems on medical problems, and for the creation of clinical trials — based on patient data. New trials for already known treatments can be created to test their therapeutic potential on other diseases (Groopman, 2007) and for identifying better treatments for a particular medical case by looking at other cases that follow a similar clinical path. Moreover, identifying relations between medical entities in clinical data can help in stratifying patients by disease susceptibility and response to therapy, reducing the size, duration, and cost of clinical trials leading to development of new treatments, diagnostics, and prevention therapies.

The fact that textual data represents one of the main sources of knowledge for attaining a personalized medical practice, the capabilities that natural language processing and machine learning offer, represent a natural solution for achieving it. Free-form clinical textual data is hardly acknowledged by automatic systems as being an important source of information. The majority of the medical systems seem to work on rigid data entry interfaces with predefined types of information. Automatic systems that are able to collect information from free clinical data, extract valuable evidence to present to healthcare providers, and infer knew knowledge from already existing data, are the tools required to achieve this new medical practice[6].

While the evidence-based medicine tries to integrate the latest medical discoveries in the clinical practice, personalized medicine is delivering care through the use of relevant information coming from patient-specific information, current medical discoveries, and most-relevant clinical knowledge. Research discoveries published in prestigious venues represent a valuable source of information for the medical practice. It is only recently

[6]http://symposium2010.amia.org/program

that, through mining techniques, free clinical textual data starts to be used as an additional and important knowledge resource. By using automatic techniques that combine information from various medical sources, extensive studies that link patient data to clinical practice and clinical outcomes can be created. The benefit of such insights lie in the ability to develop alarms, alerts, or methods to improve clinical practice and patient outcome. The key factor of having a personalized medicine stands in the ability to bring together heterogeneous reliable information for a better assessment and a better tailored therapy.

With the purpose of building automatic tools that can assist care providers, technologies that can cope with large amounts of textual data are required. Natural language processing techniques are directed to understanding text, representing it, identifying and extracting relevant information. Machine learning techniques try to solve problems in an automatic way, most of the times by learning the task from provided data. The ML field has gained its momentum in almost any domain of research and just recently has become a reliable tool in the medical domain. The empirical domain of automatic learning is used in tasks such as medical decision support, medical imaging, protein-protein interaction identification, extraction of medical knowledge, and for overall patient management care. ML is envisioned as a tool by which computer-based systems can be integrated in the healthcare field in order to obtain a more efficient medical care. Taking into account the capabilities and promises of these two fields, a marriage of the two represent the best means of achieving this thesis's goal: identify and extract technical and clinical relevant information to be used in the medical practice.

To conclude my argument, the ultimate aim of this thesis is to facilitate the extraction of healthcare information by means of using natural language processing and machine learning techniques on technical and clinical data. As stated in Swensen et al. (2010) the use and availability of such techniques are becoming crucial "[...] healthcare overvalues local autonomy and undervalues disciplined science not because of inattention or incompetence among doctors or nurses but because it is difficult for the human mind

to keep up with the explosion of medical knowledge". Due to the diverse nature of the problems that need to be solved in a personalized medical setting, this thesis, tackles a few of the puzzle pieces that need to be put together in order to achieve a personalized medicine practice, graphically described in Figure 1.1.

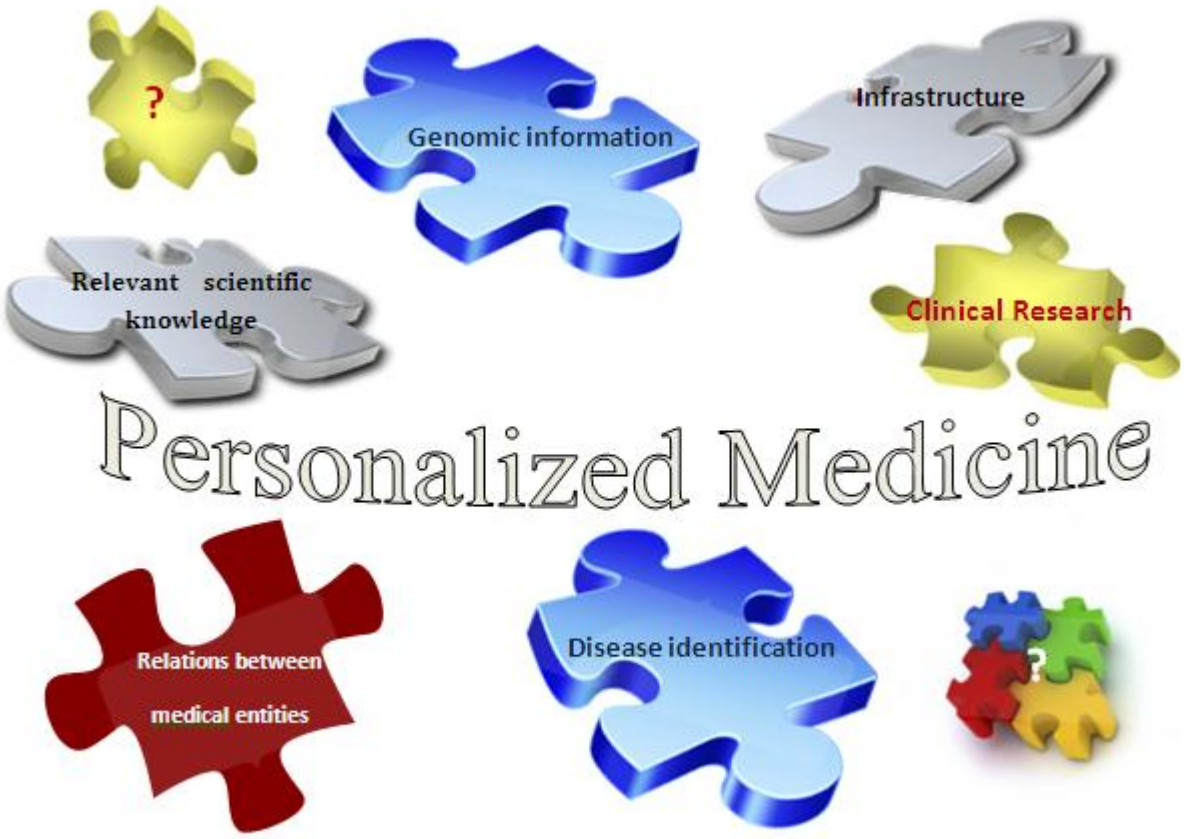

Figure 1.1: Putting the puzzle pieces together for a personalized medicine.

As the figure shows, there are a lot of interdependent research and application tasks that need to be brought together in order to achieve the vision of the personalized medicine. Genomic-markers along with phonotypical observations, test results, medical-scientific discoveries, and patient-provided observations will become the main-stream medical data to be handled by medical practitioners with the help of automatic tools

that will facilitate the management, display, and extraction of the relevant information. The goal of this thesis is to demonstrate that the earlier-mentioned techniques offer a viable solution for the following four tasks that I address:

1. Identification of obesity-related disease in clinical records, work described in Chapter 5.

2. Identification of relevant abstracts to be used for building systematic reviews, work presented in Chapter 6.

3. Identification of genes functions based on textual data, research presented in Chapter 7.

4. Identification and classification of important medical relations between medical concepts, presented in Chapter 8.

1.2 Thesis Goal

This thesis's goal is to research the problem of automatic identification and extraction of relevant medical information from technical and clinical data through natural language processing and machine learning tools.

From the wealth of tasks that need to be addressed for extracting relevant information to be used in a medical setting, the goal of my current work is directed on the tasks mentioned in Chapter 1, Section 1.1. The tasks that I tackle contribute to the larger goal of achieving a personalized medicine. When researching the best automatic models that can be used for solving these tasks I performed an in-depth analysis of data representation and of suitable learning algorithms for each particular task. I further integrated the human-specific protocol with the proposed automatic tools. I investigated the potential of both technical and clinical data, and then combined technical and clinical data to solve the task of relation identification.

1.2.1 SubGoal1: Identify obesity-related diseases in clinical data

The development of automatic systems tailored to identify diseases started with the development of the expert systems. The new generation of such systems needs to be able to mine free-form textual data for a better understanding of the patient's history, present, and future assessment.

This type of free-form textual data is a valuable source of knowledge for personalized medicine. In this setting, an automatic system needs to be able to determine and/or assist the medical practitioner in identifying the presence of diseases in a medical record. After this step is performed, further information can be identified and used, e.g., the latest research discoveries made on particular diseases (systematic reviews), genetic information (information about the genes that are linked to medical conditions), and relations that the diseases have with available treatments, tests, other diseases, and medical cases that went through a similar clinical path.

Since the identification of medical conditions is one of the first steps required in an automatically-assisted PM, a part of my thesis is dedicated to building models that can automatically predict the presence of diseases in a patient's medical record. Following the hypothesis of this thesis that NLP and ML tools are a reliable solution to medical specific tasks, the models that I develop use only the textual data of a medical record. Chapter 5 of this thesis presents in details the automatic models that I built for the identification of the presence of diseases in clinical data. The models are oriented towards obesity-related diseases and are able to identify 16 diseases in medical discharge summaries. The models are trained using two different types of annotations performed on longitudinal medical records. The first type of annotations direct the models to identify the presence of diseases based on purely textual information, while the second type of models capture the intuitive annotations of the medical experts. Both types of models use a lexical, syntactic, and domain-specific data representation in combination with suitable learning

algorithms.

1.2.2 SubGoal2: Identify relevant abstracts to be used for building systematic reviews

Systematic reviews represent summaries of the latest discoveries on different medical topics that bring the best medical evidence to medical researchers and practitioners. By building and using these types of tools, the evidence-based knowledge is brought into the personalized medical practice.

For evidence-based medical practice, they represent fundamental tools for decision making. The reviews encompass relevant and quality research published on life science topics. The process of building these reviews is a tedious process that requires human knowledge to curate thousands and thousands of documents and identify which articles are relevant and which are not. Automatic means that can help in the process of identifying articles to be included in the review would facilitate the human effort and reduce the time and cost of building the reviews.

The research goal of the work performed on this topic in Chapter 6 is to develop models that can identify abstracts of published articles as being relevant or not to the topic of a systematic review. Two novel methods are proposed for this purpose: one that uses a general machine learning setting, and another one that tries to emulate the way human judges decide on the relevance of a particular abstract. Both methods take into account the specifics of the task and particularities of the data. The models use features that alleviate the sparsity problem and learning methods that cope with large amounts of data and skewed distribution of the classes. The method that is guided by the human-specific protocol is more robust. This thesis also describes a symbiotic model that combines the human-effort with the automatic tool for achieving the best results. The methods are topic independent and can be used for building other systematic reviews.

1.2.3 SubGoal3: Predict functional properties of the genes by only using biomedical abstracts

This part of the thesis aims to automatically bring genetic information closer to the medical practice by using natural language processing and machine learning techniques. The experiments that I perform show that functional genetic knowledge can be extracted from Medline abstracts that contain mentions of genes. The aim of this work is to find means by which genetic information can be delivered in an automatic way to healthcare providers.

Knowledge repositories have been created to link genetic information with medical information, e.g., diseases[7]. The diagnosis and treatment procedure often entails knowledge coming from different sources: the disease itself; the cause of the disease — can be genetic information; environmental causes — stress, location, age group, life condition, physical activities; observations made on the patient, e.g., symptoms; and available treatments, e.g., drug treatments, chemotherapy, physiotherapy. Through the identification of the genes' function, medical practitioners can get a better understanding of the dynamic of the disease and the way it manifests for different individuals, (Scharfe et al., 2009).

Chapter 7 presents the work done on the task of identifying functional properties of genes by only using textual information. The research experiments involve the yeast gene regulation prediction. The identification of the proteins that affect the activity of Aryl Hydrocarbon Receptor (AHR) is done through models that are trained only on Medline abstracts that mention the genes involved in the biological experiments. Various textual representations and classification algorithms are used to predict the activity of the AHR system when single genes are knocked out. The best results are obtained by a lexical-based model in combination with an adaptive boosting algorithm.

[7]http://coot.embl.de/g2d/

1.2.4 SubGoal4: Identify and classify relations between medical entities in technical and clinical texts

One of the main objectives of this thesis is to identify relations between important medical concepts, e.g., diseases, treatments, and medical tests. I address this thesis's goal by means of a twofold setting: initially, I automatically identify instances that contain related concepts, while in a second step the automatic classification of the relation between the pairs of concepts is performed.

The outcome of mining such relations can help a better-targeted therapy. Spear et al. (2001) present percentages of patient populations for which drugs in a class are not effective (on average). The numbers presented, 38% for antidepressants drugs, 50% for arthritis drugs, 70% for cancer drugs, etc., suggest the need for a patient-stratification therapy based on characteristics and sensitivity to certain treatments. These results also portray the necessity of a thorough evaluation of the patients, the need of using data coming from various sources, e.g., genetic, test results, patient testimonies, cultural background, environmental and behavior information, etc., in order to identify, cure, and prevent medical conditions.

This current work addresses the task of relation identification and discrimination between medical concepts by researching the problem on both technical and clinical data, with more emphasis on the later. The hierarchical methodological settings that are used are influenced by the data type, the length of the context and the specifics of the task. Domain-knowledge is integrated in the feature set that are used for representation, and the algorithms are chosen such that they can offer reliable results. Chapter 8 presents all these results along with the ones for combining the technical and clinical data.

1.3 Contributions

The contributions of this thesis start with its aim, namely the use of natural language processing and machine learning methodologies to solve important medical tasks. The

tasks are focused on the identification and extraction of relevant information. The thesis addresses four problems by adapting the above mentioned techniques to the specifics of the tasks and to the domain.

The general contributions of this work are: the way data is represented and fed to the automatic classifier, the way medical-specific knowledge is integrated in the classification setting, the way the integration of the human workload and the automatic tools is performed, the proposed hierarchical approach of solving the relation identification and classification task, and the insights drawn from the experiments where technical and clinical data are combined to solve the relation-identification task.

The fact that the tasks that I address make use of both technical and clinical data and the fact that thorough investigations are performed on clinical-related tasks represent an important advancement of the field. Also, the genetic information, represents the capital "P" in the personalized medical setting. In this thesis, the genetic information is considered and used as a source of knowledge that can be automatically extracted from textual data.

The next paragraphs present in detail the specific contribution for each of the tasks investigated in this thesis.

1. **Disease identification**. The contributions that the current research performed on the task of identifying obesity-related diseases in clinical data stand in:

    - the implementation of reliable models that solve the task for annotations performed when using only the clinical textual data and the implementation of models that rely on annotations that use the human judges' intuition in addition to the clinical data;
    - the design of feature engineering steps performed towards integrating medical-specific information for data representation;
    - the evaluation and critical comparison of relevant learning algorithms adapted to the characteristics of the data;

- proposing a lexical-based model that uses an adaptive boosting algorithm as the best solution for the task. The obtained results, comparable to related work performed on the same data sets, suggest that the combination of lexical features and an adaptive-learning classifier used for highly skewed data sets with long texts is the best.

2. **Building systematic reviews**. Advancements to the topic of building automatic systematic reviews stand in:

   - the development of two automatic methods for helping the process of building systematic reviews through the identification of topic-relevant published articles;
   - the design and implementation of a novel solution for adapting the general machine-learning setting to the human-specific protocol of building systematic reviews;
   - the design of a voting classification technique that offers desired results;
   - showing improved results for a proposed human-machine workflow on a real-life data set;
   - use of representation and classification algorithms that cope with high-imbalanced data sets and concept-drift phenomena;
   - analysis of the correlation between the data size and the performance of the automatic models;
   - advancement of the scarce research on automatic solutions for building systematic reviews on medical-related topics.

3. **Identification of gene functions**. The main improvement that this thesis brings to the research community represents the investigation of textual information in predicting a biological-specific task, functional genomics. Detailed contributions are as follows:

- investigation of the "power" of textual information for a biology specific task;
- the evaluation of novel classification models that use only textual data to perform the task;
- the investigation of a variety of representations and learning algorithms that are used for this problem;
- the proposed domain-specific concept-based model that uses an adaptive boosting classifier;
- the novel integration of medical-specific and concept-relations information in the representation techniques;
- the fact that the proposed model obtains improved results for the task compared to other related research that uses the same data and more than just the textual information.

4. **Medical relation identification and classification**. A major part of the thesis is dedicated to the identification and classification of semantic relations between medical concepts. The fact that the majority of the research performed on this topic involves clinical data represents a significant step forward for the clinical research community. New research discoveries on clinical data are of great value, since it is only recently that clinical data has been made available to the computational linguistics community. This current work supports the computational linguistics field to gain momentum in the clinical informatics domain through:

   - investigations into adapting straight-forward techniques to achieving improved results;
   - an innovative hierarchical methodology for solving the task;
   - a thorough evaluation of the hierarchical methodology for both technical and clinical data;

- presenting meaningful results for rich representation settings that overcome the use of short contexts for classification;
- investigating the use of verb-based features for the classification task;
- critical comparison of various representation and learning algorithms that take into account the characteristics of the data;
- insights into using unlabeled data;
- pioneering experiments and conclusions when combining technical and clinical data for solving the given task.

The methodologies that I propose in this thesis are building blocks that can be used in an automatic or semi-automatic way to perform a more reliable and better-informed medical practice. Research shows that hospitals that use automated notes and records and clinical decision support systems are correlated with fewer patient complications, lower mortality rates, and ultimately lower costs (Amarasingham et al., 2009). The same vision is shared by the Canadian government that is focused on fostering and accelerating the development and adoption of electronic health record systems throughout all territories[8].

All these goals are inline with the current view of how we can achieve a personalized medicine, by building and using "IT-based clinical decision support systems that are integrated with electronic medical records and can be accessed as part of routine practice workflow. Such systems draw on the information present in the medical record to give the physician patient- and situation-specific information on the diagnostics and therapeutics relevant to the patients care[9]".

Research discoveries from this thesis have been published in the following book chapter: (Frunza and Inkpen, 2012), journal articles: (Frunza et al., 2011b), (Frunza et al., 2011a), (Matwin et al., 2010), (Matwin et al., 2011), and conferences or workshop

[8]https://www.infoway-inforoute.ca/

[9]http://www.whitehouse.gov/files/documents/ostp/PCAST/pcast_report_v2.pdf

papers: (Frunza and Inkpen, 2008a), (Frunza and Inkpen, 2008b), Kouznetsov et al. (2009), (Frunza and Inkpen, 2010b), (Frunza et al., 2010), (Frunza and Inkpen, 2010a), (Frunza and Inkpen, 2011).

1.4 Outline

The thesis is organization as follows.

Chapter 2 presents previous research in the literature that is pertinent to the tasks addressed in this thesis.

Chapter 3 describes all the data sets used for experimentation. The chapter is divided by the type of medical data used, technical and clinical.

The common textual representation techniques and learning algorithms used in the thesis are described in Chapter 4. Methodological particularities for each task are further detailed in each distinct chapter.

Chapter 5 presents the study done on identifying and classifying obesity-related diseases in patient data.

The research performed towards building automatic models that can triage Medline abstracts according to their relevance to a medical topic is presented in Chapter 6.

Chapter 7 is dedicated to research related to the biology-specific task, functional properties of the genes.

Chapter 8 is dedicated to presenting the main tasks of this thesis, the identification and classification of relations between medical concepts.

Finally, the last chapter presents the conclusions and future work.

Chapter 2

Related Work

This chapter presents relevant previous work done on the tasks that I address in this thesis. In the past few years, the computational linguistics community showed its interest on biomedical problems. A lot of research has been done on biology-specific tasks using as source of information abstracts or full-length published articles, while less work has been done on clinical data, e.g., discharge summaries, radiology reports, nurses and doctors' notes, etc.

I organized the relevant work in sections that correspond to the tasks that I have worked on, while also mentioning state-of-the-art research that has been done in the domain.

2.1 Biomedical text mining

Understanding the role of genetics in diseases, and not only, is one of the most important goals of the latest life science research. Researchers from various domains try to help advance findings on this issue; proof stands in the alliance that the computational linguistics made with the life-science domains. The emerged discipline is called Biomedical Natural Language Processing (BioNLP) and touches the fields of information science, bioinformatics, and computational linguistics.

The majority of the work that has been done in this field uses Medline abstracts as source of information. This is a positive outcome of the freely available and large repository of Medline abstracts. This huge amount of information related to the life science domain represents a great knowledge resource that can be used by the computational linguistics community (Hunter and Cohen, 2006).

Other important resources that contain more detailed life science knowledge are:

1. controlled vocabularies
    - MeSH[1];
    - UMLS[2].
2. ontologies
    - Gene Ontology[3] (GO);
    - Open Biological and Biomedical Ontologies (OBO) (Smith et al., 2007);
    - Disease ontologies [4].
3. annotated corpora
    - Genia corpus, (Kim et al., 2003);
    - OHSUMED, (Hersh et al., 1994);
    - BioCreative[5].
4. domain-specific tools
    - part-of-speech taggers: Genia tagger (Tsuruoka, 2005) and MedPost (Smith et al., 2004);

[1]http://www.ncbi.nlm.nih.gov/mesh
[2]http://www.nlm.nih.gov/research/umls/
[3]http://www.geneontology.org/
[4]http://born.nii.ac.jp/?page=ontology
[5]http://biocreative.sourceforge.net/bio_corpora_links.html

- name entity recognizers: MetaMap (Aronson and Lang, 2010), LingPipe[6], NLProt (Sven and Burkhard, 2004).

In my current work, I used UMLS as a domain-specific controlled vocabulary. UMLS is a knowledge source developed at the U.S. National Library of Medicine (NLM) that contains a metathesaurus, a semantic network, and the specialist lexicon for the biomedical domain. The Metathesaurus is a very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them. The Metathesaurus is organized by concepts (meanings) its purpose being to link alternative names and views of the same concept together and to identify useful relationships between different concepts.

The semantic network offers a categorization of all concepts represented in the UMLS Metathesaurus and provides a set of useful relationships between these concepts. The specialized lexicon is intended to be a general English lexicon that includes many biomedical terms. It includes both commonly occurring English words and biomedical vocabulary. It contains an entry for each word or term with syntactic, morphological, and orthographic information.

From the tools that were built around the UMLS resource, I use the MetaMap system. The system is capable to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in a text. MetaMap uses symbolic, natural language processing, and computational linguistic techniques.

Another domain-specific tool that I have uses is the Genia tagger. The part-of-speech tagging, shallow parsing, and named entity recognition capabilities are especially designed for biomedical texts. The tagger analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. Both the MetaMap system and the Genia tagger are presented in more details in Chapter 4, Section 4.1.

[6]http://alias-i.com/lingpipe/

Research competitions In the past years, research competitions were organized around biomedical problems offering a huge incentive to advances in research. In 2002, the Knowledge Discovery in Databases (KDD) Challenge Cup[7] was organized around the task of curation. The teams used the FlyBase data set to determine if an article should be considered relevant, based on the presence of experimental evidence of *Drosophila* gene products. Yeh et al. (2003) gives insides of the challenge. An F-score of 78% was achieved by a set of manually constructed rules that used part-of-speech (POS) tags combined with lexical and semantic constrains (Regev et al., 2002).

In the Text Retrieval Conference (TREC) 2004, a special track was dedicated to genomics (Cohen and Hersh, 2006). The task was intended to mimic the process that the human annotators in the Mouse Genome Informatics (MGI) system go through, in order to find documents that contain experimental evidence about genes. The genes were annotated using the Gene Ontology codes.

In 2007, the competition also had a genomic touch (Hersh et al., 2007). Participants were required to nominate passages of text from a collection of full-text biomedical journal articles that represent the answer to a question.

The BIOCreAtIvE[8] challenge took place in 2004 and 2006. The one in 2004 was focused on two tasks: identification of gene mentions in text and functional annotation of gene products extraction. The results for both tasks varied between 80% and 90% F-measure (Hirschman et al., 2005). The second BioCreAtIvE challenge took place in 2006 and focused on: gene mention tagging (GM), gene normalization (GN), and extraction of protein-protein interactions from text (PPI).

In 2009, the BioNLP shared task concerned the recognition of bio-molecular events that appear in biomedical literature. Kim et al. (2009a) present a general overview of what the task was and what were the results. The teams used both supervised and rule-based methods, with a combination of lexical and syntactic resources.

[7] http://www.biostat.wisc.edu/ craven/kddcup/tasks.html

[8] http://biocreative.sourceforge.net/biocreative_1.html

Named entity recognition The task of identifying biomedical entities has been a major focus of the research community because of the potential utility and complexity of the problem, yielding a tremendous amount of published research on this topic. The approaches that are being used generally fall into three categories: lexicon-based, rules-based and statistical, while combined approaches have also been applied. The entities that are discovered vary from biology-specific entities, e.g., genes and proteins to medical entities, e.g., malignancies, therapies, drug names. Cohen and Hersch (2005) published a survey on the work done on biomedical text in the past years; a big part of it is devoted to name entity recognition (NER). Another review work on biomedical literature tasks and their results is presented in Zweigenbaum et al. (2007).

Sasaki et al. (2008) describe a novel way to improve the NER performance by adding NEs to an NE dictionary without any retraining. NEs are identified in parallel with part-of-speech (POS) tags based on a general word dictionary and an NE dictionary. Then, the statistical NER system is trained on the POS and protein tagger outputs with correct NE labels attached.

An active learning approach for reducing the human effort required to create named entity annotations on a corpus was done by Tsuruoka et al. (2008). In this framework, the annotation work is performed as an iterative and interactive process between the human annotator and a probabilistic named entity tagger.

Cai and Cheng (2009) present a novel technique based on tri-training learning technique that uses three classifiers: CRFs, SVMs, and ME to identify entity annotations on the Genia[9] corpus.

Biology-specific tasks The main tasks the BioNLP community is focused on are: bio-entity recognition, protein/gene normalization, curations of information for biomedical databases, interaction and relation extraction, etc., all around automatic extraction of information about genes, proteins, and their functional relationships from text

[9]http://www-tsujii.is.s.u-tokyo.ac.jp/ genia/topics/Corpus/

documents.

The most known methodologies used to solve bio-medical related tasks are based on:

- **Co-occurrence-based methods** — identify concepts that occur in the same unit of text: a sentence, an abstract, or sometimes even a full document.
- **Knowledge-based approaches** — make use of some knowledge about the language (e.g., how it is structured) and about biologically relevant facts and the way they are mentioned in the biomedical literature.
- **Statistical/machine-learning approaches** — operate by building classifiers that are trained and then deployed to solve the task.

For the medical domain, genetics has become a mainstream source of information, a way to achieve the medicine's goal nowadays, a personalized medicine. Özgür et al. (2008) proposed an approach that can be used to extract known gene-disease associations and to infer unknown ones. They discovered that based on their methodology, a total of 95% of the top 20 ranked genes were confirmed to be related to prostate cancer (their subject of study).

Relevant work for the biology-specific task that I tackle in this thesis in functional genomics represents the work done by the teams that participated in the KDD Cup 2002[10]. The task and the data that I use are the same as in the second shared task of this challenge. The organizers of the competition provided data obtained from experiments performed on a set of yeast strains in which each strain contains a single gene that is knocked out (a gene sequence in which a single gene is inoperative). Each experiment had associated a discretized value of the activity of the aryl hydrocarbon receptor (AHR) system when a single gene was knocked out. The textual information that was available in the task was considered as an auxiliary source of information and not the primary one, as it is in my work. In contrast, with the work presented by the teams, the work that I

[10]http://www.sigkdd.org/kddcup/index.php?section=2002&method=task

have done is focused only on using the textual information for solving the task. A more detailed description of the data set and task are presented in Chapter 3, Section 3.2.

The winners of the task, Kowalczyk and Raskutti (2002), have used the textual information as additional features to the ones extracted from the other available information for the genes. They used a "bag-of-words" representation, removed stop words (function words that appear in every document, e.g., "the", "it", "of", "an", etc.), and words with low frequency. They also used a stemming algorithm as a solution to the sparsity of the data when representing the textual features. The learning algorithm was a variation of the Support Vector machine (SVM) classifier, called One Class SVM.

Krogel et al. (2002), another team that participated in the task, used the textual information with an information extraction system in order to extract missing information (function, localization, protein class) for the genes in the released data set.

Vogel and Axelrod (2002) used the Medline abstracts to extract predictive keywords, and added them to their global system. Forman (2002) used the abstracts to extract binary features for several machine learning algorithms: the Naïve Bayes, SVM and AdaBoost classifiers were used from the Weka[11] tool. These features were added to the global feature set extracted from the main sources of information.

In the work that I present, I investigate different textual representations and many learning algorithms for solving the biology task. The motivations were provided by a follow-up study of Krogel et al. (2002), who mentioned in their paper that future investigation into different representations of the textual information has a potential for an overall improvement of the performance on this task. They present a set of experiments using only the information extracted from text. In this thesis, I want to investigate and suggest what textual representations and what methods are best for this task.

The majority of the work done on Medline abstracts is framed as a text classification task, identify articles that present a certain problem, or talk about a medical issue,

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

a molecular biomarker, or is relevant to a certain topic. Identifying articles that are relevant to a certain medical topic represents the task of building systematic reviews (SR), a problem that I address in this thesis, in Chapter 6. SRs are highly structured summaries of existing research in any given field and particularly in the medical domain. In the literature, there is not much work done in automatically or semi-automatically constructing these summaries of research but relevant work has been done by Cohen et al. (2006), Aphinyanaphongs and Aliferis (2005), and Haynes et al. (1994).

The traditional way to collect and triage the abstracts in a systematic review begins with the use of simple query search techniques based on MeSH or keyword terms. The queries are usually Boolean and are optimized either for precision (the articles that are retrieved are all most likely relevant) or for recall (try to retrieve the entire collection of possible relevant articles, which might mean that articles that are not relevant can be retrieved as well), as studies (Haynes et al., 1994) show that it is difficult to obtain high performance for both measures.

Although the task of selecting articles for a systematic review is a natural application of a well-developed area of automatic text classification, prior efforts to exploit this technology for such reviews has been limited. The research done by Aphinyanaphongs and Aliferis (2005) appears to be the first such attempt. In that paper, the authors experimented with a variety of text classification techniques, using the data derived from the ACP Journal Club as their corpus. They found that SVM was the best classifier according to a variety of measures, but could not provide a comprehensive explanation as to how SVM decides whether a given abstract is relevant. The authors emphasized the difficulties related to the predominance of one class in the datasets (i.e., the number of relevant abstracts is only a small portion of the total), along with the difficulty of achieving both good recall and good precision.

Further work was done by Cohen et al. (2006), focused mostly on the elimination of non-relevant documents. As their main goal was to save work for the reviewers involved in systematic review preparation, they defined a measure called work saved over sampling

(WSS), that captured the amount of work that the reviewers would save with respect to a baseline of just sampling for a given value of recall. The idea is that a classifier can return, with high recall, a set of abstracts, and that the human needs to read only those abstracts and weed out the non-relevant ones. The savings are measured with respect to the number of abstracts that would have to be read if a random baseline classifier were used. Such a baseline corresponds to uniformly sampling a given percentage of abstracts (equal to the desired recall) from the entire set. In the work done by Cohen et al. (2006), the WSS measure was applied to report the reduction in reviewers' work when retrieving 95% of the relevant documents, but the precision was very low. The focus of my work is on developing a classifier for systematic review preparation, relying on characteristics of the data that were not included in the method used by Cohen et al. (2006), and therefore I cannot perform a direct comparison of results here. Also, the data sets that Cohen et al. (2006) used in their experiments are significantly smaller than the one that I used in my research.

2.2 Clinical text mining

Patient data, even though just recently made available to the large community has been an important and in the same time challenging source of information. The BioNLP community is achieving state-of-the art results for various tasks when using Medline articles, while the unstructured and free-from text of clinical narratives still represents a big challenge.

Before describing work that has been done using clinical texts I would like to point out some particularities of this type of texts that make the research work more challenging but more valuable in the same time.

Challenges in clinical text mining

Some of the characteristics that make clinical data different from biomedical technical data and more challenging for NLP research are:

- ungrammatical and short telegraphic phrases — the diversity of clinical texts, e.g., discharge summaries, radiology reports, clinical notes, etc., and the way they are collected creates a huge diversity in their grammatical structure. Discharge summaries and reports are often intended for communication, while other clinical texts, like progress notes, have a documentation purpose.

- abbreviation abundance — clinical data might be one of the most highly-charged abbreviation texts, with 33% of the time having more than one sense (Liu et al., 2001).

- misspellings — the fact that it is quite common for second-language learners of English to type the discharge summary of a patient creates the opportunity for grammatical errors to be present. These errors make the NLP tools trained on grammatically-correct English to choke and have trouble running.

- non-word tokens — clinical narratives contain any character that can be typed or passed (a copy-paste scenario). Laboratory and test results get passed in the clinical narratives making even the simplest NLP tasks to have difficulty, e.g., token boundary detection, sentence splitting, etc.

- a combination of structured and unstructured data — each medical institution tends to have their own "template" for creating clinical data. We can observe cases when templates and pseudo-tables (plain text looking tabular) are combined with section and subsection headings, all customized for a specific institution. Clinical textual data tends to be a combination of customized prose and non-prose.

- highly-versatile vocabulary — the medical language is different from department

to department and from a type of medical practitioner to another, requiring the NLP systems to deal with extra issues.

All the above mentioned problems create new challenges for the NLP techniques and make the research on clinical data be more demanding and valuable (Meystre et al., 2008).

For a better understanding of the clinical textual data, Appendix A presents an example of discharge summary, an example of progress notes, and an example of radiology report[12].

Information extracted from patient data represents a great value for doctors as Moore et al. (2003) show. Clinical data is of tremendous use for identifying adverse events, (Moore et al., 2003) and, as Demner-Fushman et al. (2009) state, for building computerized clinical decision support (CDS) systems. Demner-Fushman et al. (2009) present a good up-to-date overview of the natural language processing success in clinical decision support.

Castro (2007) states that electronic health records can improve the health care by bringing better patient safety. He mentions few benefits that EHR can bring:

- having available the best scientific evidence;
- giving the opportunity to patients to have multiple doctors. This way a central file is constructed and contains the entire medical history;
- having a better management of various treatment outcomes;
- reducing medical errors through being aware of allergies and medication history;
- improving the drug utilization;
- making patients being in control of their own care *e.g.,* identifying conflicting recommendations, errors in their files, etc.;

[12]Available from: http://nlp.dbmi.pitt.edu/report_repository.html

- allowing chronic patients to be alerted about new treatments, vaccinations, screening, they can monitor their health, and be more flexible for traveling.

Besides research done on clinical narratives, e.g., patient records, and discharge summaries, the radiology, emergency, and pathology reports have also been of important value to the research community.

One of the most known resources of radiology reports was put together during the 2007 Computational Medicine Center (CMC) challenge. Pestian et al. (2007) give an overview of the corpus collection and annotation process along with the description of the participating teams. The challenge also touched an important problem in clinical informatics: coding, which represents the process of assigning disease codes to patient records. Roberts et al. (2008) describe an annotation framework for clinical data aimed to identify semantic medical entities and relations between them.

The BioNLP'07 shared task involved the assignment of international standard on classification of diseases (ICD) codes to radiology reports. The overview of the task is presented by Kim et al. (2009b). The task used the data set that was made available in the CMC Challenge, one of the first challenges on clinical data. The availability of this clinical data facilitated the advances in research in this field.

The same task was addressed in the work published by Patrick et al. (2007). A classification system that uses features including negation, different strategies of measuring gloss overlaps between the content of clinical records and ICD code descriptions together with expansion of the glosses from the ICD hierarchy are used. The best classifier achieved an overall F-measure value of 88.2% on a data set of 978 free text clinical records.

Pneumonia and heart problems using radiology reports were the focus of the work done by Elkin et al. (2008) and Friedlin and McDonald (2006). Elkin et al. (2008) encoded the reports in the SNOMED CT Ontology and then a set of SNOMED CT-based rules was created for the identification of these radiological findings and diagnoses. They report a sensitivity (recall) of 100% and a specificity of 98%.

Chapman et al. (2001) found that expert rules performed better than Bayesian networks or decision trees at automatically identifying chest X-ray reports that support acute bacterial pneumonia.

Friedman et al. (2004) reported a method for encoding concepts from health records using the UMLS. In this study, the investigators used their system, MedLEE[13], to abstract concepts from the record and reported a recall of 77% and a precision of 89%

Pandemic issues have been a big concern for national security and ultimately gave rise to systems that can automatically detect this information from Internet-available data. Biocatser[14] represents an ongoing project that provides advanced search and analysis of Internet news and research literature for public health workers, clinicians and researchers interested in communicable diseases.

Negation Negation, hedging, information about the historicity of a disease and experiencer identification are also some of the tasks that attracted the researchers' attention. Chapman et al. (2007) developed a stand-alone algorithm, ConText, for identifying three contextual features: negation, e.g., *no pneumonia*; historicity — the condition is recent, occurred in the past, or might occur in the future; and experiencer — the condition occurs in the patient or in someone else, such as parents abuse alcohol.

Solt et al. (2009) present an algorithm for determining whether a medical condition is absent, present, or uncertain.

A comparison between rule-based and machine learning methods for assertion classification is presented in the study done by Uzuner et al. (2009). The authors find that the statistical approaches outperform the rule-base system when a context of 4 words around the target word is used.

A good overview of current research on information extraction task from textual documents in patient narratives is presented in Ananiadou and McNaught (2006) and in Meystre et al. (2008). The later work represents a survey of 174 publications on the topics

[13]http://zellig.cpmc.columbia.edu/medlee/

[14]http://sites.google.com/site/nhcollier/projects/biocaster

of methods used, pre-processing of textual documents, contextual features detection and analysis, extraction of information in general, extraction of codes and of information for decision-support and enrichment of the electronic health records, information extraction for surveillance, research, automated terminology management, data mining, and de-identification of clinical text.

i2b2 research competitions In the past three years a major contribution in advancing research discoveries on clinical text is due to the research challenges organized by the Informatics for Integrating Biology and the Bedside (i2b2), a National Center for Biomedical Computing[15].

The first challenge was focused on two tasks: de-identification of discharge summaries and automatic evaluation of the smoking status of patients, based on medical records.

Data consisted in discharge summaries from Partners HealthCare that were de-identified, tokenized, broken into sentences, and converted into XML format. There were five types of annotations for each summary corresponding to five possible smoking status categories:

- **a past smoker** the patient was a smoker one year or more ago, but who has not smoked for at least one year;
- **a current smoker** is a patient whose discharge summary asserts explicitly that the patient was a smoker within the past year;
- **a smoker** is a patient who is either a current or a past smoker, but whose medical record does not provide enough information to classify the patient as either;
- **a non-smoker** the discharge summary indicates that the patient never smoked.
- **an unknown** is a patient whose discharge summary does not mention anything about smoking.

[15]https://www.i2b2.org/NLP/Relations/PreviousChallenges.php

The teams that participated in the challenge used various approaches, from supervised and unsupervised classifiers to hand-crafted rules. Despite the diversity in the approaches, many of the systems produced good results. The best results were surpassing an F-measure of 84%. The majority of the systems showed that discharge summaries express the smoking status using a limited number of key textual features (e.g., smok, tobac, cigar, Social History, etc.), while the lack of them strongly identified the non-smokers' records. A complete overview of this challenge can be found in (Uzuner et al., 2008).

The second challenge took place in 2008, year in which I participated as well. The shared task logo was "Who's obese and what co-morbidities do they (definitely/likely) have?". The task to be solved was a multi-class, multi-label classification task focused on obesity and its co-morbidities. A detailed description of the challenge and the data is presented in Chapter 3, section 3.1 and in Chapter 5.

The annotation process had the data divided in two: the experts were given a textual task, which asked them to classify each disease as Present, Absent, Questionable, or Unmentioned, based on explicitly documented information in the discharge summaries, and the experts were also given an intuitive task, which asked them to classify each disease as Present, Absent, or Questionable by applying their intuition and judgment to information in the discharge summaries.

For the textual annotation, the best-ranked systems made use of lexical and domain knowledge resources: additional lexical resources, precompiled dictionaries of disease, symptom, treatment, and medication terms, to find sentence matches with the training data (Yang et al., 2009), rule-based classifiers that identified the disease-labels after divided the text in the sections, (Solt et al., 2009), regular expressions with additional keywords (Ware et al., 2009), and decision trees classifiers with features that included signs, symptoms, and medication names related to each disease (Patrick and Asgari, 2008).

For the intuitive annotations, the best systems used the knowledge from the textual

annotations, (Solt et al., 2009), (Szarvas et al., 2009), (Childs et al., 2009), while the top ones used rules that incorporated disease-specific, non-preventive medications and their brand names, disease-related procedures, and symptoms highly correlated with diseases (Solt et al., 2009), numeric expressions corresponding to measurements (Szarvas et al., 2009), and medication names (Ware et al., 2009). Machine learning approaches have also been used (Ambert and Cohen, 2009), (DeShazo and Turner, 2008). Uzuner et al. (2009) presents the full overview of the challenge.

The third i2b2 challenge was set around medication extraction. The scope of the challenge was the development of natural language processing systems for the extraction of medication-related information from narrative patient records. The information to be extracted includes medications, dosages, modes of administration, frequency of administration, and the reason for administration. Compared to previous tasks, the data set that was released in this task was unannotated, giving a chance for semi- and un-supervised systems to be built.

From the teams that participated in the competition, the vast majority used rule-based systems, while the others used statistical or hybrid approaches. The top ranked team used a hybrid approach, while the rest of the leading teams used rule-based systems. Their results scored in the mid 80s for the F-measure.

Patrick and Li (2009), the winning team used two machine learners, the conditional random field (CRF) and support vector machine (SVM), combined with a rule-based system. The complete system was a cascade of applying the classification algorithms and feature generation methods. The results that they obtained scored better than 90% accuracy on 5 out of 7 entities used in the study.

Halgrim et al. (2010) also used a hybrid system, machine learning and rule-based modules for medication information extraction. The statistical classifiers were complemented by a few template-filling rules. It achieved good performance that was comparable to the top systems in the i2b2 challenge, showing that statistical approaches can perform as well or better than systems with many sophisticated rules.

The forth i2b2 challenge was focused on three tasks: extraction of medical problems, tests, and treatments; classification of assertions made on medical problems; and relations between medical problems, tests, and treatments. I participated in the third task of this competition.

The data set released in the competition, presented in Chapter 3, Section 3.1, includes discharge summaries from Partners HealthCare and from Beth Israel Deaconess Medical Center (MIMIC II) Database[16]), as well as discharge summaries and progress notes from University of Pittsburghs Medical Center. All the records have been fully deidentified and manually annotated for concept, assertion, and relation information.

From the 16 teams that participated in this track, from the top 10, 8 systems were supervised, 1 hybrid (a combination of rule-based and supervised system), and one semi-supervised (used unlabeled data). The top system used a supervised technique, the second a semi-supervised, and the third a hybrid system. The difference in results between the first two systems was not significant.

In (Roberts et al., 2010), we can find the description of the top-ranked system. The system is using an SVM classifier for a 9-class classification task, the 8 relations in question and an additional one that shows that the pair of concepts stands in no relation. The type of features that they used are: (1) contextual features consisting in the words between the concepts, (2) information related to one of the concepts, (3) the type of the nearby concepts, (4) information extracted from Wikipedia (if there is a direct link from one page-concept to another and the shortest common ancestor of the two pages in the Wikipedia category hierarchy), (5) contextual similarity to training concept pairs. The authors concluded that domain specific tools and information provided by UMLS, MetaMap and NegEx did not help the task. They achieved an F-measure value of 73.65%.

The second team, de Bruijn et al. (2010) used a semi-supervised system that was using maximum entropy (ME) classifier. They deployed three types of classifiers based on the types of concepts present in the pair, treatment-problem, test-problem, and

[16]http://www.physionet.org/pn5/mimic2db/

problem-problem. The types of features they used are based on parsing trees information, Pointwise Mutual Information (PMI) between the concept-pair, MetaMap and cTakes outputs. They also used additional data labeled by means of bootstrapping and downsampled the data to decrease the skeweness of the data. Similar to the observations made by the first team, the authors concluded that the textual features better improved the results compared to knowledge-rich resources like MetaMap, cTakes[17] and MedLine. They obtained an F-measure value of 73.13%.

The third team, (Grouin et al., 2010) use a systems that is a combination of rule-based and SVM classification techniques. The rule-based system was used for the under-represented classes, while SVM for the other classes. The features that they used are a combination of lexical features from a 3-word window context, surface features — order of the concepts, distance between them, etc., and syntactic features — part-of-speech, prepositions, conjunctions, etc. This team reported a 70.90% F-measure.

The other teams mostly used SVM, CRF, and MaxEntropy algorithms as classifiers in combination with feature spaces that resemble the ones mentioned earlier. Divita et al. (2010) used additional human annotators to label some of the unannotated data to add as training data for few underrepresented classes, but with no significant improvement.

Solt et al. (2010) used a variety of kernels based on deep and shallow linguistic information and a suite of SVM classifiers to predict relations between concepts. As a baseline, they used a word and n-grams pattern matching approach.

2.3 Relation classification in biomedical texts

The relation classification task represents a major focus for the computational linguistic research community. The domains on which this task was deployed vary wildly, but the major approaches used to identify the semantic relation between two entities are the following:

[17] http://sourceforge.net/projects/ohnlp/files/cTAKES/

1. rule-based methods and templates to match linguistic patterns;

2. co-occurrence analysis;

3. statistical or machine-learning based approaches.

Due to the fact that the domain of the research this thesis is focused on is the bioscience domain, I will describe relevant previous work done in this domain.

The task of relation extraction or relation identification is previously tackled in the medical literature, but with a focus on biomedical tasks: sub cellular-location (Craven, 1999), gene-disorder association (Ray and Craven, 2001), and diseases and drugs (Srinivasan and Rindflesch, 2002). Usually, the data sets used in biomedical specific tasks use short texts, often sentences. This is the case of the first two related works mentioned above. The tasks often entail identification of relations between entities that co-occur in the same sentence.

2.3.1 Rule-based methods

In the biomedical literature, rule-based approaches have been widely used for solving relation extraction tasks. The main sources of information used by this technique are either syntactic: part-of-speech and syntactic structures; or semantic information in the form of fixed patterns that contain words that trigger a certain relation. One of the drawbacks of using methods based on rules is that they tend to require more human-expert effort than data-driven methods (though human effort is needed in data-driven methods too, to label the data). The best rule-based systems are the ones that use rules constructed manually or semi-automatically — extracted automatically and refined manually. Rule-based systems tend to obtain good precision results, while the recall levels tend to be low.

Syntactic rule-based relation extraction systems are complex systems based on additional tools used to assign POS tags or to extract syntactic parse trees. It is known

that in the biomedical literature such tools are not yet at the state-of-the-art level as they are for general English texts, and therefore their performance on sentences is not always the best (Bunescu et al., 2006). Representative works on syntactic rule-based approaches for relation extraction in Medline abstracts and full-text articles are presented by Thomas et al. (2000), Yakushiji et al. (2001), and Leroy et al. (2003). Even though the syntactic information is the result of tools that are not 100% accurate, success stories with these types of systems have been encountered in the biomedical domain. The winner of the BioCreative II.5 task was a syntactic-rule based system, OpenDMAP described by Hunter et al. (2008). A good comparison of different syntactic parsers and their contribution to extracting protein-protein interactions can be found in (Miyao et al., 2009).

In Ng and Wong (1999) a set of rules that define simple patterns for identifying protein-protein interactions are build. A more constrained method is presented in the Blaschke et al. (1999), the user defines both protein names for which the relations needs to be determined and if one of the words from a pre-defined set is also present, the relation is identified by using simple rules.

The semantic rule-based approaches suffer from the fact that the lexicon changes from domain to domain and new rules need to be created each time. Certain rules are created for biological corpora, medical corpora, pharmaceutical corpora, etc. Systems based on semantic rules applied to full-text articles are described by Friedman et al. (2001), on sentences by Pustejovsky et al. (2002), and on abstracts by Rindflesch et al. (2000). Some researchers combined syntactic and semantic rules extracted from Medline abstracts in order to obtain better systems with the flexibility of the syntactic information and the good precision of the semantic rules, e.g., (Gaizauskas et al., 2003) and (Novichkova et al., 2003).

Feldman et al. (2002) use a mixture of syntactic and lexical features combined with semantic constrains for finding relations between genes, proteins, drugs and diseases.

Pattern generation with the scope of identifying pairs of entities in a given relation is the focus of Agichtein and Gravano (2000)'s work. They use an initial set of pairs from

which to build contextual patterns and then to identify new pairs of entities given the learned patterns. A strong constraint is used; once a pair of entities is identified there is only one relation they are assigned to be in. They work with People an Location entities.

In (Soderland et al., 1995), a system called CRYSTAL is described. The system automatically constructs dictionaries for information extraction (IE) by learning extraction patterns from labeled data. The task is to extract *diagnosis* and *sign* or *symptom* from hospital discharge reports. Syntactic patterns are extracted from labeled data, and semantic information from the semantic network of UMLS is used to identify the semantic class of the entities.

2.3.2 Co-occurrence-based methods

The co-occurrences methods are mostly based only on lexical knowledge and words in context, and even though they tend to obtain good levels of recall, their precision is low. Good representative examples of work on Medline abstracts include (Jenssen et al., 2001) and (Stapley and Benoit, 2000). The latest work uses a graphical co-occurrence approach and states that the genes that get clustered together implicitly define the type of relation they are in.

Stephens et al. (2001) make use of co-occurrence frequency in a collection to find if there is a strong association between the pair of genes and then, using some predefined relationship words, the relation is identified.

Erhardt et al. (2006) use a combination of co-occurrence and graphical technique to identify which genes are more likely to be associated with a certain disease in the biomedical literature.

2.3.3 Statistical or machine-learning based approaches

Statistical methods tend to be used to solve various NLP tasks when annotated corpora are available. Rules are automatically extracted by the machine learning algorithm.

In general, statistical techniques can perform well even with little training data. For extracting relations, the rules are used to determine if a textual input contains a relation or not. Taking a statistical approach to solve the relation extraction problem from abstracts, the most-used representation technique is bag-of-words. It uses the words in context to create a feature vector (Donaldson et al., 2003) and (Mitsumori et al., 2006). Other researchers combined the bag-of-words features, extracted from sentences, with other sources of information like POS (Bunescu and Mooney, 2005). Giuliano et al. (2006) used two sources of information: sentences in which the relation appears and the local context of the entities, and showed that simple representation techniques bring good results.

Various learning algorithms have been used for the statistical learning approach, with kernel methods being the popular ones applied to Medline abstracts (Jiexun et al., 2008). Kernel functions are also used by Culotta and Sorensen (2004) on parse trees of the entity pair in order to identify the relation between the pair. They report an average F-measure result of 45% for classifying relations that exist between locations, a person's functional role, e.g., *affiliate, founder, etc.*, and social status, *e.g., associate, grandparent, etc.*.

The task of identifying informative sentences is addressed in the literature mostly for the tasks of summarization and information extraction, and typically on such domains as newswire data, novels, medical, and biomedical texts. In the latter-mentioned domains, Goadrich et al. (2004) used inductive logic techniques for information extraction from abstracts, while Ould et al. (2003) experimented with bag-of-word features on sentences.

Bundschus et al. (2008) use the conditional random fields technique to identify semantic relations between diseases and treatments, and relations between genes and diseases.

In her doctoral thesis, Rosario (2005) uses graphical models specifically customized for extracting semantic relations from bioscience text. She compares five generative graphical models and a neural network, using lexical, syntactic, and semantic features.

In (Sibanda, 2006) semantic relationships between medical concepts are extracted.

The thesis is also focused on de-identification, semantic category recognition and assertion classification. A combination of lexical and syntactic features along with an SVM classifier represents the proposed solution for relation classification. The relations of interest for this task are between: diseases and symptoms, diseases and treatments, symptoms and treatments, and diseases and tests. An overall micro F-measure of 84.5%, and a macro F-measure of 66.7% are obtained.

Chen et al. (2008) use a mixture of biomedical and clinical documents along with statistical techniques, in order to identify disease-drug associations. The authors are using statistical methods to identify associations between diseases and drugs in both the literature and patient records. There is no attempt to identify what relations hold between the entities.

In her doctoral thesis, Caropreso (2008) looked at semantic and syntactic ways to enrich textual representation by means of dictionaries and syntactic analysis obtained from automatic parsing, for the sentence selection task. On two biomedical and one legal-judgments data set, she shows that the enriched representations outperforms the bag-of words one when using both statistical and relational learners (Aleph[18]).

[18]http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph_toc.html

Chapter 3

Data Sets

This chapter describes the data sets used in this thesis. One of the contributions that this thesis brings is the fact that it uses technical and clinical textual data. Technical data is represented by Medline[1] abstracts. Medline is the most well-known database of published articles covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences. Clinical data is represented by discharge summaries and patient records. Appendix A presents examples of clinical data.

3.1 Clinical Data

The clinical data sets that are used in the thesis are annotated for the task of disease presence identification and the task of identifying semantic relations that exist between medical concepts.

[1] http://medline.cos.com/

3.1.1 Obesity-related discharge summaries

This data set represents the data used in the i2b2-08 Shared-Task[2], Challenges in Natural Language Processing for Clinical Data focused on obesity and its co-morbidities. The data released for the competition consists of discharge summaries of patients annotated with obesity and related diseases. The data for the challenge were randomly selected from the Research Patient Data Repository (RPDR) using a query that drew records of patients who were evaluated for either obesity or diabetes. Each of the records in the data includes zero to more than ten occurrences of the stem "obes".

Each patient record went through a process of de-identification. The procedure consisted in one automatic and two manual passes for every record. The disagreements were solved using a third manual annotator. The doctor and patient names were replaced by random names extracted from the US Census Bureau names dictionary. The phone numbers, ID numbers, and ages were replaced by randomly drawn digits for each of the numbers. The dates were replaced by surrogate ones. The new locations and hospital names were obtained by permuting syllables of the original ones. Valid cities, state names, and zip codes were added to the newly generated locations.

The records released as data for the competition were document-level annotated with obesity information and its co-morbidities. Each document is a longitudinal record of a patient; the information that it contains is collected over time. The goal of the challenge is the development of systems that accurately identify obese patients and the co-morbidities correlated with the disease.

Two types of annotation were supplied for each record: **textual annotation** — judgments made only on texts; and **intuitive annotation** — judgments made on implicit information in the narrative text (e.g., computing the BMI (Body Mass Index) is considered part of an intuitive annotation process). Two doctors annotated the data in parallel and the disagreements were resolved with the help of a third expert. The classification tasks that needs to be performed at document level requires the

[2]https://www.i2b2.org/NLP/Obesity/

identification of two types of judgments the doctors performed on the records.

The Kappa, (Cohen, 1968) value for the agreement between the annotators (before any tie-breaking) on each of the co-morbidities is:

Co-morbidity	**Textual Kappa**	**Intuitive Kappa**
Obesity	0.9068	0.8617
Diabetes mellitus (DM)	0.9072	0.8685
Hypercholesterolemia	0.8732	0.6772
Hypertriglyceridemia	0.7123	0.7154
Hypertension (HTN)	0.8150	0.6719
Atherosclerotic CV disease (CAD)	0.7790	0.8140
Heart failure (CHF)	0.8122	0.7394
Peripheral vascular disease (PVD)	0.9423	0.7328
Venous insufficiency	0.7882	0.4369
Osteoarthritis (OA)	0.7551	0.7587
Obstructive sleep apnea (OSA)	0.9193	0.9178
Asthma	0.9041	0.7638
GERD	0.8856	0.5865
Gallstones / Cholecystectomy	0.9198	0.8987
Depression	0.9170	0.8613
Gout	0.9318	0.9218

Table 3.1: The kappa values for the agreement between annotators, the obesity-related dataset.

For the textual judgments the possible classes are: “Y” that stands for “Yes, the patient has the co-morbidity”, “N” that stands for “No, the patient does not have the co-morbidity”, “Q” that stands for “Questionable whether the patient has the co-morbidity”, “U” that stands for “the co-morbidity is not mentioned in the record”. For the intuitive annotations, only the “Y”, “N”, and “Q” class labels are used, “U” is

irrelevant as an intuitive judgment.

The annotations made the challenge a multi-class, multi-label classification task focused on obesity and its co-morbidities. The 15 most representative co-morbidities of obesity were considered in this task. The challenge consisted in two tracks, one for the intuitive annotations, and one for the textual annotations. The participating teams had the choice to submit up to three runs for each of the annotations, or for both.

Table 3.2 and 3.3 present the training data sets for each type of annotation, textual and intuitive, for all 16 diseases (obesity with its 15 co-morbidities). The training data sets represent, 60% of the original data, and the remaining of 40% represents the test data.

3.1.2 Clinical data annotated with medical relations

This data set was released in the i2b2-10 Shared-Task[3]. The data set includes discharge summaries from Partners HealthCare and from Beth Israel Deaconess Medical Center (MIMIC II Database). I describe this database in detail in the following subsection, as well as the discharge summaries and the progress notes from University of Pittsburgh's Medical Center. All the records have been fully de-identified and manually annotated for concept, assertion, and relation information.

Concept annotation The concepts that were annotated in the data set are:

Medical problems, represented by 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.

Treatments, represented by 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 types therapeutic or preventive procedure, medical device, steroid, pharmacologic substance, biomedical or dental material, antibiotic, clinical drug,

[3]https://www.i2b2.org/NLP/Relations/

Disease	**Class Label**				**Total**
	Y	**N**	**Q**	**U**	
Asthma	93	3	2	630	728
CAD	399	23	7	292	721
CHF	310	11	0	399	720
Depression	104	0	0	624	728
Diabetes	485	15	7	219	726
Gallstones	109	4	1	615	729
GERD	118	1	5	599	723
Gout	90	0	4	634	728
Hypercholesterolemia	304	13	1	408	726
Hypertension	537	12	0	180	729
Hypertriglyceridemia	18	0	0	711	729
OA	115	0	0	613	728
Obesity	298	4	4	424	730
OSA	105	1	8	614	728
PVD	102	0	0	627	729
Venous Insufficiency	21	0	0	707	728

Table 3.2: Training data sets for textual annotation.

and drug delivery device.

Tests, represented by 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.

Assertion annotation For this type of annotation, each medical problem was assigned to one of six assertion categories. The assertion categories are: *present*, *absent*, *possible*, *conditional*, *hypothetical*, and *not associated* with the patient.

Disease	**Class Label**				**Total**
	Y	**N**	**Q**	**U**	
Asthma	86	596	0	NA	682
CAD	391	265	5	NA	661
CHF	308	318	1	NA	627
Depression	142	555	0	NA	697
Diabetes	473	205	5	NA	683
Gallstones	101	609	0	NA	710
GERD	144	447	1	NA	592
Gout	94	616	2	NA	712
Hypercholesterolemia	315	287	1	NA	603
Hypertension	511	127	0	NA	638
Hypertriglyceridemia	37	665	0	NA	638
OA	117	554	1	NA	672
Obesity	285	379	1	NA	665
OSA	99	606	8	NA	713
PVD	110	556	1	NA	667
Venous Insufficiency	54	577	0	NA	631

Table 3.3: Training data sets for intuitive annotation.

Relation annotation The relation annotation track of the competition required the teams to determine the type of relationship that exists between two concepts in a sentence (if any). Relations can exist only between medical problems and treatments, medical problems and tests, and medical problems and other medical problems.

There were a total of 8 possible relation types between these medical concepts:

1. treatment improves medical problem (TrIP);

2. treatment worsens medical problem (TrWP);

3. treatment causes medical problem (TrCP);

4. treatment is administered for medical problem (TrAP);

5. treatment is not administered because of medical problem (TrNAP);

6. test reveals medical problem (TeRP);

7. test conducted to investigate medical problem (TeCP);

8. medical problem indicates medical problem (PIP).

These annotations are made at sentence level. Sentences that contain these concepts, but without any relation between them, were not annotated.

The training data set consisted in 349 records, divided by their type and provenance, while the test set consisted of 477 records. Table 3.4 presents the training data distribution.

Data Set	Partners	Beth-Israel Deaconess Medical Center	University of Pittsburgh Medical Center	
Type	Discharge Summaries (DS)	Discharge Summaries (DS)	DS	Patient Progress Notes
Training	97	73	98	81
Test	133	123	102	119

Table 3.4: Summary of the training and test data set (the columns describe the providing medical institution) in the i2b2-10 shared task.

Table 3.5 presents the class distribution for the relation annotations in the training and the test data. Besides the annotated data, a number of 827 unannotated records were also released.

Relation	Training Sentences	Test Sentences	Example
PIP	1,239	1,989	*Chronic diarrhea with multiple admissions for dehydration*
TeCP	303	588	*She eventually underwent a cardiac cath to assess coronary disease*
TeRP	1,734	3,033	*Chest x-ray revealed mild pulmonary edema*
TrAP	1,423	2,487	*For her depression , continued on Citalopram 10 mg daily*
TrCP	296	444	*Left ankle was notable for swelling secondary to status post left ankle surgery*
TrIP	107	198	*Her chest pain was controlled with morphine.*
TrNAP	106	191	*She was initially started on Bactrim , but this was stopped as she was afebrile*
TrWP	56	143	*fevers recurred on this antibiotic regimen*

Table 3.5: The number of relations of each kind, from the training data set in the i2b2-10 shared task.

3.2 Technical Data

In this current work, technical data represents Medline abstracts. The abstracts' annotations are specific to each task, and the data sets are described bellow.

3.2.1 Abstracts used for building systematic reviews

Systematic reviews are highly structured summaries of existing research in a particular field. They are a valuable tool in enabling the spread of evidence-based practices especially in the medical domain as the amount of information in medical publications continues to increase at a tremendous rate. Systematic reviews help to parse this growing

body of information and distill targeted knowledge from it.

The data set that I used for this task consists of Medline abstracts, a collection of 47,274 abstracts with titles collected from MEDLINE. This data set is part of a systematic review done by the McMaster University's Evidence-Based Practice Center using TrialStat Corporation's Systematic Review System[4], a web-based software platform used to conduct systematic reviews.

The initial set of abstracts was collected using a set of Boolean search queries that were run for the specific topic of the systematic review (which was: *"the dissemination strategy of health care services for elderly people of age 65 and over"*). Normally, such queries are on purpose very general in order not to miss any relevant abstracts. The goal of the queries is to provide close to 100% recall, perhaps at the expense of the precision. It is not known if this bound of recall performance is actually achievable. The explicit terms that created the query are not disclosed; we only have the collection of abstracts and the label attached by the reviewers to each of the abstracts. The process by which labels were assigned is described below, as it follows a specific systematic review protocol. The label determines if the abstract is relevant or not to the topic of interest.

In the protocol applied, two reviewers work in parallel. They read the entire collection of 47,274 abstracts and answered a set of questions to determine if an abstract is relevant or not to the topic of review. Table 3.6 presents the set of questions used by reviewers to determine if an abstract was relevant or not. The set of questions, or screening protocol, is prepared during the setup-up stage of a systematic review when the topic of the review and the search space is defined.

The data that I used is a collection of abstracts that are labeled as relevant or not after the first screening process and after the final label is determined from the two reviewers' opinions. From the entire collection of labeled abstracts, only 7,173 are relevant, and the rest of 40,101 are non-relevant. Usually, in the process of building systematic reviews the number of non-relevant documents is much higher than the number of relevant ones.

[4]http://www.trialstat.com/

ID	Question	Type of response
1	Is this article about a dissemination strategy or a behavioral intervention?	YES/NO
2	Is this paper going to be relevant as background although it does not meet inclusion criteria?	YES/NO
3	Is the population in this article made of individuals 65-year old or older or does it comprise individuals who serve the elderly population needs (i.e. health care providers, policy makers, organizations, community)?	YES/NO
4	Is this a primary study?	YES/NO
5	Is this article about[5]:	Text
6	Other reason for exclusion. Please specify	Text
7	Is this a review?	YES/NO
8	Is there an abstract? Answering this question does not create an exclusion rule.	YES/NO

Table 3.6: The set of questions used by reviewers to determine if an abstract is relevant to the systematic review's topic

The initial retrieval query is purposefully very broad, to not miss any relevant articles. Consequently, the result of the query contains a small percentage of relevant documents, while ensuring that most of the relevant documents are retrieved. When the reviewers were not sure or when the information from the abstract was not sufficient to make a decision, the reviewers preferred to keep the document, giving it the benefit of doubt. Some of the documents that passed the first screening process were in fact discarded in the second screening process, when the full-text of the article offered more information to support the decision. The abstracts that are considered as irrelevant in the initial screening phase do not enter the next screening steps; they are removed from the systematic review.

The final goal of using this data set is to find and train a learning model that will

classify any abstract into one of the two possible classes: Included (relevant) or Excluded (non relevant). The end result of finding such a model is to reduce the human effort needed in the systematic review process and, if possible, to reduce the workload of one reviewer by using an automatic classifier.

In order to run experiments based on machine learning techniques, I randomly split the original data set, 47,274 abstracts, into a training set and a test set. I used the former part of the split for training and the latter one to evaluate the classifiers' performance in discriminating an abstract as having one of the two possible classes, Included (relevant) or Excluded (non relevant). I decided to work with a training set smaller than the test set, because ideally good results need to be obtained without using too much training data. One thing that needs to be taken into consideration is that, if we want to train a classifier for a particular topic of review, human effort is required to annotate at least part of the collection of abstracts.

From the collection of 47,274 abstracts, 20,000 were randomly taken to be part of the training data set and the remaining 27,274 represents the test set. Table 3.7 presents a summary of the data along with the class distribution in the training and test data sets. I randomly sampled the training and test data sets and the original distribution of 1:5.6 between the two classes holds in both sets.

Data set	No. of abstracts	Class distribution (included : excluded) (ratio)
Training	20,000	3,056 : 16,944 (1:5.6)
Testing	27,274	4,117 : 23,157 (1:5.6)

Table 3.7: Training and test data sets for building SRs.

3.2.2 Abstracts annotated with gene mentions

As mentioned in Chapter 1, subsection 1.2.3, one of the sub goals of this thesis is to identify and select biomedical texts for predicting functional properties of the genes.

The textual data set used for achieving this goal consists of Medline abstracts annotated with gene mentions information. The data was used in the KDD[6] Cup 2002 competition that had as goal the building of models for yeast gene regulation prediction.

The aryl hydrocarbon receptor (AHR) is a member of the protein family that mediates the biological response of the cell to various agents. Knowing which genes encode proteins that are involved in the activity of the AHR system is useful information for new drug research. While it is difficult for biologists to interpret thousands of measurements per experiment, an automatic system that can assist specialists in the process of understanding the results would be a powerful tool.

The organizers of the KDD Cup competition provided data obtained from experiments performed on a set of yeast strains in which each strain contains a single gene that is knocked out (a gene sequence in which a single gene is inoperative). Each experiment had associated a discretized value of the activity of the AHR system when a single gene was knocked out.

Each gene/experiment that was present in the task was classified into one of the three possible classes that describe the systems' response. The "**nc**" label indicates that the activity of the hidden system in this instance (i.e., yeast strain) was not significantly different than the baseline (the wild-type yeast); the "**control**" label indicates that the activity of the hidden system was significantly different than the baseline for the given instance, but that the activity of another hidden system (the control) was also significantly changed compared to its baseline; the "**change**" label describes instances in which the activity of the hidden system was significantly changed, but the activity of the control system was not significantly changed.

The sources of information that are provided by the organizers of the task are as follows:

1. hierarchical information about the function and localization of the genes;

[6]http://www.biostat.wisc.edu/ craven/kddcup/winners.html

2. relational information describing the protein-protein interactions;

3. textual information in the form of scientific abstracts from Medline that mention the genes. The abstracts date 2002 and earlier, and were collected from the repository by using a simple keyword search (the gene name).

The data that was collected for all the experiments with the yeast genome was divided into a training set and a test set. Each set contains instances/genes with the information mentioned above and a label of the AHR system's activity.

Table 3.8 describes the number of instances (abstracts) for each class of activity of the hidden system, for both the training and the test set.

	Class: nc	Class: control	Class: change
Training	2,934	46	37
Test	1,445	24	19

Table 3.8: Training and test data sets for identifying gene functions.

Using only the textual data provided in this competition, the task that I address in this thesis is to build a model that can predict which knocked out genes will have a significant effect on the AHR signaling path.

3.2.3 Abstracts annotated with disease-treatment relations

The end result of building and deploying models that can identify existing relations between medical entities can be beneficial both for healthcare providers and laypeople. Healthcare providers could have access to more precise and focused information in making a better-informed decision on which treatment a patient should be administrated; they can use the information to build clinical trials for certain treatment to identify potentially new positive effects. Laypeople can benefit from staying informed and being more in charge of their health.

The data set that I used in order to identify the best learning model able to select texts that contain disease and treatment entities and relations between them is the same as the one used by Rosario and Hearst (2004).

The authors of this paper are the ones who created and distributed the data set used in our research. The data set consists of sentences from Medline abstracts annotated with disease and treatment entities and with 8 semantic relations between diseases and treatments. The sentences were extracted from the first 100 titles and the first 40 abstracts from each of the 59 files that are part of the Medline database from 2001. Sentences that had only the two entities, *diseases* and *treatments* were considered.

The main focus of Rosario and Hearst (2004)'s work is on entity recognition for diseases and treatments. The authors use Hidden Markov Models and maximum entropy models to perform both the task of entity recognition and the relation discrimination. Their representation techniques are based on words in context, part-of-speech information, phrases, and a medical lexical ontology — Mesh terms.

Table 3.9 describes the number of annotated sentences and the relations used for annotation, as presented in Rosario and Hearst (2004). The numbers in parentheses represent the training and test set size. For example for the Cure relation, out of 810 sentences present in the data set, 648 are used for training and 162 for testing.

Relationship	Definition and Example
Cure 810 (648, 162)	TREAT cures DIS Intravenous immune globulin for recurrent spontaneous abortion
Only DIS 616 (492, 124)	TREAT not mentioned Social ties and susceptibility to the common cold
Only TREAT 166 (132, 34)	DIS not mentioned Flucticasome propionate is safe in recommended doses
Prevent 63 (50, 13)	TREAT prevents the DIS Statins for prevention of stroke
Vague 36 (28, 8)	Very unclear relationship Phenylbutazone and leukemia
Side Effect 29 (24, 5)	DIS is a result of a TREAT Malignant mesodermal mixed tumor of the uterus following irradiation
NO Cure 4 (3, 1)	TREAT does not cure DIS Evidence for double resistance to permethrin and malathion in head lice
Total relevant: 1724 (1377, 347)	
Irrelevant 1771 (1416, 355)	Treat and DIS not present Patients were followed up for 6 months
Total: 3495 (2793, 702)	

Table 3.9: Data set description. In brackets are the numbers of instances used for training and for testing, respectively

Chapter 4

Research Methodology

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) concerned with the development of computer systems that can process, understand, or communicate in natural language. Major applications of NLP include information retrieval and web search, information extraction, question answering, machine translation, sentiment analysis, text mining, and speech recognition.

Machine Learning (ML) is a subfield of AI that is focused on constructing computer systems that learn from experience, (Mitchell, 1997). It is also a field of empirical studies that gained its momentum in various domains due to its power of applicability. Off-the-shelf implementations of various algorithms are ready to be deployed for solving a large spectrum of computational problems. The real gain of using ML techniques stands in the ability to identify the right algorithms and the right data representations for particular tasks and domains. The data and the way it is used in the process of training the ML models represent one of the key factors that drive the strength of machine learning techniques.

We have to bear in mind that there are at least three challenges that can be encountered while working with ML techniques. One, is to find the most *suitable model* for prediction. The ML field offers a suite of predictive models (algorithms) that can deployed. The task of finding the most suitable one relies heavily on empirical studies

and knowledge expertise.

The second challenge represents the identification of the right features to represent the data. Identifying the right and sufficient features both for large and small data sets represents a crucial aspect of research.

The third challenge is the identification and use of the right performance evaluation measure. This step requires understanding the meaning of the results and the awareness of the data distribution.

It is still a matter of experimental discovery which algorithms and which representations techniques are suitable to use when solving a task, especially when it comes to new domains of interest. The two challenges mentioned above are addressed in this thesis by trying various types of predictive algorithms and by using various textual representation techniques that take into account the domain of the tasks.

Guided by relevant literature and individual experience, the methodologies that I am going to present in this chapter are the ones that I use to achieve my goals.

4.1 Representation techniques

One of the challenges encountered when working with ML techniques is data representation. The features chosen to represent the data strongly influence the performance of the models.

When deciding how data should be represented, known representation techniques, e.g., bag-of-words, should be used in combination with other features that take into account the nature and the domain of the task. Contributions to the research community stand in the ability to identify the way data should be represented in combination with the suitable learning algorithms.

In this thesis, several representation techniques are used, as described in the following subsection. Some are influenced by the literature, while others are innovative ways to capture the domain of the research and the task in question.

4.1.1 Bag-of-words

The bag-of-words (BOW) representation is commonly used for text classification tasks. It is a representation in which features are chosen among the words that are present in the training data. Selection techniques are used in order to identify the most suitable tokens as features.

In my work, the selected features are words delimited by spaces and simple punctuation marks such as: (,) , [,] , . , ' . Common decisions made for selecting the features are: keep only words that appeared at least three times in the training collection, contain at least one alpha-numeric character, and are longer than three characters. In some experiments, I used the frequency threshold of three, commonly used for text collections, because it removes non-informative features and also strings of characters that might be the result of a wrong tokenization when splitting the text into words. Words that have length of two or one characters are not considered as features because possible incorrect tokenization and problems with very short acronyms in the medical domain that could be highly ambiguous (could be an acronym or an abbreviation of a common word). For some experiments, a list of stop words[1] is used in order to remove common English functional terms. These function words appear in every document and therefore they do not help in the classification. The list of general English stop-words consists of approximately 700 words.

After the feature space is identified, each training and test instance is mapped to this feature representation by giving values to each feature for a certain instance.

Two most common feature value representations for BOW are: binary feature values — the value of a feature can be either 0 or 1, where 1 represents the fact that the feature is present in the instance and 0 otherwise; or frequency feature values — the value of the feature is the number of times it appears in an instance, or 0 if it did not appear.

Term-frequency inverse document frequency (tf-idf) is another feature-value representation commonly used in combination with BOW features. This representation technique

[1] http://www.site.uottawa.ca/~diana/csi5180/StopWords

is influenced by research done in Information Retrieval (IR). This weighting technique evaluates how important a word is in a document. More importance is given to terms that are frequent in a particular document, but not in the whole collection. The formula used for this representation technique is:

$tf = \frac{n_i}{\sum_k n_k}$

$idf = \log(\frac{|D|}{|t_i \epsilon d_i|})$

$tf/idf = tf * idf$ where:

tf is the term's frequency;

n_i is the number of occurrences of the word t_i;

n_k is the number of occurrences of all words in the document;

idf is the term's inverse document frequency;

|D| is the number of all documents in the system;

d_i is the document being considered;

$|t_i \epsilon d_i|$ represents the number of documents that contain the word t_i.

The feature values for the BOW representation should be chosen by taking into consideration the characteristics of the data. For example, for short texts a frequency value representation might be appropriate since the difference between a binary value representation and a frequency value representation is not large. This has the advantage that if a feature appears more than once in a short text, this means that it is important and the frequency value representation will capture this — the feature's value will be greater than that of other features.

4.1.2 Medical concepts - UMLS features

In order to work with a representation that provides features that are more general than words from the source text, I used Unified Medical Language System (UMLS) concept

representations. UMLS is a knowledge source developed at the U.S. National Library of Medicine (NLM) that contains a metathesaurus, a semantic network, and the specialist lexicon for the biomedical domain.

The metathesaurus is organized around concepts and meanings; it links alternative names and views of the same concept and identifies useful relationships between different concepts. UMLS contains over 1 million biomedical concepts, as well as over 5 million concept names which are hierarchically organized. Each unique concept that is present in the thesaurus has associated multiple text strings variants (slight morphological variations of the concept). NLM created this knowledge base by unifying hundreds of other medical knowledge bases and vocabularies (around 100 different source vocabularies like MeSH, SNOMED CT, etc.) to create an extensive resource that provides synonymy links, as well as parent-child relationships, among single or multi-word concepts. All concepts are assigned at least one semantic type from the semantic network; this provides a generalization of the existing relations between concepts. There are around 135 semantic types in the knowledge base that are linked through 54 relationships.

In addition to the UMLS knowledge base, NLM created a set of tools that allow easier access to the useful information. MetaMap, (Aronson and Lang, 2010) is a tool created by NLM that maps free text to biomedical concepts in the UMLS, or equivalently, it discovers metathesaurus concepts in text. With this software, text is processed through a series of modules that, in the end, will give a ranked list of all possible concept candidates for a particular noun-phrase. For each of the noun phrases that the system finds in the text, variant noun phrases are generated. For each of the variant noun phrases, candidate concepts (concepts that contain the noun phrase variant) from the UMLS metathesaurus are retrieved and evaluated. The retrieved concepts are compared to the actual phrase using a fit function that measures the text overlap between the actual phrase and the candidate concept (it returns a numerical value). The best of the candidates are then organized according to the decreasing value of the fit function. I used as features the semantic type for each phrase identified by the MetaMap system.

The UMLS concept representation is similar to a multi-word expression representation. For the simple BOW representation, each feature is represented by a single word. Using UMLS concepts, features are represented by a sequence of words. Another reason to use a UMLS concept representation is the **concept drift** phenomenon that can appear in a BOW representation. Especially in the medical domain texts, this is a frequent problem, as stated by Cohen et al. (2004). New articles that publish new research on a certain topic bring with them new terms that might not match the ones that were seen in the training process in a certain moment on the time line. Using a more general representation can be a step forward in solving the concept **drift phenomenon**. By using these features, I am also trying to capture in the representation specialized medical information present in the texts.

In Figure 4.1, I present an example of the output of the MetaMap system for the phrase *"to an increased risk"*. The information present in the brackets, *"Qualitative Concept, Quantitative Concept"* for the candidate with the fit function value 861 is the concept used as feature in the UMLS representation.

```
Meta Candidates (6)

-----------------------------------------------------------------
861 Risk [Qualitative Concept, Quantitative Concept]
694 Increased (Increased (qualifier value)) [Functional Concept]
623 Increase (Increase (qualifier value)) [Functional Concept]
601 Acquired (Acquired (qualifier value)) [Temporal Concept]
601 Obtained (Obtained (attribute)) [Functional Concept]
588 Increasing (Increasing (qualifier value)) [Functional Concept]
-----------------------------------------------------------------
```

Figure 4.1: Example of MetaMap system's output.

4.1.3 Semantic information about medical entities

Semantic information about particular medical entities is extracted with the ConText tool (Chapman et al., 2007). The system is capable to provide three types of contextual information for a medical condition:

1. Negation: ConText determines whether a condition is negated or not.

2. Temporality: ConText can identify if a medical condition is recent, historical or hypothetical.

3. Experiencer: ConText assigns conditions that can be experienced by someone other than the patient.

The tool uses trigger terms, pseudo-trigger terms, and terminations, in order to identify the values for these types of contextual information.

4.1.4 Syntactic features and biomedical entities

Another type of representation that I use is based on syntactic information: noun phrases, verb phrases, and biomedical concepts. In order to extract this type of information, I use the Genia[2] tagger tool. The tagger analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. The tagger is specifically tuned for biomedical text such as Medline abstracts. Figure 4.2 presents an example of the output of the Genia tagger for the sentence: *"Inhibition of NF-kappaB activation reversed the anti-apoptotic effect of isochamaejasmin"*. The noun and verb phrases identified by the tagger are features used for this representation technique.

The following preprocessing steps are applied in order to identify the final set of features to be used for classification: removing features that contain only punctuation, removing stop-words (using the same list of stop words as for our BOW representation), and considering valid features only the lemma-based forms. I chose to use lemmas because

[2]http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

```
-------------------------------------------------------------------
Inhibition          Inhibition          NN      B-NP    O
of                  of                  IN      B-PP    O
NF-kappaB           NF-kappaB           NN      B-NP    B-protein
activation          activation          NN      I-NP    O
reversed            reverse             VBD     B-VP    O
the                 the                 DT      B-NP    O
anti-apoptotic      anti-apoptotic      JJ      I-NP    O
effect              effect              NN      I-NP    O
of                  of                  IN      B-PP    O
isochamaejasmin     isochamaejasmin     NN      B-NP    O
.                   .                   .       O       O
-------------------------------------------------------------------
```

Figure 4.2: Example of Genia tagger output including for each word: its base form, its part-of-speech, B (beginning), I (inside), O (outside) tags for the word, and the final tag for the phrase.

there are a lot of inflected forms (e.g., plural forms) for the same word and the lemmatized form (the base form of a word) will give us the same base form for all of them. Another reason is to reduce the data sparseness problem. Dealing with short texts, very few features are represented in each instance; using lemma forms alleviates this problem.

In the experiments performed when using as features the result of the Genia tagger tool, only the set of identified noun phrases, verb phrases and biomedical entities are used.

Throughout the thesis all feature representations mentioned in this chapter have been tried. Experiments were also performed when these feature types were merged. When combining the features, the feature vector for each instance is a concatenation of all features.

The reason for combining various features to create the final feature space for representation was driven by the idea of creating a rich informative representation. The logic that stands behind this idea is that a rich representation combined with the right learning model can achieve good results.

4.2 Classification algorithms

In ML, most of the times the acquired expertise and knowledge from previous research guides the approach used to tackle new tasks. The models chosen to perform the task should be reliable. The research experiments need to be guided such that the desired performance is obtained. The experimental settings presented in this thesis are directed such that they are adapted to the domain of study (medical) and to the type of data that is used.

As classification algorithms, I use a set of 6 representative models: decision-based models (Decision trees), probabilistic models (Naïve Bayes (NB) and Complement Naïve Bayes (CNB)), which is adapted for text with imbalanced class distribution), adaptive boosting (AdaBoost), a linear classifier support vector machine (SVM) with polynomial kernel, and a classifier that always predicts the majority class in the training data (used as a baseline). I decided to use these classifiers because they are representative for the learning algorithms in the literature and they proved to work well on both short and long texts.

CNB modifies the standard MNB classifier by applying asymmetric word count priors, reflecting skewed class distribution (Frank and Bouckaert, 2006). Decision trees are decision-based models similar to the rule-based models that are used in hand-crafted systems, and are suitable for short texts. Probabilistic models, especially the ones based on the Naïve Bayes theory, are the state of the art in text classification.

Adaptive boosting algorithms are the ones that focus on hard-to-learn concepts, usually under-represented in the data, a characteristic that appears in short texts and

imbalanced data sets. SVM-based models are acknowledged state-of-the-art classification techniques on texts, (Joachims, 1998). All classifiers are part of a tool called Weka. One can imagine the steps of processing the data (in the current case textual information) for ML algorithms as the steps required to obtain a database table that contains as many columns as the number of features selected to represent the data, and as many rows as the number of data points from the collection.

The most difficult and important step is to identify which features should be selected to represent the data. In this thesis, I look at several feature selection techniques that try to take into account the specifics of the data, e.g., long texts, short texts, sparse features, etc.

A special column in this table represents the label of each instance. An instance represents a row that contains values for the selected features. The ML algorithms that are using this data representation to create predictive models should capture correlations between features, feature values, and labels, in order to obtain good prediction labels on future test data. The innovation and contribution that immerge form these experimental settings stand in identifying the most informative features for the task, while using a suitable predictive algorithm in order to increase the chance of predicting correct labels for new texts processed in the future.

4.3 Evaluation measures

Similar to the representation and classification techniques used with ML methodologies, the evaluation measures need to take into account the characteristics of the data and the tasks in question, as well. Existing evaluation measures should be chosen with caution in order to better show which methodology is the one that yields the best results.

The evaluation measures are computed from the confusion matrix (Kohavi and Provost, 1998) that contains information about the actual classes and the classes predicted by the classifier. Table 4.1 presents an example of a confusion matrix.

		Actual	
Predicted		Positive	Negative
	Positive	a	b
	Negative	c	d

Table 4.1: Example of confusion matrix.

where:

Let a = the number of true positives;

Let b = the number of false negatives;

Let c = the number of false positives;

Let d = the number of true negatives.

The evaluation measures commonly used in ML are: Precision, Recall, and F-measure.
Accuracy (Acc) = the total number of correctly classified instances = $\frac{(a+d)}{(a+b+c+d)}$.
Precision (P) = the ratio of correctly classified positive instances to the total number of classified as positive = $\frac{a}{(a+c)}$.
Recall (R) = the ratio of correctly classified positive instances to the total number of positives = $\frac{a}{(a+b)}$. This evaluation measure is known to the medical research community as sensitivity.
F-measure = is the weighted harmonic mean of precision and recall = $\frac{2PR}{(P+R)}$. Macro-average F-measure gives equal weight to each class, being influenced by the performance on rare classes. Micro-averaged F-measure is the harmonic mean of micro-averaged precision and recall. Micro-averages give equal weight to each sample, regardless of its class. They are dominated by those classes with the greatest number of samples (Yang and Liu, 1999).

Taking into consideration that the task is done on medical domain texts, the commonly used evaluation measures in medicine are referred as :

1. Sensitivity (SENS) = the ratio of correctly classified positive instances to the total

number of positives = Recall in ML terminology.

2. Specificity (SPEC) = the ratio of correctly classified negative instances to the total number of negatives= 1 - the rate of false positives.

When looking at the confusion matrix:

Sensitivity of positives = $\frac{a}{(a+b)}$ = recall of the positive class R_P;

Specificity of positives = $\frac{d}{(c+d)}$ = 1 - false positive rate;

Sensitivity of negatives = $\frac{d}{(c+d)}$ = recall of the negative class R_N;

Specificity of negatives = $\frac{a}{(b+a)}$ = 1 - false negative rate.

Observations: Sensitivity of positives = the specificity of negatives; Specificity of positives = the sensitivity of negatives.

Another measure that I use in this thesis is related to the Receiver Operating Characteristic (ROC) curve, more exactly the area under an ROC curve (AUC), (Mandraekar and Mandraekar, 2005). This measure is commonly used in the medical domain to summarize the overall diagnostic accuracy of a test.

An ROC curve plots SENS versus (1-SPEC) of a test where the different points on the curve correspond to different cut points used to determine if the test results are positive. AUC measures the area under this curve. An AUC value of 1 is the best case scenario, when both SENS and SPEC are 1.

As a statistical significant measure, I used the corrected paired t-test with a confidence level of 95%. This level of confidence is considered to be sufficent to show statistical difference in results.

Chapter 5

SubGoal1: Identify Obesity-related Disease in Clinical Data

5.1 Obesity-related Diseases

Disease profiling represents a challenging task in the medical domain and it can benefit from the tremendous amount of available information. A valuable source of information for understanding the way medical conditions appear, progresses, and get cured is represented by clinical narratives. Besides scientific discoveries that get published in technical journals, new information that can be extracted from medical records becomes accessible to the research community, facilitating progress both in medical and computational linguistics fields.

Timeline is a very important factor, especially in the medical domain, both from a bedside perspective and from a research point of view as well. Medical conditions evolve and change throughout the years, making the task of understanding a disease more challenging. Current medical records that were recently made available to the research community are being augmented by past paper-based records. Optical Character Recognition (OCR) tools are used in order to transform dusty medical records into electronic files that can be used as a source of additional information in clinical research.

This chapter presents research discoveries on building automatic models that solve the task of identifying obesity-related diseases in clinical records. One of the benefits of having such models stand in the ability to help in clinical trials preparations by identifying past and current records that contains mention of specific diseases.

The data set that I used for this task is described in Chapter 3, subsection 3.1.1. It consists of records annotated with 16 obesity-related diseases. The data set was released as part of the Informatics for Integrating Biology and the Bedside (i2b2) NLP 2008 shared task, in which I also participated (Frunza and Inkpen, 2008a).

To remind the reader, there were two types of annotations on the data set of records: **textual annotation** — judgments made only on texts; and **intuitive annotation** — judgments made on implicit information in the narrative text (e.g., computing the BMI (body mass index) is considered part of an intuitive annotation process). The type of annotations made the task a multi-class, multi-label classification problem focused on obesity and its co-morbidities. The goal is to evaluate systems on their ability to recognize whether a patient is obese and what co-morbidities he/she exhibits.

5.2 Experimental Setting

I considered this task to have a NLP and ML-based solution. The models that I built combined various feature-representation techniques that captured the nature of the task in combination with different learning algorithms.

5.2.1 Data representation

The first representation technique tried was the BOW approach. The chosen features were words delimitated by the following characters: space, (,) , [,] , . , ' , _. I do not consider valid features words that have a length smaller than 3. I further reduce the feature space by removing stop-words, the same as described in Chapter 4.

The second type of features used for this task represents UMLS features. The same

preprocessing steps as the ones presented in Chapter 4, section 4.1 are used, the top-ranked MetaMap extracted semantic concepts.

The last representation method is based on noun phrases that are identified by the Genia tagger. The tagger analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. The tagger is specifically tuned for biomedical text such as MEDLINE abstracts. Noun-phrases (NPs) identified by the tagger are considered as features in the representation technique. The following preprocessing steps were applied: I removed the following punctuations: [< > . , ' ( ) # $ % & + * / = [ ] - _], and I removed stop-words (the same list as for our BOW representation was used), and I considered valid features only the lemma-based forms of the identified noun-phrases.

Frequency values were chosen for all types of features.

5.2.2 Classification algorithms

As classification algorithms the Complement Naïve Bayes (CNB) and the AdaBoost algorithms from the Weka tool were used.

Both CNB and AdaBoost are known to perform well on text-classification tasks with class imbalance, (Rennie et al., 2003). AdaBoost is intended to boost the performance of a weak classifier, the Decision Stump classifier in these current experiments.

Since the AdaBoost results were consistently and t-test statistically significantly better during the 10-fold cross validation experiments for both the textual and intuitive tracks, I decided to take into consideration only the AdaBoost classifier when submitting results in the competition.

5.3 Results

In order to identify which of the classifiers and representation techniques best fit the task, 10-fold cross validation experiments were performed on the training data set, both

for the intuitive and textual annotations. As evaluation measures, I report accuracy and macro F-measure. The baseline for the task is considered the classifier that always predicts the majority class.

The evaluation measures used in the competition were: precision, recall, and F-measure both micro and macro.

Even though this task could be seen as a ranking task and AUC can be used as an evaluation measure, the competion was focused on F-measure that it easier to interpret and more meaningful for the possible medical-trained users.

5.3.1 Results for intuitive annotations

The preliminary experiments done on the training data set measured the accuracy of the classifiers, and based on these results the most promising models where used for the later-released test set. For the intuitive annotations, the results on the training set (10-fold cross validation) for the three representation techniques, BOW, UMLS, and noun phrases with the CNB and AdaBoost classifier are presented in tables 5.1, 5.2, and 5.3.

According to the paired t-test measure, the AdaBoost F-measure results for the intuitive track were significantly better than the CNB results for 14 datasets using the BOW representation and for 11 datasets using the UMLS representation.

Based on the results obtained on the training data set, the algorithm chosen to be used on the test set is AdaBoost. The models that were deployed on the test set for the intuitive annotations are using the three representation techniques mentioned earlier in this chapter. Table 5.4 presents the results for all 16 diseases.

5.3.2 Results for textual annotations

Similar to the experiments done on the intuitive annotation, models were trained and evaluated on the training data set for the textual annotations. Tables 5.5, 5.6, and 5.7 present these results.

According to the paired t-test measure, the AdaBoost F-measures results for the textual track were statistically better than the CNB results for 15 datasets using the BOW representation and for 12 datasets using the UMLS representation.

The models that were deployed on the test set for the intuitive annotations are using the three representation techniques mentioned earlier in this chapter and AdaBoost as classifier. Table 5.8 presents the results for all 16 diseases.

The best results were obtained with the BOW representation using the AdaBoost classifier. The results that we obtained with our best methods put us in the 18 rank for the intuitive track and 20 for the textual track in the list of 28 participating teams.

5.4 Discussion and Conclusions

5.4.1 Discussion

The task described in this chapter represents one of the main computational tasks addressed in the medical domain, for decision support systems. Expert systems represent one of the early implementations of computational systems aimed to address the task. While these systems tried to integrate the knowledge an expert has on a domain, the new aim is to develop systems that can bring knowledge and additional information to support their decisions. Clinical data is a source of tremendous and important information, it represents the truth of how medical practices, conditions, treatments, etc. perform in real life scenarios.

Systems that take their knowledge from doctors' and nurses' notes, medical test results, and monitoring systems' output make them be much closer to how information and knowledge is stored and passed along in the medical domain.

Medical diagnosis is not an easy task even for humans — take as example the amount of medical errors human expert make, but the vision of having computational systems aid to improve the outcome seems now more real and achievable.

The techniques that I used try to incorporate the straightforward textual repre-

sentation (BOW) with medical knowledge (UMLS concepts) and a more linguistic representation (noun phrases). Looking at the results obtained on the test set, both for the intuitive and textual annotations, the simple BOW representation outperformed the other two representations. A possible improvement over the current results might have been obtained with a combination of representations.

A possible explanation for the good performance of the BOW representation can be the fact that clinical and medical texts in general use a highly-versatile vocabulary. The same medical concept can be lexicalized in many ways. Due to this factor, the noun-phrase representation suffers from a high-sparsity problem, which in the end is reflected in the algorithms' performance.

The UMLS concepts that get identified in the text are mapped with high-level medical concepts, giving rise to a representation that finds it hard to discriminate between classes. Due to the nature of the text, the MetaMap tool tends to annotate abbreviations with medical concepts, and in this way a lot of false positive annotations will be part of the feature space.

When identifying obesity-disease mentions in clinical data, lexical features that capture the mention of the diseases performed well. Features that capture the lexical unit of a particular disease, e.g., obese, gout, etc., represent strong indicators for the classifiers that a particular disease is present. In the i2b2-07 challenge, in the identification of the smoking status of patients, similar trends were observed, the lexical features extracted from the text represented the best type of features. In a similar task, identification of the ICD codes in radiology reports, the lexical features that captured the overlap between the clinical text and the definition of the disease codes performed the best (Patrick et al., 2007).

In some cases, the simple baseline classifier outperforms the other classifiers. This is the due to the fact that the data set is highly imbalanced and as mentioned above, some representations do not capture enough information to discriminate between the classes.

5.4.2 Conclusions and future work

The best systems in the competition were the ones that incorporated additional resources, human knowledge and manual rules, showing that human expert knowledge complemented by automatic inferred information represent the best combination. The rule-based systems perform well on document-level tasks and on tasks on which few data points are given. The rules that get extracted/created seem to be more robust and efficient when larger texts are available, with more information to go through they are capable to correct some of the errors.

The current work that I presented in this chapter is just the beginning of the work required to build reliable computational systems for the goal of having automatic systems help in the diagnosis process. Never-the-less, it moves the research direction to horizons where natural language processing and machine learning techniques can make a life changing difference.

Co-morbidity	**Baseline** F-measure/Accuracy (%)	**CNB** F-measure/Accuracy (%)	**AdaBoost** F-measure/Accuracy (%)
Asthma	0.00/87.76	36.93/77.34	89.09/**97.50**
CAD	73.38/58.88	81.46/77.90	**93.27**/90.94
CHF	0.00/53.44	79.30/80.02	85.92/**87.16**
Depression	0.00/79.04	42.82/67.35	71.23/**89.35**
Diabetes	81.81/69.23	85.62/78.92	**93.55**/90.44
Gallstones	0.00/85.33	35.28/76.04	79.90/**95.25**
GERD	0.00/76.23	38.57/65.98	71.64/**89.71**
Gout	0.00/86.62	35.12/72.58	86.30/**96.86**
Hypercholesterolemia	68.49/52.09	71.90/70.88	89.66/**89.89**
Hypertension	89.25/80.60	78.29/68.92	**95.85**/93.41
Hypertriglyceridemia	0.00/94.38	11.31/92.88	50.40/**96.53**
OA	0.00/82.51	39.88/71.74	58.21/**89.88**
Obesity	0.00/56.68	64.56/67.96	96.90/**97.28**
OSA	0.00/84.76	40.16/70.38	86.93/**96.15**
PVD	0.00/84.20	49.92/82.20	71.67/**90.73**
Venous Insufficiency	0.00/91.64	36.56/89.91	44.20/**94.07**

Table 5.1: 10-fold cross validation results for the intuitive track, BOW representation.

Co-morbidity	**Baseline** F-measure/Accuracy (%)	**CNB** F-measure/Accuracy (%)	**AdaBoost** F-measure/Accuracy (%)
Asthma	0.00/87.76	34.93/88.34	86.94/**97.17**
CAD	73.38/58.88	80.60/75.14	**84.37**/82.62
CHF	0.00/53.44	77.64/77.98	80.47/**82.33**
Depression	0.00/79.04	45.05/76.43	74.93/**91.05**
Diabetes	81.81/69.23	82.73/73.64	**85.07**/76.44
Gallstones	0.00/85.33	24.15/79.46	74.79/**94.24**
GERD	0.00/76.23	36.64/72.30	69.10/**88.90**
Gout	0.00/86.62	39.68/81.49	85.34/**96.66**
Hypercholesterolemia	68.49/52.09	72.03/68.93	77.13/**80.34**
Hypertension	89.25/80.60	86.54/78.23	**94.67**/91.72
Hypertriglyceridemia	0.00/94.38	0.00/94.01	34.02/**95.71**
OA	0.00/82.51	39.40/80.31	54.46/**89.26**
Obesity	0.00/56.68	61.06/64.61	83.18/**87.55**
OSA	0.00/84.76	39.33/76.05	68.05/**92.63**
PVD	0.00/84.20	46.55/83.33	82.33/**95.33**
Venous Insufficiency	0.00/91.64	0.06/90.54	29.15/**93.03**

Table 5.2: 10-fold cross validation results for the intuitive track, UMLS representation.

Co-morbidity	**Baseline** F-measure/Accuracy (%)	**CNB** F-measure/Accuracy (%)	**AdaBoost** F-measure/Accuracy (%)
Asthma	0.00/86.72	40.46/85.04	59.50/**95.89**
CAD	73.38/55.96	45.80/61.42	51.50/**77.01**
CHF	0.00/57.31	45.36/60.97	50.56/**75.97**
Depression	0.00/85.39	53.20/79.51	51.33/**83.56**
Diabetes	81.81/66.56	41.90/64.55	26.66/**66.62**
Gallstones	0.00/84.10	31.70/76.81	45.73/**88.88**
GERD	0.00/82.95	35.16/65.55	48.23/**85.13**
Gout	0.00/87.69	44.26/80.41	58.66/**94.89**
Hypercholesterolemia	68.49/56.58	41.40/57.81	49.46/**75.93**
Hypertension	89.25/73.81	40.10/69.78	50.10/**81.18**
Hypertriglyceridemia	0.00/97.38	33.63/89.71	39.43/**93.14**
OA	0.00/84.57	36.00/75.10	45.10/**87.20**
Obesity	0.00/57.94	38.60/53.93	48.56/**77.24**
OSA	0.00/84.26	41.76/77.04	52.43/**90.87**
PVD	0.00/86.23	35.90/78.06	55.83/**92.68**
Venous Insufficiency	0.00/97.21	32.86/87.22	35.80/**89.14**

Table 5.3: 10-fold cross validation results for the intuitive track, Genia NPs representation.

Model	Precision (%)		Recall (%)		F-Measure (%)	
	Micro	Macro	Micro	Macro	Micro	Macro
BOW	92.39	94.59	92.39	60.21	92.39	**60.69**
UMLS	89.01	92.60	89.01	56.74	89.01	57.76
NPs	73.21	78.86	73.21	43.82	73.21	44.33

Table 5.4: Results on the test set for the intuitive annotations where BOW - bag-of-words features; UMLS - features; NPs - Genia tagger noun phrases.

Co-morbidity	**Baseline** F-measure/Accuracy (%)	**CNB** F-measure/Accuracy (%)	**AdaBoost** F-measure/Accuracy (%)
Asthma	0.00/86.72	37.00/75.11	92.73/**98.03**
CAD	71.76/55.96	78.50/73.86	**90.35**/87.60
CHF	0.00/57.31	73.91/75.80	86.83/**88.81**
Depression	0.00/85.39	37.15/73.21	83.50/**95.29**
Diabetes	79.91/66.56	82.71/73.79	**87.39**/83.17
Gallstones	0.00/84.10	37.25/73.26	78.56/**94.26**
GERD	0.00/82.95	34.29/66.07	80.16/**94.21**
Gout	0.00/87.69	32.84/71.99	93.47/**98.52**
Hypercholesterolemia	0.00/56.58	65.13/67.85	68.77/**79.60**
Hypertension	16.98/73.81	71.98/62.20	**95.83**/93.29
Hypertriglyceridemia	0.00/97.38	0.00/96.72	33.16/**98.11**
OA	0.00/84.57	34.61/68.31	68.73/**92.76**
Obesity	0.00/57.94	63.99/67.23	**98.14**/97.23
OSA	0.00/84.26	40.03/69.93	85.36/**95.74**
PVD	0.00/86.23	45.13/83.33	83.09/**95.69**
Venous Insufficiency	0.00/97.21	1.16/96.60	47.80/**97.97**

Table 5.5: 10-fold cross validation results for the textual annotation, BOW representation.

Co-morbidity	**Baseline** F-measure/Accuracy (%)	**CNB** F-measure/Accuracy (%)	**AdaBoost** F-measure/Accuracy (%)
Asthma	0.00/86.72	38.26/87.31	88.20/**97.21**
CAD	71.76/55.96	77.64/70.55	**76.37**/75.00
CHF	0.00/57.31	74.09/75.78	79.88/**85.00**
Depression	0.00/85.39	36.38/84.09	86.17/**96.04**
Diabetes	0.00/66.56	80.55/69.93	**79.99**/66.77
Gallstones	79.91/84.10	28.31/77.23	75.65/**93.77**
GERD	0.00/82.95	36.07/75.65	77.71/**93.55**
Gout	0.00/87.69	37.61/82.99	92.72/**98.36**
Hypercholesterolemia	0.00/56.58	64.88/66.23	62.46/**76.28**
Hypertension	16.98/73.81	78.68/69.13	**94.83**/91.98
Hypertriglyceridemia	0.00/97.38	0.00/97.21	29.33/**97.98**
OA	0.00/84.57	33.90/77.85	63.19/**91.92**
Obesity	0.00/57.94	60.21/64.30	84.53/**88.05**
OSA	0.00/84.26	39.64/75.44	67.97/**92.30**
PVD	0.00/86.23	36.80/84.28	85.80/**96.70**
Venous Insufficiency	0.00/97.21	0.00/97.05	29.03/**97.52**

Table 5.6: 10-fold cross validation accuracy results for the textual annotation, UMLS representation.

Co-morbidity	**Baseline** F-measure/Accuracy (%)	**CNB** F-measure/Accuracy (%)	**AdaBoost** F-measure/Accuracy (%)
Asthma	0.00/86.72	29.60/83.10	44.77/**95.46**
CAD	71.76/55.96	37.53/62.74	36.32/**71.19**
CHF	0.00/57.31	34.92/63.10	40.35/**81.41**
Depression	0.00/85.39	40.25/84.35	41.15/**87.67**
Diabetes	79.91/66.56	28.09/61.89	20.02/**66.85**
Gallstones	0.00/84.10	21.75/74.79	34.7/**88.63**
GERD	0.00/82.95	36.72/74.44	41.55/**91.94**
Gout	0.00/87.69	32.8/83.72	45.97/**96.55**
Hypercholesterolemia	0.00/56.58	30.85/54.97	39.9/**80.93**
Hypertension	16.98/73.81	34.00/96.84	40.12/**82.46**
Hypertriglyceridemia	0.00/97.38	24.60/96.84	37.30/**98.35**
OA	0.00/84.57	29.07/75.99	36.10/**89.43**
Obesity	0.00/57.94	30.29/56.12	33.65/**72.90**
OSA	0.00/84.26	29.97/76.13	39.95/**91.58**
PVD	0.00/86.23	25.17/80.44	44.25/**95.17**
Venous Insufficiency	0.00/97.21	26.77/81.07	44.2/**95.19**

Table 5.7: 10-fold cross validation accuracy results for the textual annotation, Genia NPs representation.

Model	Precision (%)		Recall (%)		F-Measure (%)	
	Micro	Macro	Micro	Macro	Micro	Macro
BOW	93.09	96.35	93.09	45.37	93.09	**45.79**
UMLS	90.15	94.45	90.15	43.39	90.15	43.84
NPs	73.55	83.03	73.55	31.96	73.55	32.24

Table 5.8: Results on the test set for the textual annotations where BOW - bag-of-words features; UMLS - features; NPs - Genia tagger noun phrases.

Chapter 6

SubGoal2: Identify Relevant Articles to Be Used for Building Systematic Reviews

Systematic reviews (SRs) are highly structured summaries of existing research in any given field, particularly in the medical domain, where they are a fundamental tool for evidence-based medicine (EBM). As stated in Sackett et al. (1996), EBM represents "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence-based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research."

SRs are valuable resources in health care. Clinical trials[1], which are research studies from where vital information in the medical domain emerges, are often the consequence and the cause of a SR. The conclusions a SR brings can lead to a new clinical trial design, while the results of various clinical trials end up being present in a SR. The information SRs contain can be used in diagnosis procedures, in the pharmaceutical domain for designing new drugs, and also in the hospital management field for improving the heath

[1] http://clinicaltrials.gov/

care system.

SRs are the result of a tedious process, due to the fact that the amount of information in medical publications continues to increase at a tremendous rate. Systematic reviews help to parse this growing body of information and distill targeted knowledge from it. The systematic review process, though typically more economical than primary research, is expensive in both time and human effort. Human reviewers are needed to manually screen references to determine their relevance to each given review. This process often entails reading thousands or even tens of thousands of abstracts from prospective articles. As the body of available articles continues to grow, this process is becoming increasingly difficult.

Common systematic review practices stipulate that two reviewers are used at the screening phases of a systematic review to review each abstract of the documents retrieved after a simple query-based search. After a final decision is made for each abstract (the two reviewers decide if the abstract is relevant or not the topic of review), in the next phase, further analysis (more strict screening steps) on the entire article is used in order to determine if the article is clinically relevant or not, to extract information, etc. The initial broad screening process is done using only the abstracts of the documents, whereas in the second one the reviewers might have access to the entire documents (often the documents are not available for free).

A systematic review must be exhaustive; the accidental exclusion of a potentially relevant abstract can have a significant negative impact on the validity of the overall review (Cohen et al., 2006). Thus the process is extremely labor-intensive.

The cost of human effort is high: reviewers need to read in parallel the entire collection of abstracts of the retrieved articles. My goal is to propose an automatic system that can help human judges in the process of triaging articles by looking only at abstracts and not the entire documents (that is, I focus on the initial screening phase). One reviewer will still read the entire collection of abstracts, while the other one will benefit from the help of the automatic system; this reviewer will have to label only the articles that will be

used to train the classifier, the rest of the articles will be labeled by the classifier. Ideally the proportion of articles that the reviewer needs to label for training the classifier will be small, in order to achieve a higher workload reduction.

I envision two ways to obtain the labels of the abstracts that will be used in training the classifier. The labels could be based only on the decisions made by the assisted reviewer, or they could represent the final decisions resulting from the work of both reviewers. Usually, if either reviewer believes that an article should receive further screening, it is labeled for inclusion. For the experiments performed in this paper, I followed this latter approach, using the labels obtained after the two reviewers' decisions are combined.

The decision process for the final labels when the two reviewers' opinions are used can be the same as the one used in the initial screening phase: if at least one reviewer agreed to include the abstract, the abstract will be labeled as included (the benefit of doubt plays an important role in the decision process). For the experiments performed in this thesis, I used the labels obtained after the two reviewers' decisions are combined, the latter-mentioned scenario. This approach should both maintain reliability of the systematic review and reduce the overall workload. With regard to reliability, even if one of the reviewers is assisted by an automatic classifier, the chances that both the human judge and the classifier exclude the same abstract will be approximately the same as if two human judges had directly reviewed the abstract. The reduction in workload is from the time required for the usual two passes through the whole collection of abstracts (by both humans) to only one full pass plus a lesser amount of activity by the classifier-assisted reviewer.

Figure 6.1 graphically presents the human-classifier workflow of building a systematic review, when the labels for training the classifier are taken after combining the decisions of both reviewers. Other scenarios in which a human-classifier workflow is used can be tried. For example, some of the abstracts labeled by the classifier can be double-checked by the assisted human reviewer, in this case the final label being given by the human

reviewer.

Other scenarios in which a human-classifier workflow is used can be tried. For example, some of the abstracts labeled by the classifier can be double-checked by the assisted human reviewer, Reviewer 2 in Figure 6.1, in this case the final label being given by the human reviewer. The focus and the workflow that I propose in this paper is the one described in Figure 6.1, in which there is no need for a human reviewer to check the labels given by the classifier.

An automatic system helping with the tedious process of deciding the relevance or non-relevance of each abstract could make systematic reviews easier, faster, more scalable, and more affordable to complete. Machine learning techniques could fulfill this need (Aphinyanaphongs and Aliferis, 2005). Specifically, the subfield of machine learning, called automatic text categorization is highly relevant for an intelligent systematic review system. The machine learning task that needs to be solved in order to help the systematic review process is a text classification task intended to classify abstracts as relevant or not relevant to the topic of review.

The methods presented in this work outline the use of machine learning in preparing systematic reviews. The hypothesis that guides this research is that it is possible to save time for the human reviewers and obtain good performance levels, similar to the ones obtained by humans, by replacing some of the manual screening of abstracts with an automatic classifier that will be able to determine the relevance of abstracts while requiring little training overhead. The goal is to show that such an automatic classifier can achieve acceptable levels of recall with a small number of false positives in the screening process. Also it is required to show that, when the system is combined with an appropriately-designed human reviewer workflow, a net reduction in manual effort can be achieved, while maintaining the required level of systematic review rigor. An easy way of integrating the machine learning capabilities in the workflow of creating systematic reviews is needed. These research experiments also want to determine whether the automatic classification of documents can be useful in systematic reviews on medical

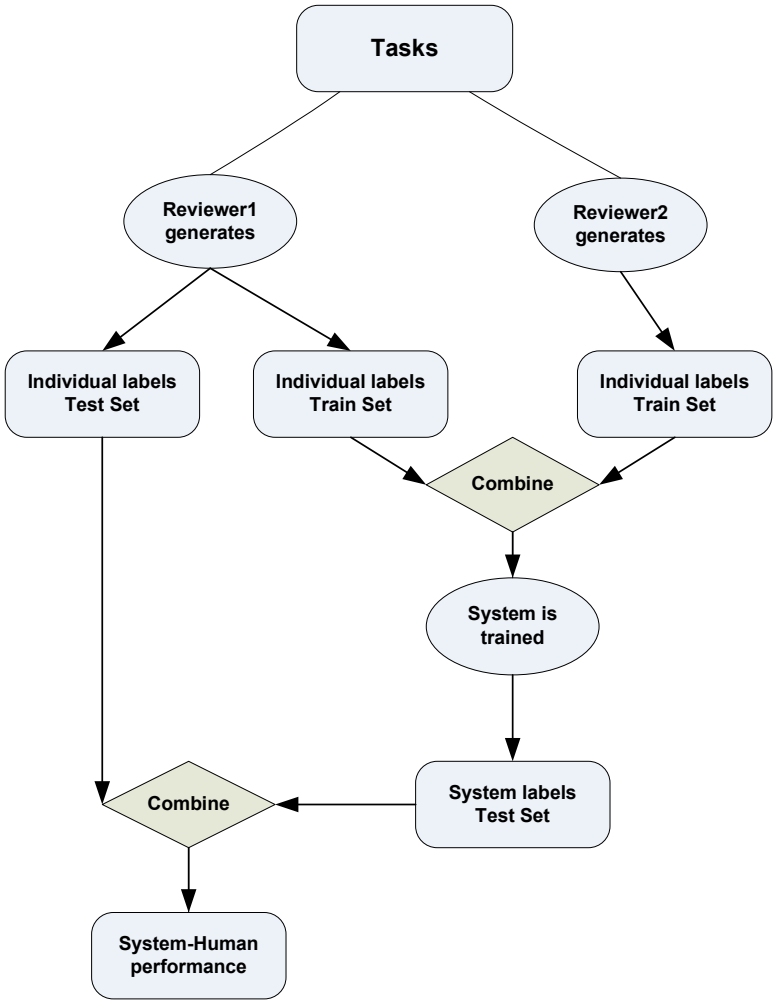

Figure 6.1: Embedding automatic text classification in the process of building a systematic review.

topics, and specifically if the performance of the automatic classification can be enhanced by using the particular protocol of questions employed by the human reviewers.

While this current work is focused on automatic ways of building SR, I acknowledge that SR, can help as additional source of information for other biomedical NLP-oriented tasks, e.g., diagnosis, identification of relations between medical concepts, disease classification, etc.

Part of this work has been published in (Frunza et al., 2010) and in (Frunza et al., 2011a). I contributed with my expertise to (Kouznetsov et al., 2009), (Matwin et al., 2010), (Matwin et al., 2011), but none of the experiments presented in these papers are part of this thesis, because the datat set was a different one and the methodology was different and focused on other issues.

6.1 Experimental Setting

Using machine learning as a technique to help in the automation of the systematic review process is a difficult task because of several issues that need to be taken into account. Some possible problems that might appear for this task are pointed out in the next lines.

Distribution of the data - in most systematic reviews, the number of relevant documents is much smaller than the number of non-relevant documents creating a class imbalance which can cause problems for the machine-learning algorithms.

Noise - when reviewers are not sure if the abstract is relevant or not, a final decision will be made during a second screening process where the entire document is reviewed. The benefit of doubt will affect the quality of the data used to train the classifier, since a certain amount of noise is introduced: abstracts that are in fact non-relevant can be labeled as being relevant in the first screening process due to the benefit of doubt.

Labeling cost - the labeling process is expensive both in human effort and money, and the machine learning approach will try to ease the human effort that is required.

Misclassification cost - if an abstract does not pass the first level of screening, it will not be taken into consideration for the second level of screening in the information extraction phase. Misclassification of a relevant abstract in the first screening process may have a profoundly negative impact on the validity of systematic review results (Cohen et al., 2006). It is, therefore, of high importance and it has to be taken into account as a critical factor in the evaluation technique.

Representation - due to the high number of abstracts, a consequence of the increasing rate of adding knowledge to the medical domain repositories, the machine learning representation has to take into account the huge number of attributes / words that can be extracted.

Training cost - if the manual overhead associated with training and tuning a machine learning classifier is too high, this may negate the value of the solution in terms of the overall amount of screening work saved. Similarly, if the training interface and process for the machine learning classifier is not simple enough to be accessible to a non-technical user, its relevance to the systematic review community will be negligible.

With the above-mentioned problems acknowledged this current work tries to address them with the exception of the misclassification cost.

The first experimental setting is based on standard text classification techniques as in (Cohen et al., 2006), while the second one is a new technique that takes into account the specifics of the systematic reviews - the set of questions and answers that are commonly part of the systematic review protocol. I believe that this approach in which the specifics of the task are taken into account represents a promising choice for obtaining desirable results.

The data set used for this task was presented in Chapter 3, subsection 3.2.1, Table 3.7.

6.1.1 Global text classification method

The first method that I propose in order to solve the text classification task that is intended to help a systematic review process is a straightforward machine learning approach. I trained a classifier on a collection of abstracts and then evaluated the classifier's performance on a separate collection of abstracts representing the test data set. The power of this classification technique stands in the ability to use a suitable classification algorithm and a good representation for the text classification task; Cohen et al. (2006) also used this approach. In (Matwin et al., 2010), our research group performed experiments with the data set used by Cohen et al. (2006) with a similar approach as my global text classification method. Further discussions on the choice of classification algorithms were published in (Matwin et al., 2011) and in the end the results suggested that the CNB classifier is suitable to use for such setting and task.

The same machine learning algorithms are used to train several classifiers, for the second method that is proposed, described in detail in subsection 6.1.2. With this method, additional knowledge from the systematic review protocol is used with an ensemble of classifiers trained on different data sets and with a final classification decision for each testing instance taking into account the predictions of the ensemble of trained classifiers. The idea of using the data divided into subsets that correspond to questions that are asked in the protocol came into consideration when looking at the broad topics and subjects that could be involved in building systematic reviews. When looking at the text of the questions that are involved, I noticed that sometimes they tackle different concepts. Dividing the problem and combining the end results seemed to us to be a suitable choice for the task. As I will show later in the results section, the intuition was right.

For the global text classification method, I used the entire collection of 47,274 abstracts were 20,000 were randomly taken to be part of the training data set and the remaining 27,274 represents the test set. This data set is described in 3, subsection 3.2.1. Cross-validation evaluation technique is not appropriate for this task since splitting the

data using a time-line stamp is a more realistic scenario when integrating a trainable automatic system in the human workflow. The fact that the size of the data is large is another reason why cross-validation does not necessarily suits the task. Cross-validation is usually used when the data set is not large enough to split it into training and test sets.

In order to evaluate the effect of the training set size, additional experiments were performed for different splits of the data set. The size varied from 10-50% in 10% increments. I decided not to include more than 50% of the data for training, because it would increase the workload of the human judge, and the final goal is to use as little training data as possible.

Feature engineering

A lot of effort was put in identifying what representation technique yields the best results. Various feature selection and data undersampling techniques were used.

InfoGain Feature Selection (Quinlan, 1986), is a method that evaluates the worth of an attribute by measuring the information gain with respect to each class. In the experiments when I used InfoGain as a feature selection method, I varied the number of the top selected features. I used the InfoGain feature selection implementation technique from the Weka package choosing a number of 50, 100, 150, 250, 300 and 500 top features (words that were present in the training data).

χ^2 Feature Selection (Yang and Pedersen, 1997) evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to each class. For the χ^2 feature selection experiments, I used different ratios of features for the Included and Excluded class, from the training data. I selected the top k1 χ^2 features that are exclusively Included (appeared only in the training abstracts that are classified as Included) and the top k2 χ^2 features that are exclusively Excluded (appeared only in

the training abstracts that are classified as Excluded) and used them as a representation for the data set. I varied the k1 and k2 parameters: from 10 to 150 for the parameter k1 and from 5 to 150 for the parameter k2 having a total number of attributes from a minimum of 20 to a maximum of 300. Similarly to the InfoGain feature selection, the Weka package was used to determine the top χ^2 word features.

Bi-Normal Separation Feature Selection (BNS) is a feature selection technique that measures the separation between the threshold occurrences of a feature in one of the two classes. It is described in detail in Forman (2002). The algorithm works with a ratio of features; it determines which are the most representative features for each class. I used a ratio of features that varies from 10 to 150 for the most representative features for the Included class and from 5 to 150 for the Excluded class. For some experiments the number of features for the Included class is higher than the number of features for the Excluded class. I have chosen to do so because I wanted to balance the distribution of classes in the training data set. This representation technique was used in our research group and showed to work well with binary feature values (Ma, 2007), representation that I used in my experiments as well.

I also ran additional experiments with this feature selection technique when I balanced the training data set. For these experiments, I used an undersampling technique. The Weka package supports this preprocessing technique. I decided to use the undersampling technique because it is demonstrated to work better than the over-sampling technique for classification tasks with data sets that have a class that is predominant (Drummond and Holte, 2003). The undersampling technique creates a balanced training data set from the original data set that I worked with. The instances of the minority class are kept as they are, and the number of instances in the majority class is reduced to the number of instances in the minority class. The selection of the instances of the majority class that will be part of the new subset is done randomly.

Results for the experiments that were performed for these methods are presented in

section 6.2.

6.1.2 Per-question classification method

I made the transition from the first method to the second method by first combining the global method with knowledge from the answers that the reviewers gave to the questions, for each abstract. I used additional features in addition to the features extracted by the global method. The three additional features correspond to the possible answers a question can have: "Yes", "No", "N/A". Results obtained for this experiment are described in section 6.2.

The second method proposed takes into account the specifics of the systematic review process. More exactly, it takes advantage of the set of questions that the reviewers use in the process of deciding if an abstract is relevant or not. These questions are created in the design step of the systematic review and almost all systematic reviews have such a set of questions, similar to the ones presented in Table 3.7 from Chapter 3, section 6.2, e.g., "Is this a primary study?", "Is this a review?". For instance, Cohen et al. (2006) represent such information in the form of a closed set of reasons for the exclusion of a specific abstract, but this information was not used by the classifier developed in their work.

In the worst case, the design of a systematic review has only one question and the method will be similar to the global text classification technique that I presented earlier. The algorithm used for the per-question method is described bellow.

1. Choose a subset of questions - the ones that have a YES/NO response.

2. Create a training data set for each question Q using only the abstracts that have an answer to the question and removing abstracts for which the answers did not agree.

3. Train a CNB classifier for each question using the new training data sets.

4. Combine the prediction accuracies of the CNB classifiers to predict the class of each testing instance.

I have chosen to use only the questions that have inclusion/exclusion criteria, because they are the ones that are important for reviewers to make a decision for an abstract. To collect training data for each question, I used the same training data set as in the previous method (but note that not all the abstracts have answers for all the questions; therefore the training set sizes differ for each question). I also kept the same test data set.

In order to create a training data set for each question separately, I removed the abstracts for which I had a disagreement between the human experts - two different answers for a specific question, because they represent noise in the training data. I need to train classifiers only on reliable data, when possible. Table 6.1 presents the subset of questions that I selected and the number of instances (abstracts) present in each training data set for each class.

Q1 - Is this article about a dissemination strategy or a behavioral intervention?	14,057	1,145	12,912
Q3 - Is the population in this article made of individuals 65-year old or older or does it comprise individuals who serve the elderly population needs (i.e. health care providers, policy makers, organizations, community)?	15,005	7,360	7,645
Q4 - Is this a primary study?	8,825	6,895	1,930
Q7 - Is this a review?	6,429	5,640	789

Table 6.1: Data sets for the per-question classification method.

Combining the per-question classifiers

For each of the questions from Table 6.1, I trained a CNB classifier on the corresponding data set. I used the same representation for the per-question classifiers, as I did for the

global classifier: BOW, UMLS (the concepts that appeared only in the new question-oriented training data sets), and the combination BOW+UMLS. I used each trained model to obtain a prediction for each instance from the test set; therefore each test instance was assigned four prediction values of 0 or 1. The predictions have values of 0 or 1. To assign a final class for each test instance, from the prediction of all four classifiers, the class of a test instance is decided according to one of the following four schemes:

1. If any one vote is *Excluded*, the final class of a test instance is *Excluded*. This is a 1-vote scheme, e.g., any one classifier voted excluded.

2. If any two votes are *Excluded*, the final class of a test instance is *Excluded*. This is a 2-vote scheme.

3. If any three votes are *Excluded*, the final class of a test instance is *Excluded*. This is a 3-vote scheme.

4. If all four votes are *Excluded*, the final class of a test instance is *Excluded*. This is a 4-vote scheme.

With the final prediction for each test instance, I computed the confusion matrix by comparing the predicted class with the actual class assigned by the reviewers, and I calculated the performance measures. When I combined of the classifiers, I gave each classifier an equal importance, when using the majority vote to decide the final classification.

For all experiments, the results for both methods are reported for the CNB classifier. Experiments on the training data set using 10-fold cross validation were performed for several classifiers, but the CNB was the one that consistently outperformed the others.

6.2 Results

This section presents the results obtained with the two methods described above.

When performing the evaluation for the task of classifying an abstract into one of the two classes included (relevant) or excluded (non relevant), two objectives are of great importance:

1. Ensure the completeness of the systematic review (maximize the number of relevant documents included). It translates in good levels of recall.

2. Reduce the reviewers' workload (maximize the number of irrelevant documents excluded). It translates in good levels of precision.

I observe that objective 1 is more important than objective 2 and this is why I decided to report recall and precision for the included class. I also report F-measure, since I am dealing with imbalanced data sets.

The results are reported on the initially data split of the 47,274 abstracts: 20,000 were randomly taken to be part of the training data set and the remaining 27,274 represents the test set.

Besides the evaluation measures commonly used for reporting results on similar tasks, I also used a measure introduced by Cohen et al. (2006). Their work is mostly focused on the elimination of non relevant documents. As their main goal is to save work for the reviewers involved in systematic review preparation, they define a measure, called work saved over sampling (WSS), that captures the amount of work that the reviewers will save with respect to a baseline of just sampling for a given value of recall.

The formula is:

$$WSS = \frac{(TE + FE)}{(TE + FE + TI + FI)} - 1 + \frac{TI}{(TI + FE)}$$

.

where:

Let TI = the number of true inclusions;

Let FE = the number of false exclusions;

Let FI = the number of false inclusions;

Let TE = the number of true exclusions.

The idea is that a classifier returns, with high recall, a set of abstracts, and only those abstracts need to be read to weed out the non-relevant ones. The savings are measured with respect to the number of abstracts that would have to be read if a random baseline classifier was used. Such baseline corresponds to uniformly sampling a given percentage of abstracts (equal to the desired recall) from the entire set. In (Cohen et al., 2006), the WSS measure is applied to report the reduction in reviewer's work when retrieving 95% of the relevant documents, but the precision was very low. This work is focused on developing a classifier for systematic review preparation, relying on characteristics of the data that were not included in the Cohen et al. (2006)'s method and therefore I cannot perform a direct comparison of results here. Also, the data sets that they used in their experiments are significantly smaller than the one used in this study.

Baselines For comparison purposes I decided to present two extreme baselines and a random-baseline classification that takes into account the distribution of the two classes in the training data set. They are not reasonable to be used in real-life applications therefore I need to make sure the classifiers obtain better results than these baselines.

The random-baseline classifier will randomly generate a real number between 0 and 1 and uses the prevalence of the two classes to decide. If the generated number is less or equal to the threshold the label is relevant, otherwise it is non-relevant.

The extreme baseline that will classify everything as included can be seen as a simple query-based system used for retrieving articles relevant to a certain topic. In fact, these types of systems are commonly now in place for building systematic reviews, since everything that the retrieved by the query is considered relevant. An automatic classification system must, obviously, exceed the results attained by these two baselines to be of any value.

Table 6.2 presents the results for the baselines.

Baseline	Recall	Precision	F-measure	WSS
Include_All	100%	15%	26.2%	0%
Exclude_All	0%	100%	64.2%	0%
Random baseline	8.9%	15.4%	67.8%	0.23%

Table 6.2: Results for the baseline model.

6.2.1 Experimental results for various splits of the training and test sets

Global method

As mentioned earlier, experimental results are performed in order to determine the best training-testing split using the global method. These results are presented in Table 6.3.

Train-test split	True Inc	False Inc	True Exc	False Exc	Recall	Precision	F-measure
10%-90%	1,137	1,179	34,898	5,332	17.58%	49.09%	25.89%
20%-80%	2,721	3,665	28,404	3,029	47.32%	42.61%	44.84%
30%-70%	2,969	5,087	23,019	2,017	59.55%	36.85%	45.53%
40%-60%	2,692	5,022	18,135	1,425	65.39%	34.90%	45.51%
50%-50%	2,316	4,524	14,615	1,092	67.96%	33.86%	45.20%

Table 6.3: Results for the global method using the BOW representation technique for various splits of the train and test data sets.

Per-question method

Similar to the experiments performed with the global method when varying the training data sizes, experiments with various data sizes are executed for the per-question technique. The classifier used for these experiments is the 2-vote scheme with BOW+UMLS feature representation (the classifier that obtained the best performance

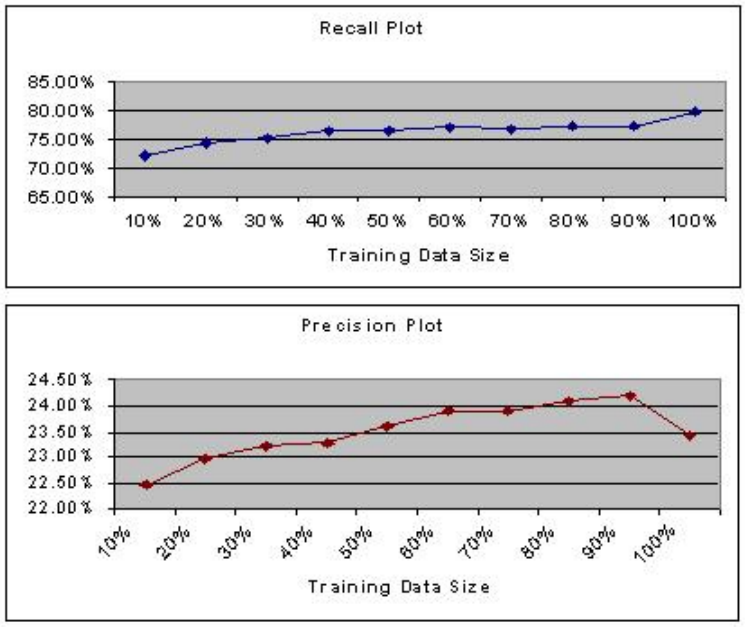

Figure 6.2: Recall and precision plots when varying the training size for per-question technique.

level, see Table 6.6). These experiments are intended to show how much variation exists in the performance levels when the training data size varies. The original size of the training data set for each question used by the per-question method is the one presented in Table 6.1. The experiments that I performed use an increment of 10% of the training data when starting with only 10% of the original size. For all experiments, the test data remains the same. Figure 6.2 presents the recall and precision plots for the experiments described above.

The results of Figure 6.2 show that the level of precision and recall tends to increase in correlation with a bigger training data. For both the recall and precision measure, the differences in results are rather small. A jump in recall is observed when I used the collection of abstracts on which I run most of the experiments; this validates the choice of training data size.

Global method combined with answers from the protocol questions. For the experiment in which I combined the global method with the answers for the questions involved in the systematic review protocol, I obtain a recall value of 67.5%, a precision value of 37.9%, and an F-measure value of 48.5%. Comparing these results with the ones that I obtained with the global method, we can observe that the recall value, the main evaluation measure, did not improve. The precision result is one of the best, but still the recall is not acceptable.

6.2.2 Results for the global method

In this subsection, I present the results that were obtained using the global method with the three representation techniques: bag-of-words (BOW), UMLS features, and a combination of the two with CNB as classification algorithm and all the feature selection techniques that were tried.

I experimented with other classifiers from Weka as well (such as decision trees, support vector machine, instance-based learning, and boosting), and the results obtained with CNB were better than those with the other classifiers.

Naïve Bayes (NB) is a classifier that is known to work well with text. It is fast and easy to integrate in a more complex system. Its performance is comparable to that of the SVM classification, and a previous study (Colas and Brazdil, 2006) showed that the NB classification algorithm works better than SVM with a high number of features. In that study the authors compared SVM, NB, and k-nearest neighbor (kNN) classification algorithms for text classification tasks. They found that all classifiers obtained comparable results, but that NB worked best for various sizes of the training data set. The difference in results between SVM and NB was not significant, but, in view of the running times and parameter settings, the NB classifier is definitely a more suitable choice than SVM for the task.

To get a clear image of the results, I show the numbers in the confusion matrix as well to give the reader a better understanding of the workload reduction when using the

automatic classifiers.

Results for the feature selection techniques

InfoGain From the results obtained when using InfoGain as a feature selection algorithm, I choose one set of representative results to present in Table 6.4. The number of features that I used for the experiment is 300 and the classifier is CNB.

As we can observe from the reported results, almost all abstracts were classified as Included. This performance results in a good recall, but with the cost of having a poor precision. Form the practical point of view, these types of results will not reduce too much the workload of the human reviewers. They will have to read almost the entire initially retrieved collection and remove a large number of abstracts that are not relevant to the topic of review.

I also run experiments with the same feature selection settings when using a Naïve Bayes classifier. I mention some of the results that I obtained with the Naïve Bayes classifier for a comparison in performance with the most commonly used classification algorithm for text data sets. The results obtained were in the opposite direction to ones obtained for CNB, all abstracts were classified as Excluded. This is similar to the extreme baseline when everything is classified as non relevant to the topic of review, and it is a result that is not realistic to be used in a systematic review. I believe that this is the case because the CNB classifier tries to compensate for the class distribution and gives more credit to the minority class, while the Nave Bayes classifier will let the majority class overwhelm the classifier, resulting in almost all abstracts being excluded.

χ^2 Table 6.4 shows a representative result for all the experimental results that I obtained using these BOW representations and with k1=150 and k2=150 feature ratio. The results that I obtained are similar to the ones obtained using InfoGain feature selection. Again the CNB classifier favored the Included class; this triggers the same observations for the systematic review process results. The Naïve Bayes classifier performed similarly to the

extreme baseline in which all the abstracts are classified as Excluded.

BNS Representative results for all these experiments using the CNB classifier are presented in Table 6.4 second part, the third line of results. The feature ratio that was used is 10:8.

The results obtained with this feature selection for the BOW representation are slightly better (the balance between recall and precision) than the ones that I obtained with the two previous methods for both CNB and Naïve Bayes classifiers (a high recall and a low precision).

In some additional experiments when I used the BNS feature selection from a balanced training data set (I used the undersampling technique as explained to create a balanced training data set), the results improved (in terms of average recall-precision and WSS) compared to the results of the initial BNS feature selection. Table 6.4 also presents representative results obtained with the BNS feature selection and the Naïve Bayes classifier from a balanced training data set with 50:25 feature ratio.

	χ^2 (150:150)	**InfoGain** (300)	**BNS** (10:8)	**BNSB** (50:25)
True Inc.	3,819	3,875	2,690	2,220
False Inc.	19,233	19,638	13,905	8.741
True Exc.	3,924	3,518	9,253	14,416
False Exc.	298	242	1,427	1,897
Recall	92.8%	94.1%	65.3%	53.9%
Precision	16.6%	16.5%	16.2%	62.2%
F-measure	28.0%	28.0%	25.0%	57.7%
WSS	8.2%	7.9%	4.5%	13.7%

Table 6.4: Representative results obtained for various feature selection techniques.

Besides all the results presented above for various feature selection techniques, I also

tried to boost the representative feature for the Included class hoping to re-balance the imbalance present in the training data set. To perform these experiments, I selected the top k χ^2 word features and then added to this set of features the top $k1$ χ^2 representative features only for the Included class. The parameter k varied from 50 to 100 and the parameter $k1$ from 30 to 70. I performed experiments when using the original imbalanced training data set and using a balanced data set as well, with both CNB and Naïve Bayes classifier. The results that I obtained for these experiments were similar to the ones when I used the previous feature selection techniques.

In addition, to boosting the Included class features, I also boosted the Excluded class features, with a number of features in the same range as for the case when I boost the Included class features. There was no significant difference in the results.

From all the feature selection algorithms and the different representation techniques that I tried, a general conclusion that I observed is that, due to the noise and the imbalance in the data sets the results obtained are not satisfying for the problem of keeping the relevant documents and discarding many of the non-relevant ones. This conclusion is similar to previous research done for the same task when using different data collections.

Global method

The results obtained with the whole set of features for all three representation techniques are presented in Table 6.5.

For the BOW representation, 23,906 features were selected, without any feature selection technique. To determine the UMLS concepts I used the Meta-Map system. From the whole training abstracts collection, 459 UMLS features were identified. Analyzing the results from Table 6.5, in terms of recall, the UMLS representation obtained the best recall results, 67.8% for the global method but much lower precision, 23.8% than BOW representation, 34.9%. The hybrid representation, BOW+ UMLS features had similar results with the BOW alone. Recall increased a bit for the hybrid

	BOW	**UMLS**	**BOW+UMLS**
True Inc.	2,692	2,793	2,715
False Inc.	5,022	8,922	5,086
True Exc.	18,135	14,235	18,071
False Exc.	1,425	1,324	1,402
Recall	65.3%	67.8%	65.9%
Precision	34.9%	23.8%	34.8%
F-measure	45.5%	35.2%	45.5%
WSS	37.1%	24.9%	37.3%

Table 6.5: Results for the global method.

representation compared to BOW alone, 0.6% but its value is still not acceptable. I conclude that the levels of recall, the main objective for this task, were not acceptable for a classifier to be used as replacement of a human judge in the workflow of building a systematic review. The levels of precision that I obtained with the global method are acceptable but they cannot substitute the low level of recall.

6.2.3 Results for the per-question method

Compared with the global method, the results obtained by the per-question method, especially the ones for 2 votes are the best so far in terms of the balance between the two objectives. A large number of abstracts that should be excluded are classified as Excluded, whereas wrongly excluding very few abstracts that should have been included (a lot fewer than in the case of the global classification method).

Table 6.6 presents all results obtained with the second method proposed in this current work, the per-question method.

The 2-votes scheme performs better than the 1-vote schemes, because of potential classification errors. When the classifiers for two different questions (that look at two different aspects of the systematic review topic) are confident that the abstract is not

relevant, the chance of correct prediction is higher; a balance between excluding an article and keeping it as relevant is achieved. When using the classifiers for 3 or 4 questions, the performance goes down in terms of precision; a higher number of abstracts get classified as included because some abstracts do not address all target question of the review topic.

For the per-question technique the recall value peaked at 99.2% with the 4-vote method BOW and BOW+UMLS representation techniques when using CNB as classifier. At the same time the lowest values of precision for the per-question technique, 15.6% is obtained with the same experimental setting. It is important to aim for a high recall, but not to dismiss the precision values. The difference of even less than 2% in precision values can cause the reviewers to read additional thousands of documents, as observed in the confusion matrices for 2-vote, 3-vote and 4-vote methods.

From the confusion matrix in Table 6.5 for the 2-vote method and the 3- and 4-vote method, I observe the high difference in the number of documents a reviewer will have to read (the falsely included documents). The difference in precision from 24.1% for the 2-vote method to 15.6% for the 4-vote method makes the reviewer go through 11,988 additional abstracts. The best value for the WSS measure for the per-question method is achieved by the 2-vote scheme. The result it is lower than the one obtained by the global method, but the recall level is too low for the global method. Therefore, I still keep as a potential winner the 2-vote scheme for the per-question classification technique.

6.2.4 Results for human-machine workflow

In Figure 6.1, I envisioned the way we can use the automatic classifier in the workflow of building a systematic review. In order to determine the performance of the human-machine workflow that I propose, I computed the recall values when the human reviewer's labels are combined with the labels obtained from the classifier. The same labeling technique is applied as for the human-human workflow: if at least one decision for an abstract is to include it in the systematic review, then the final label is included.

I also calculated the evaluation measures for the two reviewers. The evaluation

measures for the human judge that is kept in the human-machine workflow, Reviewer 1 in Figure 6.1, are 64.29% for recall and 75.20% for precision. The evaluation measures for the reviewer that is to be replaced in the human-machine classification, Reviewer 2 in Figure 6.1 are 59.66% for recall and 85.09% for precision. The recall value for the two human judges combined is 85.26% and the precision value is 85%. As we can observe, the recall value for the second reviewer, the one that is replaced in the human-classifier workflow is low. Like the conceptual idea that stands behind all committees of judges, the power of many is stronger than the power of one, the results when both reviewers participate are much higher than the results for each of them individually.

The results that I obtain when both reviewers are used are much higher than each of the results of them individually. In Table 6.7, I present precision and recall results for the symbiotic model, for both the methods. In these results we can clearly see that the 2-vote technique is superior to the other voting techniques and to the global method. For almost the same level of precision, the level or recall it is much higher. These observations support the fact that the extra effort spent in identifying the most suitable methodology pays off.

The fact that I keep a human in the loop makes the method acceptable as a workflow for building a systematic review. I always aim for a high level of recall while keeping in mind that a good level of precision is important; low precision means more human effort in the next stage of the systematic review.

6.3 Discussion and Conclusions

6.3.1 Discussion

The global method achieves good results in terms of precision, while the recall (objective 1) is low. In terms of recall, almost none of the representations (UMLS or a combination of UMLS and BOW features) outperformed the recall values obtained by the per-question method. The best results for the task were obtained using the per-question method with

the 2-vote scheme.

The results obtained by the 3-vote scheme UMLS representation are close to the results obtained by the 2-vote scheme, but looking at F-measure results the 2-vote scheme outperforms the 3-vote one. Even thought there is not a completely clear decision for the results obtained by the per-question method, in my view the 2-vote scheme with or without UMLS features is the best method for the task. There are other per-question settings that obtained better levels of recall.

The per-question technique is more robust and it offers the possibility to choose the desired type of performance. If the reviewers are willing to read almost the entire collection of documents, knowing that the recall is high, then a 3 or 4-vote scheme can be the set-up (though the 3 or 4- vote method is not likely to achieve 100% recall because it is very rare that an abstract contain answers to three or four of the questions associated with the systematic review). If the reviewers will like to read a small collection being confident that almost all are relevant, then a 1-vote scheme can be the set-up required. The per-question method that I described confirms the fact that a committee or an ensemble of classifiers is better than one classifier; this conclusion is supported in the machine learning literature (Dietterich, 1997).

When I combine the human and the system results, I obtain a major improvement in terms of recall. I base the discussion for the human-machine results for the experiment that is performed for the classifier that obtained the best results, the 2-vote scheme with a BOW+UMLS representation technique. From the total of 834 falsely excluded abstracts, 537 are identified as relevant by the first human reviewer, the one that reads and labels the entire collection of retrieved abstracts. These results show that the overall recall of the human-machine system is 92.7%, while the precision level is 17%. When combining the human and classifier decisions, the precision level decreased a bit compared to the one that the machine obtained. I believe that this is the case because some of the abstracts that the classifier excluded were included by the first human reviewer and, with this decision process in place, the level of precision dropped. The goal of improving the recall

level from the first level of screening is achieved, since when both the classifier and the human judge are integrated in the workflow, the recall level jumps from 79.7% to 92.7%.

As stated by previous studies done in our research group by Kouznetsov et al. (2009), the level of recall of human judges usually varies from 90-95%. I can conclude that the two human judges involved in the process of building the review corrected each other's mistakes, and in the end, together they were able to identify all relevant articles that ended up being the building blocks for the systematic review. The low levels of precision are corrected in the second level of screening where more refined questions are asked so that only the true relevant articles are identified as being relevant.

I believe that the low level of precision that is obtained for the human reviewer, for the human-classifier workflow, and for the classifier, is due to the fact that I run the experiments for the first screening phase. As mentioned earlier in this chapter, two screening phases are implied when building a systematic review. The objective is to achieve high performance levels of recall in the initial screening phase, knowing that once the abstract will be labeled as non-relevant in the initial screening phase, it will not be reviewed afterwards and it will be definitively excluded from the systematic review. The reviewers' performance, especially the precision values will raise in the second screening phase, because the information available to make the decision is richer, and the set of questions that the reviewers need to answer are more fined-grained in details and relevance to the topic of review. The aim for this work is to show that in a human-machine workflow used in the first screening phase a good level of recall can be obtained and the workload of this tedious process is reduced. The results obtained especially the ones when I combine the human and classifier decisions, show that the goals are met.

I believe that further investigations are required to fully replace a human reviewer with an automatic classifier, but the results that I obtained with the per-question method encourage us to believe that this is a suitable solution for reaching the final goal. In this paper, I looked at various methods that I thought are suitable for the task. As the results show, the task is not as trivial as it might seem, and it can be considered as a second

level of triage for abstracts that were retrieved using a simple query search. The task can be a challenge even for human experts. This is supported by the number of reviews that are released every year by the companies that work in this field (e.g., the Cochrane Collaboration[2]).

6.3.2 Conclusions and future work

In this chapter, I presented two methods that can be used in automatic classification of abstracts for a systematic review task. The two methods are: 1) the global method, where a classifier is built based on the labeled training set of relevant and non-relevant abstracts, and 2) an alternative method, exploiting the protocol often used by human reviewers in screening abstracts for systematic reviews. This protocol consists of asking a series of simple questions for each abstract and aggregating the answers to these questions into a binary decision.

In the current work, in order to achieve this particular subgoal of this thesis, I addressed most of the issues mentioned earlier in this chapter. The skewed class distribution is taken into account by using special classifiers (CNB). Noise in the data is reduced by removing unreliable instances, when possible (in the per-question technique). The evaluation measures pay attention to the importance of misclassifications.

Throughout the experiments, I showed that automatic classification can help in the process of preparing systematic reviews, and that the performance of the machine learning can be improved by exploiting the specific protocol. More work needs to be done before replacing a human judge with a fully-automatic system, however, the results are encouraging enough to partly replace one of the two human experts with a classifier. The best results were obtained without any feature selection and I see this as a positive aspect since the classifier is easier to integrate in the systematic review system.

I also show that for a real-life data set, the systematic-review protocol can help in building better automatic models to help the tedious process of human experts when

[2]http://www.cochrane.org/index.htm

building systematic reviews. I show that machine learning techniques can be used in the workflow and that a reduction in the workload is achieved. The fact that I propose a human-machine workflow that keeps one judge in the loop guarantees the fact that the papers the human judge includes will not be rejected by the workflow. It also makes the proposed system be accepted as a valid alternative to the human-human process of building systematic reviews.

Possible future work can be focused on reducing the amount of training data required, by using active learning techniques to select the best cases for training. The system will ask a human reviewer to label a small amount of training data that would be most useful in training a classifier. Another direction of future work is to use the classifier in updating existing systematic reviews. The classifiers can de deployed to identify newly published articles from Medline that are relevant to the topic of the systematic review. This way, updated information can be added to an already existent review on a certain topic.

Other directions of future work are to run experiments when I weight in a different way the decision of each classifier in the per-question method, and to combine the classifiers' decisions by using a meta-classifier instead of a voting technique.

Finally, it will be interesting to extend the Machine Learning approach to the second stage of systematic review process, in which the full texts of the papers are screened for relevance. I want to point out that since the full text representation will result in much less sparseness of the BOW representation, the results are likely to be significantly stronger than in the first phase.

1-Vote	**BOW**	**UMLS**	**BOW+UMLS**
True Inc.	1,262	1,222	1,264
False Inc.	745	2,266	741
True Exc.	22,412	20,891	22,416
False Exc.	2,855	2,895	2,853
Recall	30.6%	29.6%	30.7%
Precision	62.8%	35.0%	63.0%
F-measure	41.2%	32.1%	41.2%
WSS	23.2%	16.8%	23.3%
2-Vote	**BOW**	**UMLS**	**BOW+UMLS**
True Inc.	3,181	2,603	3,283
False Inc.	9,976	9,505	10,720
True Exc.	13,181	13,652	12,437
False Exc.	936	1,514	834
Recall	77.2%	63.2%	79.7%
Precision	24.1%	21.5%	23.4%
F-measure	36.8%	32.0%	36.2%
WSS	29.0%	18.8%	28.4%
3-Vote	**BOW**	**UMLS**	**BOW+UMLS**
True Inc.	3,898	3,480	3,890
False Inc.	18,915	16,472	18,881
True Exc.	4,242	6,685	4,276
False Exc.	219	637	227
Recall	94.6%	84.5%	94.4%
Precision	17.0%	17.4%	17.0%
F-measure	28.9%	28.9%	28.9%
WSS	11.0%	11.3%	11.0%
continued on next page			

continued from previous page			
4-Vote	**BOW**	**UMLS**	**BOW+UMLS**
True Inc.	4,085	3,947	4,086
False Inc.	21,946	20,869	21,964
True Exc.	1,211	2,288	1,193
False Exc.	32	170	31
Recall	99.2%	95.8%	99.2%
Precision	15.6%	15.9%	15.6%
F-measure	27.1%	27.2%	27.0%
WSS	3.7%	4.8%	3.7%

Table 6.6: Results for the per-question method.

Method	**BOW**	**UMLS**	**BOW+UMLS**
Global	17.9%/87.7%	17.0%/88.6%	17.9%/87.7%
1-Vote	17.1%/75.3%	16.5%/74.8%	17.1%/75.4%
2-Vote	17.1%/91.6%	16.4%/86.6%	17.1%/92.7%
3-Vote	15.8%/97.9%	15.8%/94.2%	15.8%/97.8%
4-Vote	15.3%/99.6%	15.4%/98.3%	15.3%/99.6%

Table 6.7: Precision/recall results for the human-classifier workflow.

Chapter 7

SubGoal3: Predict Functional Properties of Genes by Only Using Biomedical Articles

In the past years, gene-specific information has become a valuable resource for the biomedical-natural language processing community. Proof stands in the research-based shared tasks like KDD Cup[1] and BioCreative[2] that are directed to problems that involve gene-type information.

Genes represent the blueprint of every being on the planet and they often represent the only source of information that contains the answers to different problems, especially medical ones. Researchers from medical data mining domains try to combine their knowledge and identify information that can help in the advance of life science.

Scientific published papers still remains one of the primary sources of information for performing research in almost any domain. An unusual mixing of bioinformatics and NLP called BioNLP has become of great interest for scientist from both ends.

BioNLP has become a domain of great interest due to the growing knowledge

[1]http://www.sigkdd.org/kddcup/index.php

[2]http://www.pdg.cnb.uam.es/BioLink/BioCreative.eval.html

repositories (e.g., PubMed, Medline, etc.), due to biologists' interest in using text as an additional source of information for solving biology specific tasks, due to the challenges that the NLP research faces (e.g., tokenization, sentence boundary detection, name entity recognition, etc.) when dealing with biological and medical domain specific texts.

The goal of the research presented in this chapter is to show the "power" of textual information when used as a single source of information for the yeast gene regulation prediction task. The task determines which proteins affect the activity of AHR system by learning a model that can accurately predict how active the AHR system is when single genes are knocked out. My thesis presents experiments and results obtained when models are trained on a single source of information: the text of the Medline abstracts that mention the genes involved in the biological experiments. Various textual representations and classification algorithms are used. This task was a competition track for the KDD Cup[3] 2002. Part of this work has been published in (Frunza and Inkpen, 2008b).

7.1 Experimental Setting

The task that is tackled in this work is a biology-specific task, yeast gene regulation prediction, solved by using textual information as the only source of information. The organizers of the KDD Cup competition provided data obtained from experiments performed on a set of yeast strains in which each strain contains a single gene that is knocked out (a gene sequence in which a single gene is inoperative). Each experiment had associated a discretized value of the activity of the AHR system when a single gene was knocked out.

The data set that I used is presented in Chapter 3, subsection 3.2.2. It consists of Medline articles annotated with three classes that describe the aryl hydrocarbon receptor systems' response: "**nc**" the activity of the hidden system was not significantly different than the baseline (the wild-type yeast); "**control**" indicates that the activity of the

[3]http://www.biostat.wisc.edu/ craven/kddcup/winners.html

hidden system was significantly different than the baseline for the given instance, but that the activity of another hidden system (the control) was also significantly changed compared to its baseline; the "**change**" label describes instances in which the activity of the hidden system was significantly changed, but the activity of the control system was not significantly changed.

Some characteristics of the data need to be taken into consideration in order to make suitable decisions for choosing the trainable system/classifier, the representation of the data, etc. Missing information is a characteristic of the data set. Not all the genes had the location and function annotation, the protein-protein interaction information, or abstracts associated with the gene name. Besides the missing information, the high class imbalance is another fact that needs to be taken into account.

The method that I propose to solve the biology task is using Machine Learning (ML) classifiers suitable for text classification and various feature representations that are known to work well for data sets with high class imbalance.

From the data that was released for the KDD competition only the abstracts will be considered as source of information to solve the task, identify if the genes are involved or not in the activity of the hidden biological system. No other information, e.g., function, location, protein-protein interaction, is considered since the focus is only on the level of knowledge that a learning algorithm can extract from abstracts.

Not all the genes that are present in the task have associated abstracts. In fact missing pieces of information is a characteristic of this data set. Due to this constraint, I will experiment only with the genes (both from the training and test sets) that have at least one abstract associated with them. This is a constraint of this data set (the constraint can extend to other tasks as well, but in general for most of the "interesting" genes, textual information can be found in the well-known repositories). Brady and Shatkay (2008) propose three methods for finding textual data for biology specific entities (e.g. proteins). These methods can be used in future work to extract information for all the genes.

The ML task that needs to be solved, in the context described above, is a two-class classification task. The two classes are "Positive" versus "Negative", with a "narrow" and "broad" definition for the positive class. The task becomes a two-class classification problem for two different distributions for the positive and negative class on the same data.

Further aspects of the task need to be taken into consideration: the representation techniques, the classification algorithms, and the evaluation method for a ML task with highly imbalanced data sets. In the following sections, I will mention how I address each of these concerns and what are the final decisions.

Table 7.1 presents a summary of the training and test sets that I use in the experiments. The new data sets are subsets of the original data sets (Table 3.8 from Chapter 3). They consist only the examples (genes) that had abstracts associated with them.

Data set	**Narrow Pos**	**Neg**	**Broad Pos**	**Neg**
Training	24 (37)	1,435 (2,980)	51 (83)	1,408 (2,934)
Test	11 (19)	715 (1,469)	30 (43)	696 (1,445)

Table 7.1: Summary of the data for the ML experiments for the two definitions of the positive class. In brackets are the original sizes of the data sets.

As a general observation, the new data sets had a loss of approximately 48% for the Negative class and 60% for the Positive class with regard to the original data. The class imbalance is on average of 1:62 for the narrow class and 1:25 for the broad class, both the training and the test data sets.

The majority of the genes have only one abstract associated. There are some genes that have many abstracts, for example, more than 200 genes have 22 abstracts each (see

Figure 7.1).

7.1.1 Data representation

BOW — For this representation, I have chosen to experiment with both, binary and frequency feature values. The selected features are words delimited by spaces and simple punctuation marks that appeared in the training collection of abstracts and are not part of an English list of stop words, the same as the one described in Chapter 4.

For the binary representation the value of a feature is either 1 or 0. 1 represents the fact that the feature word is present in an instance/abstract; the value is 0 otherwise. For the frequency representation the features present in an instance have associated the number of times they appear. If the feature does not appear the value of the feature is 0. We have chosen to run experiments with both representations because it is not clear from the literature which one can lead to better results.

UMLS — Besides the UMLS concepts extracted by the MetaMap system in the same manner as described in Chapter 4, I also used **UMLS phrases**. For this representation, features are considered to be the actual noun phrases that the MetaMap system identifies when it parses the texts. For example for the noun phrase "to an increased risk" the UMLS concept representation will have as feature "*Qualitative Concept, Quantitative Concept*" while in the UMLS phrases representation, it will have as feature the phrase to "**an increased risk**".

I also used **UMLS Relations** as representation type for this task. The Semantic Network present in the UMLS provides a generalization of the existing relations between concepts. I used as features for the third UMLS-based representation the relations between all pairs of concepts identified for a given abstract.

Similarly to the BOW representation, a binary and a frequency representation for all three UMLS types of representation is used.

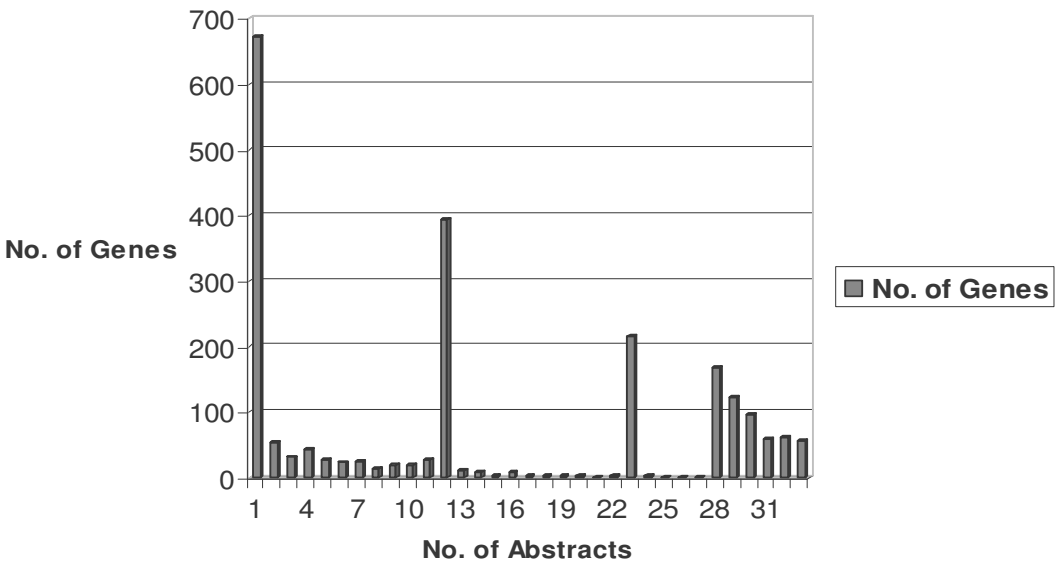

Figure 7.1: Embedding automatic text classification in the process of building a systematic review.

InfoGain Feature Selection InfoGain evaluates the worth of an attribute by measuring the information gain with respect to each class. For all three types of representations described earlier, I run experiments using only the most informative features. I used the InfoGain feature selection implementation technique from the Weka package.

7.1.2 Classification algorithms

As classification algorithm for this task, I used the Complement Naïve Bayes classifier, AdaBoost with Decision Stumps, and Support Vector Machines. All chosen classifiers are part of the Weka toolkit.

Among the classifiers with potential to work well on data sets with high class imbalance distribution, we intended to show which one could be a more suitable solution for this task.

7.2 Results

From the biology point of view, the genes that are knocked out and have a significant effect on the AHR signaling system are of greater importance than the other ones. These genes are considered the positive class for the given data. The organizers of the KDD Cup framed the task as a two-class problem with focus on the positive class, but with two definitions for the positive class. The first definition is called the "narrow" definition of the positive class and it is specific to the knocked-out genes that had an AHR-specific effect. In this case, the positive class is defined by the experiments in which the label of the system is "change" and the negative examples are the experiments that consist of those genes with either the "nc" or the "control" label. The second definition consists of those genes labeled with either the "change" or the "control" label. The negative class consists of those genes labeled with the "nc" label. The second partitioning corresponds to the "broad" characterization of the positive class genes that affect the hidden system.

With this division of the positive class, the results will be evaluated as a combination of results for the two "views" of the class of interest.

The area under the Receiver Operating Characteristic (ROC) curve was chosen as an evaluation measure. The global score for the task will be the sum of the area under the ROC curve (AUC) values for both the "narrow" and the "broad" partitioning of the data. This evaluation technique was used in the KDD Cup. In order to be consistent and be able to compare the results, I follow it for my methods as well. Because of the highly imbalanced data sets, this measure seems to be more appropriate than others. Also, because the AUC measure is suitable for a ranking-based classification evaluation and in this current task the definition of the positive class was not strict (a narrow and broad definition was provided) this evaluation measure is a suitable solution.

The random classifier with an AUC measure of 0.5 is considered as a baseline.

The results will be presented for each definition of the positive class, narrow and broad, separately, and in the end, the combination of results will also be presented. For all the tables presented the results in bold mark the best AUC result.

7.2.1 Results for the narrow classification

Tables 7.2 and 7.3 present the results obtained for the narrow class. The first table contains the results obtained without the feature selection technique, while the second one contains the results after applying the InfoGain feature selection.

The structure of the result tables is consistent: the first column represent the name of the representation technique (B stands for binary, Freq stands for frequency, UMLS stands for UMLS concepts, Rel stands for UMLS relations, UMLS_Phr stands for UMLS phrases); the second column presents the number of features that were used in the experiment, and the next three columns show the results for the three classification algorithms. The last row in the tables averages the results obtained for each classifier. The average results are intended just to give a general view of the results for each classification algorithm.

Represen tation	**No. of Features**	**CNB (AUC)**	**AdaBoost (AUC)**	**SVM (AUC)**
BOW_B	21,524	0.499	**0.613**	0.5
BOW_Freq	21,524	0.497	0.592	0.5
UMLS_B	415	0.492	0.571	0.494
UMLS_Freq	415	0.574	0.5	0.5
Rel_B	50	0.506	0.505	0.5
Rel_Freq	50	0.506	0.5	0.5

Table 7.2: Results for the narrow class definition without feature selection.

Represen tation	**No. of Features**	**CNB (AUC)**	**AdaBoost (AUC)**	**SVM (AUC)**
BOW_B	328	0.532	0.596	0.499
BOW_Freq	361	0.463	0.598	0.5
UMLS_B	16	0.51	0.571	0.499
UMLS_Freq	13	0.48	0.519	0.5
Rel_B	3,607	0.509	**0.615**	0.5
Rel_Freq	3,678	0.473	0.563	0.5

Table 7.3: Results for the narrow class definition with InfoGain feature selection.

In the experiments without feature selection, for both classes, the UMLS_Phr representation does not have results present in the table, the reason being that the number of the feature was huge (191,000) and the experiments were not feasible to run. For this type of representation, the results with feature selection technique are reported.

7.2.2 Results for the broad classification

Similar tables to the ones presented for the narrow class definition are presented for the broad class. Tables 7.4 and 7.5 present the broad class results.

Represen tation	**No. of Features**	**CNB (AUC)**	**AdaBoost (AUC)**	**SVM (AUC)**
BOW_B	21,524	0.498	0.598	0.498
BOW_Freq	21,524	0.497	0.557	0.494
UMLS_B	415	0.496	**0.607**	0.489
UMLS_Freq	415	0.477	0.606	0.5
Rel_B	50	0.542	0.547	0.5
Rel_Freq	50	0.542	0.5	0.5

Table 7.4: Results for the broad class definition without feature selection on the test data.

In the experiments without feature selection, the UMLS representation with the AdaBoost classifier was slightly superior to the other representations. The difference is not significant but one reason why this slight improvement could be that the further feature selection that the AdaBoost classifier performs, helps in correctly classifying some of the test instances. The classifier is able to select few of the most discriminative features to perform the classification.

For a better view of the results, Tables 7.6 and 7.7 present the summed results for the two classes with and without feature selection for the classifier that consistently obtained the best results, the AdaBoost classifier. AdaBoost obtained significantly better results over the baseline and over the other two classifiers according to the t-test. In these experiments the BOW representation with the AdaBoost classifier slightly improved the UMLS representation maybe due to the fact that the selected features were larger in number and discriminative enough.

When summing the results for both definitions of the classes the best results are obtained with lexical features without any feature selection technique and with AdaBoost as a classifier. From these results we can conclude that lexical features stand for a good representation technique for this particular task.

Represen tation	**No. of Features**	**CNB (AUC)**	**AdaBoost (AUC)**	**SVM (AUC)**
BOW_B	470	0.501	**0.598**	0.5
BOW_Freq	470	0.47	0.557	0.514
UMLS_B	14	0.395	0.584	0.517
UMLS_Freq	14	0.497	0.5	0.5
Rel_Freq	1	0.5	0.5	0.5
UMLS_Phr_B	6,373	0.438	0.538	0.499
UMLS_Phr_Freq	6,373	0.481	0.522	0.499

Table 7.5: Results for the broad class definition with InfoGain feature selection on the test data.

Representation	**AdaBoost (AUC) Narrow**	**AdaBoost (AUC) Broad**	**Sumed (AUC)**
BOW_B	0.613	0.598	**1.211**
BOW_Freq	0.592	0.557	1.149
UMLS_B	0.571	0.607	1.178
UMLS_Freq	0.5	0.606	1.106
Rel_B	0.505	0.547	1.052
Rel_Freq	0.5	0.5	1

Table 7.6: Sum of the AUC results for the two classes without feature selection on the test data.

7.3 Discussion and Conclusions

7.3.1 Discussion

Looking at the obtained results, a general conclusion can be made: textual information is useful for biology-specific tasks. Brady and Shatkay (2008) showed that text mining

Representation	**AdaBoost (AUC) Narrow**	**AdaBoost (AUC) Broad**	**Sumed (AUC)**
BOW_B	0.596	0.598	**1.194**
BOW_Freq	0.598	0.557	1.155
UMLS_B	0.571	0.584	1.155
UMLS_Freq	0.519	0.5	1.019
Rel_Freq	0.5	0.5	1
UMLS_Phr_B	0.615	0.538	1.153
UMLS_Phr_Freq	0.563	0.522	1.085

Table 7.7: Sum of the AUC results for the two classes with InfoGain feature selection on the test data.

techniques out-performed state-of-the-art results (using genome sequences) for the protein sub-cellular location task. Without any additional knowledge, my results are comparable with the ones obtained by the winners of the KDD competition.

The classification algorithm that was consistent in results and outperformed the other two is AdaBoost. It is known that this classifier works well for text classification. Due to the high imbalance in the class distribution, a classifier that manages to extract as much information as possible from the minority class has more chances to perform well. The fact that the classifier focuses on examples that are "hard" to learn can be the reason why AdaBoost outperformed the others.

Binary versus frequency representation is a decision that always needs to be made when working with text. For the experiments presented in this work, the binary representation outperformed the frequency one for almost all types of representations. One reason could be due to the fact that the binary representation tends to be more stable. There are classifiers that are designed to work well with frequency representation for imbalance data sets, CNB is one of them. For some experiments, CNB with frequency representation outperformed the CNB with binary representation in the experiments, but

overall binary outperformed the frequency representation for this task.

Feature selection is a technique that is mostly applied when the number of features is large. In these experiments, the feature selection technique did not obtain a significant improvement, but it was useful in reducing runtime when the number of features was huge.

The bag-of-words representation outperformed the other ones that were supposed to be more "knowledge"-charged. I believe that the reason behind these results is the fact that the UMLS features are too coarse and are not able to capture differences between classes. Based on previous experiments, the AdaBoost classifier tends to perform well when instances are represented by features that are discriminative for the classes. Technical textual data is more English-standardized than clinical data and the lexical features are able to capture the small differences between instances. These differences show to be very beneficial to some classification algorithms like AdaBoost. If we are looking at the results in terms of recall for the positive class, another measure that might be of interest besides the AUC value, the best performance by far was obtained by the UMLS_Freq representation when using CNB as a classification algorithm. CNB classifier tends to work well when the features are well represented in the data and not very sparse. Because an UMLS representation fits this description of the data, CNB also performed well on this task.

7.3.2 Conclusions and future work

As an overall conclusion, the text of the abstracts has shown its usefulness once again. Taking into consideration the characteristics of the task and of the data in making experimental decisions is important. The sum of 1.211 in the AUC measure that I obtained outperformed the 1.164 AUC value obtained by Krogel et al. (2002) with an information extraction system on the same data set that we used. The results presented in this work are comparable (it is not a direct comparison because I used the subset of the original data that had abstracts) with a sum of AUC measure of 1.23 previously

obtained by the winners of the competition after various levels of tuning the parameters. They obtained 0.62 AUC measure for the narrow class and 0.61 AUC measure for the broad class. The best results are 0.613 for the narrow class and 0.607 for the broad class.

The experiments and results obtained in this work conclude that text is a useful source of information. Not only that it can improve the results in biology-specific tasks, but can be considered a stand-alone source of knowledge in this domain.

In future work, it would be useful to investigate the addition of more biology-specific knowledge as features for the learning algorithm. New information can be used from the location, function, and protein-protein interactions annotations that come with the original data.

A combination of classifiers or a co-training setting might also help improving the results of the task, especially when the data is divided into various types of information.

Other representations that make use of more external resources, e.g., the Gene Ontology, can be useful as well for new directions of research on this task and other similar tasks. Ontologies can be used to find textual information about the genes that are not directly mentioned in published abstracts. Homologous genes can be identified from existing ontologies and information about these related genes can be used to identify missing textual data.

Chapter 8

SubGoal4: Identify Relations Between Medical Entities

People care deeply about their health and want to be, now more than ever, in charge of their health and healthcare. Life is more hectic than has ever been, the medicine that is practiced today is one in which medical expertise is not only based on years of practice but on the latest discoveries as well. Tools that can help us manage and better keep track of our health such as Microsoft HealthVault[1] are reasons and facts that make people more powerful when it comes to healthcare knowledge and management. Tools like this make the vision of Dr. William Buchan, the physician, who in 1785 published the first home health care book, Buchan (1784), a reality: "The cure of diseases is doubtless a matter of great importance; but the preservation of health is of still greater. This is the concern of every man, and surely what relates to it ought to be rendered as plain and obvious to all as possible".

The traditional healthcare system is also becoming one that embraces the Internet and the electronic world. Electronic health records (EHR) are becoming the standard in the healthcare domain. Researches and studies show that the potential benefits of

[1] http://healthvault.com/

having an EHR system are[2]:

Health information recording and clinical data repositories — immediate access to patient diagnoses, allergies, and lab test results that enable better and time-efficient medical decisions;

Medication management — rapid access to information regarding potential adverse drug reactions, immunizations, supplies, etc.;

Decision support — the ability to capture and use quality medical data for decisions in the workflow of healthcare; and

Obtain treatments that are tailored to specific health needs — rapid access to information that is focused on certain topics.

Due to the accumulating masses of text surrounding every patient case, in order to embrace the views of the EHR systems, we need better, faster, and more reliable access to information. In the medical domain, the richest and most used source of information is PubMed, a database of extensive life science published articles. All research discoveries come and enter the repository at high rate (Hunter and Cohen, 2006), making the process of identifying and disseminating reliable information a very difficult task. Moreover the presence of EHR and medical records in which patients are able to input their own observations are signs that the medical practice that is required nowadays is one tailored to precise needs of each individual.

This chapter is focused on identifying meaningful relations between medical concepts in technical and clinical data. As it is mentioned in Chapter 1, building systems that are able to identify these types of relations represent one of the puzzle pieces for building a personalized medicine. This information can be used in predicting treatment outcomes for patients with specific characteristics or in identifying patient characteristics for which a certain therapy represents an unsuccessful treatment process. From a linguistic point of view, this type of information can help in the process of building domain-specific ontologies.

[2]http://healthcaretracker.wordpress.com/

The tasks that are addressed in this work represent the foundation of an information technology framework that identifies and disseminates healthcare information. People want fast access to reliable information and in a manner that is suitable to their habits and workflow. Medical knowledge (e.g., published articles, clinical trials, news, etc.) represents a source of empowerment for both healthcare providers and laypeople. Studies reveal that people are searching the web and read medical related information in order to be informed about their health. Ginsberg et al. (2009) show how a new outbreak of the influenza virus can be detected from search engine query data.

Healthcare providers (e.g., private clinics, hospitals, medical doctors, etc.), companies that build systematic reviews, or laypeople who want to be in charge of their health by reading up-to-date medical information would find a use in such a tool. I envision a possible final product as a browser plug-in or a desktop application that will automatically find and extract the latest medical discoveries related to disease-treatment relations to present them to the user.

The objective of this work is to show which natural language processing and machine learning techniques, what type of representation and what classification algorithms are suitable to use for identifying and classifying relevant medical information in short texts. Also as part of my objectives, I propose and investigate the effectiveness of using a hierarchical, pipeline approach of performing this work's goals.

In this work, I look at identifying and classifying pairs of medical entities into several semantic relations. This research problem is using both technical and clinical data. The proposed methodology is hierarchical: in an initial step instances that contain related pairs of concepts are identified, while in a second step these related concept-pairs are further classified into one of the possible medical relations.

For comparison purposes, and not only, I also look at identifying the existing relations between the pairs of concepts by means of a flat-classification approach. The above mentioned tasks are solved in one step by adding a non-relevant class to the other relations of interest.

In this thesis, I focus on the identification and classification of relations between disease, treatments, and tests in Medline abstracts and in clinical data. I envision the potential and value of these findings as representing a big part of a framework capable to find relevant information about diseases and treatments in both technical and clinical medical repositories. The results that I obtained show that NLP and ML techniques represent a realistic scenario for building such a tool. Therefore, this current study and the future work are aimed at researching various representation techniques in combination with different learning methods for identifying and extracting medical relations from technical and clinical data.

The contributions brought with this work stand in the fact that I present an extensive study of various ML algorithms and textual representations for classifying short medical texts and identifying semantic relations between three medical entities: diseases, treatments, and tests. From a ML point of view, I show that in short texts, when identifying semantic relations between pairs of concepts, an improvement in results is obtained when using a hierarchical way of approaching the task (a pipeline of two tasks). It is in general better to identify and eliminate first the textual context that do not contain relevant information, and then classify the rest of the sentences by the relations of interest, instead of doing everything in one step by classifying sentences into one of the relations of interest plus the extra class of uninformative sentences.

Compared to previous research, I do not use specifically the information that identifies which token represents a medical entity. I consider the textual context as the only source of information.

This work has been published in (Frunza and Inkpen, 2010a), (Frunza and Inkpen, 2010b), (Frunza et al., 2011b), and (Frunza and Inkpen, 2011).

8.1 Task description and data sets

The goal of this chapter is to identify a successful methodology that can identify related pairs of medical concepts and further classify this pairs in meaningful relations. For a broader and better assessment of such tools, I propose and test my methodology on both technical and clinical data. I follow the same general approach, while in the same time I take into account the specifics of the tasks in each of the types of data I work with.

The technical data that I use consists of Medline abstracts and the mention of medical concepts is scarcer than in clinical data such as *clinical notes, discharge summaries, etc.* Both in technical and clinical data, medical concepts are annotated at sentence level, but, while in the technical data a sentence contains only one pair of concepts, in the clinical data a sentence can contain more than one pair.

The hypothesis that guides my current work for solving the task of identifying and classifying relations between pairs of concepts follows a divide-and-conquer approach. I base the premise of my methodology on the fact that, if in an initial step in the original supplied data set only related pairs are identified and then in a second step these already identified pairs will be finer-grained classified, the performance of the overall task is improved. This hypothesis also follows a natural way of having automatic tools assist humans in their tasks: an initial step will translate into a scan of the whole original data with the aim to reduce its size and retain only meaningful information, while the next step is the one that will use and perform further tasks on this newly created data set.

Figure 8.1 presents a graphical view of the methodology that I propose for solving the task of identifying and classifying relations between medical concepts.

Since the task is deployed on two types of textual data: technical and clinical the following sections will be divided according to the type of data I use.

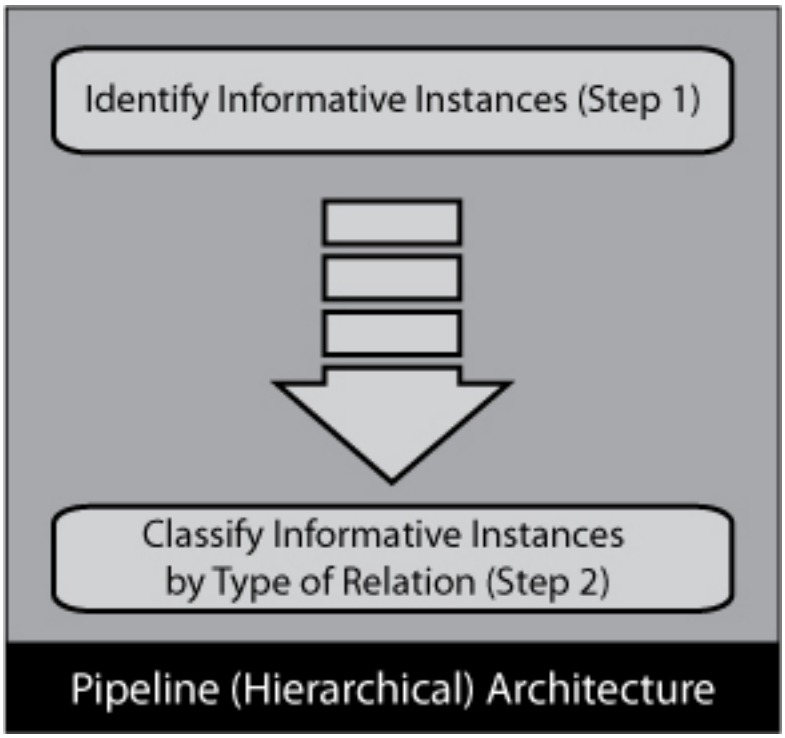

Figure 8.1: The hierarchy (pipeline) architecture.

8.1.1 Task description and data sets - technical data

The data set that I use for the experiments is presented in Table 3.9, Chapter 3, section 3.2, subsection 3.2.3. Each sentence from the annotated Medline abstracts contains a single pair of concepts and implicitly a single relation.

The first task (task 1 or sentence selection) identifies sentences from Medline published abstracts that talk about diseases and treatments. The task is similar to a scan of sentences contained in the abstract for the scope of identifying sentences that contain relevant information, related disease-treatment entities.

The second task (task 2 or relation identification) has a deeper semantic dimension and it is focused on identifying disease-treatment relations in the sentences already selected as being informative (e.g., task1 is applied first). I focus on 3 relations: *Cure*, *Prevent*, and *Side Effect*, a subset of the 8 relations that the corpus is annotated with. I decided to focus on these 3 relations because these are most represented and

more meaningful in the corpus compared with the others.

Unlike in the work of Rosario and Hearst (2004), in my current research, the annotations of the data set are used to create a different task (task 1). It identifies informative sentences that contain information about diseases and treatments and semantic relations between them, versus non-informative sentences. This allows us to see how well NLP and ML techniques can cope with the task of identifying informative sentences, or in other words, how well they can weed out sentences that are not relevant to diseases and treatments.

Extracting informative sentences is a task by itself in the NLP and ML community. Research fields like summarization and information extraction are disciplines where the identification of informative text represents a crucial task. The contributions of this research task stand in the usefulness of the results and the experimental settings insights of applying it to the medical domain.

For the first task, the data sets are annotated with the following information: a label indicating that the sentence is informative, i.e., containing disease treatment information (all sentences except the irrelevant ones in the original corpus) or a label indicating that the sentence is not informative. Table 8.1 gives an example of labeled sentences.

Label	**Sentence**
Informative sentence	*Urgent colonoscopy for the diagnosis and treatment of severe diverticular hemorrhage.*
Non-informative sentence	*In all cases a coproparasitological study was performed.*

Table 8.1: Examples of annotated sentences for the sentence selection task.

For the second task, the sentences have annotation information that states if the relation that exists in a sentence between the disease and treatment is *Cure*, *Prevent*, or *Side Effect*. These are the relations that are more represented in the original data set and also needed for future research. I would like to focus on a few relations of interest and try to identify what predictive model and representation technique bring the best

results. The task of identifying the three semantic relations is addressed in three ways:

Setting 1: build three models, each focused on one relation that can distinguish sentences that contain the relation — *Positive* label, from other sentences that are neutral — *Neutral* label, and from sentences that do not contain relevant information — *Negative* label;

Setting 2: build three models, each focused on one relation that can distinguish sentences that contain the relation from sentences that do not contain any relevant information. This setting is similar to a two-class classification task in which instances are labeled either with the relation in question — *Positive* label, or with non-relevant information — *Negative* label;

Setting 3: build one model that distinguishes the three relations - a three-way classification task where each sentence is labeled with one of the semantic relations, using the data with all the *Positive* labels.

Tables 8.2 and 8.3 present the data sets that I used for the two tasks.

	Informative sentences	**Non-informative sentences**
Training set	1,225	1,176
Test set	612	591

Table 8.2: Data sets used for the first task on technical data.

In Table 8.3 the label "Positive" represents a sentence that contains the semantic relations (i.e., *Cure*, *Prevent*, or *Side Effect*), and "Negative" a sentence that does not contain information about any of the semantic relations but contains information about either a disease or a treatment labels *Treatment_Only*, *Disease_Only* in previous research by Rosario and Hearst (2004). In this previous work, the authors use five generative graphical models and a discriminative model (a multi- layer neural network) to perform the task of entity recognition. The neural network model is used for the relation discrimination. Their representation techniques are based on words in context, part-of-speech information, phrases, and a medical lexical ontology - Mesh terms.

		Train	
Relation	**Positive**	**Negative**	**Neutral**
Cure	554	531	25
Prevent	42	531	25
SideEffect	20	531	25
		Test	
Relation	**Positive**	**Negative**	**Neutral**
Cure	276	266	12
Prevent	21	266	12
SideEffect	10	266	12

Table 8.3: Data sets used for the second task on technical data.

Compared to this work, my research is focused on different representation techniques, different classification models, and most importantly, it generates improved results with less annotated data.

Up to this point, I presented the two tasks separately as being two self-defined tasks, since they can be used for other more complex tasks as well. From a methodological point of view, and, more importantly, from a practical point of view, they can be integrated together in a pipeline of tasks as a solution to a framework that is tailored to identify semantic relations in short texts and sentences, when it is not known a priori if the text contains useful information. The proposed pipeline solves task 1 first and then processes the results in task 2, so that, in the end, only informative sentences are classified into the three semantic relations. The logic behind choosing to experiment with and report results for the pipeline of tasks is that I have to identify the best model that will get us closer to the main goal: being able to identify and reliably classify medical information. Using the pipeline of tasks, I eliminate some errors that can be introduced due to the fact that I would consider uninformative sentences as potential data when classifying sentences directly into semantic relations. I will show that the pipeline achieves better

results than a more straightforward approach of classifying in one step into one of the three relations of interest plus an extra class for uninformative sentences.

The pipeline of tasks can be a solution for identifying and disseminating relevant information tailored to a specific semantic relation, because the second task is doing a finer-grained classification of the sentences that already contain information about the relations of interest. This framework is appropriate for consumers that are more interested in an end-result that is more specific, e.g., relevant information only for the class *Cure*, rather than identifying sentences that have the potential to be informative for a wider variety of disease-treatment semantic relations.

8.1.2 Task description and data sets - clinical data

As described in Chapter 3, section 3.1, subsection 3.1.2 the clinical data consists of discharge summaries where disease, treatment, and test concepts are annotated, and 8 sentence-level relations between these concepts are identified. A sentence can contain more than two concepts and implicitly can have associated more than a single relation.

The first task (task 1 or concept-pair selection) is focused on identifying which pairs of concepts that co-existing in a sentence are related. For each sentence with already identified concepts, this task is similar to one of weeding out all pairs of concepts that are not related.

The second task (task 2 or relation identification) is the task that is applied after the first task is deployed on the whole data set. The pairs of concepts identified as related by task 1, are now being finer-grained classified in one of the 8 relations of interest.

Similar to the methodology used on the technical data, the first goal of the hierarchical approach is to weed out the non-informative instances and create a new data set for the second step in which these pairs of concepts are classified. This type of methodological approach fits well with a real-life scenario where, from all annotated medical concepts in a medical record, the most interesting pairs of concepts are the ones identified as being

related.

For the first task, the original annotated data set becomes one that is labeled with two classes: *Related* versus *Non-Related*. The annotations for the pairs of concepts are made at a sentence level, similar to the annotations performed on the technical data. If in the technical data only one pair of concepts was present is a sentence, in the clinical data a sentence can contain more than one pair of concepts, each with a possibly different class. The way the clinical data is annotated makes the task more difficult than the one on technical data, triggering different approaches compared to the ones used on the technical data. Figure 8.2 presents an example with concept and relation annotations.

Table 8.4 presents some statistics of the distribution of annotated concepts in the training and test data sets. We can observe that many sentences have more than two concepts annotated, rendering the task of relation classification a more difficult one to deal with.

Training data	No. of instances
Sentences with 1 concept	7,990
Sentences with exactly 2 concepts	3,092
Sentences with more than 2 concepts	3,289
Test data	No. of instances
Sentences with 1 concept	12,170
Sentences with exactly 2 concepts	5,306
Sentences with more than 2 concepts	5,131

Table 8.4: Statistics on concept annotation on clinical data.

If on the technical data it was obvious to use the entire sentence as a context for classification, on the clinical data, I decided not to use the whole sentence as a context, but a sub context that consists of the pair of concepts and the word tokens that exists between them. Early experiments were performed on whole sentences but with very poor results.

```
Concept annotations

--------------------------------------------------------------------------
c="an echocardiogram" 10:3 10:4||t="test"
c="hypertension" 8:2 8:2||t="problem"
c="hypertension" 4:7 4:7||t="problem"
c="diabetes" 4:12 4:12||t="problem"
c="a cardiac catheterization" 9:12 9:14||t="test"
...
--------------------------------------------------------------------------

Relation annotations


--------------------------------------------------------------------------
c="a cardiac catheterization" 9:12 9:14||r="TeCP"||c="chest pain" 9:5 9:6
c="a cardiac catheterization" 9:12 9:14||r="TeRP"||c="an occluded right
coronary artery" 9:23 9:27
c="a cardiac catheterization" 9:12 9:14||r="TeRP"||c="a 40-50% proximal
stenosis" 9:29 9:32
--------------------------------------------------------------------------
```

Figure 8.2: Example of concept and relation annotation in clinical data.

The original annotated data has for each medical record, all medical concepts identified and information about their relation provided, making all other sentence identified pairs of concepts the non-related instances. Using this data, the new data-set for the first task becomes one in which all sentence-level related concepts represent the *Related* class and the rest the *Non-related* class.

Table 8.5 and 8.6 present the clinical data sets that I used for the first and second task.

	Related instances	**Non-related instances**
Training set	5,264	11,154
Test set	9,070	17,046

Table 8.5: Data sets used for the first task on clinical data.

Relation	Training Sentences	Test Sentences
PIP	1,239	1,989
TeCP	303	588
TeRP	1,734	3,033
TrAP	1,423	2,487
TrCP	296	444
TrIP	107	198
TrNAP	106	191
TrWP	56	143

Table 8.6: Data sets used for the first task on clinical data.

8.2 Experimental Setting

8.2.1 Data representation

In ML, as a field of empirical studies, the acquired expertise and knowledge from previous research guide the way of solving new tasks. The models should be reliable at identifying related instances and discriminating between semantic relations. The research experiments need to be guided such that high performance is obtained. The experimental settings are directed such that they are adapted to the domain of study (medical knowledge) and to the type of data I deal with (short texts or sentences), allowing for the methods to bring improved performance.

In solving the proposed tasks, I used the bellow-mentioned types of features. I did not use any feature selection technique, mostly because the data set that I have is not

large and because the majority of the classifiers that I use are known to work well with a high number of features.

Besides the frequency feature representation, I also tried the tf-idf representation in some experiments, but it did not bring any further improvements.

BOW — For the BOW representation, I used a frequency-value representation because I am dealing with short texts, an average of 20 words per sentence, and the difference between a binary value representation and a frequency-value representation is not large. This representation has the advantage that if a feature appears more than once in a sentence, this means that it is important and the frequency-value representation will capture this — the feature's value will be greater than that of other features.

The features that I selected for the BOW representation follow the same rules, as I mentioned in Chapter 4.

Syntactic features and biomedical entities — The second type of representation is based on syntactic information: noun phrases, verb phrases, and biomedical concepts identified in the sentences, as described in Chapter 4 subsection 4.1.4.

Experiments were performed when using as features only the final set of identified noun phrases, only verb phrases, only biomedical entities, and with combinations of all these features. When combining the features, the feature vector for each instance is a concatenation of all features.

Medical concepts (UMLS) representation — The UMLS features used as a representation technique are identified in the same way as they were described in Chapter 4, subsection 8.2.2. The identified concepts by the MetaMap system represent the selected features.

ConceptType — These features represent semantic information about the type of medical concept of each entity. This information is represented for each concept by

one of three possible nominal values. These values correspond to each of the possible types of medical concepts: problem, treatment, and test. This feature is used only when representing clinical data, since in the technical data I had only relations between diseases and treatments.

ConText — This is a type of feature that represents information extracted with the ConText tool described in Chapter 4. I used six nominal features in order to represent the information provided by the tool: three for each concept of the pair. The first feature captures information regarding the possible negation of the term in the given context and it has two possible values: *affirmed* or *negated*; the second feature captures information regarding the temporality of each concept in the concept pair: *recent* or *historical*; the third feature is focused on the experiencer information and it has two possible values: *patient* or *other*.

Semantic vectors — Semantic vector models are models in which concepts are represented by vectors in some high-dimensional space. Similarity between concepts is computed using the analogy of similarity or distance between points in this vector space. The main idea behind semantic vector models is that words and concepts are represented by points in a mathematical space and this representation is learned from text in such a way that concepts with similar or related meanings are near to one another in that space.

In order to create these semantic vectors I used the Semantic Vectors Package[3] of Widdows and Ferraro (2008). The package uses indexes created by applying a Random Projection algorithm to term-document matrices created using Apache Lucene[4].

I used the semantic vectors to extract the top 300 terms correlated with each type of relation and to determine the cosine distance between a pair of concepts and the relations that I use. These semantic vector features are used on the clinical data.

[3]http://code.google.com/p/semanticvectors/

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

8.2.2 Classification algorithms

As classification algorithms, I use a set of 6 representative models: decision-based models (Decision trees), probabilistic models (Naïve Bayes (NB) and Complement Naïve Bayes (CNB), which is adapted for text with imbalanced class distribution), adaptive boosting (AdaBoost), a linear classifier — support vector machine (SVM) with polynomial kernel (other kernels were tried with no improved results), and a classifier that always predicts the majority class in the training data (used as a baseline). The justifications for using these algorithms are the ones described in Chapter 4.

8.3 Results

As evaluation measures, because I use imbalanced data sets, I chose to report, in addition to accuracy, the macro-averaged F-measure. I decided to report macro and not micro-averaged F-measure because the macro measure is not influenced by the majority class, as the micro measure is. The macro measure tends to focus on the performance the classifier has on the minority classes; these classes create most of the problems for the ML classifiers.

8.3.1 Results on the technical data

Results for the task of identifying informative sentences (task 1)

This subsection presents the results for the first task, the one of identifying whether sentences are informative, e.g., they contain information about diseases and treatments or not. The ML settings are created for a binary classification task and the representations are the ones mentioned in the previous subsection, while the baseline on which I need to improve is given by the results of a classifier that always predicts the majority class.

Figure 8.3 presents the results obtained when using as representation features verb phrases identified by the Genia tagger. When using this representation the results are

close to baseline. The reason why this happens for all algorithms that I use is the fact that the texts are short and the selected features are not well represented in an instance. I have a data sparseness problem: it is the case when a lot of features have value 0 for a particular instance.

Figure 8.4 presents the results obtained using as representation features noun phrases selected by the Genia tagger. Compared to previous results, I can observe a slight improvement in both accuracy and F-measure. The best results are obtained by the CNB classifier. I believe that the slight improvement is due to a reduction of the sparseness problem: noun phrases are more frequently present in short texts than verb phrases.

Figure 8.5 presents the best results obtained so far. An increase of almost 5 percentage points, for both accuracy and F-measure is obtained when using as representation features biomedical entities extracted by the Genia tagger and the CNB as classifier. An increase in results for the other classifiers can be also observed.

This increase can be caused by the fact that, when present in sentences, the biomedical entities have a stronger predicting value. The entities identify better if a sentence is informative or not and this is something that I would expect to happen. If a sentence contains a good proportion of biomedical entities, this should trigger a higher chance of labeling a sentence as informative.

Figure 8.6 presents accuracy and F-measure results for all classifiers when noun phrases, identified by the Genia tagger, and biomedical entities are used as representation features. Compared to previous representation techniques, the results presented in Figure 8.6 follow what will become a trend, an increase in results when more informative and diverse representation techniques are used. With a representation that combines noun phrases and biomedical entities and CNB as classifier, an increase of 2 percentage points is obtained compared to the results when only biomedical concepts are used (Figure 8.5). An increase of 8 percentage points is obtained compared to only a noun-phrase representation (Figure 8.4).

In Figure 8.7, I use as representation technique UMLS concepts - medical domain-

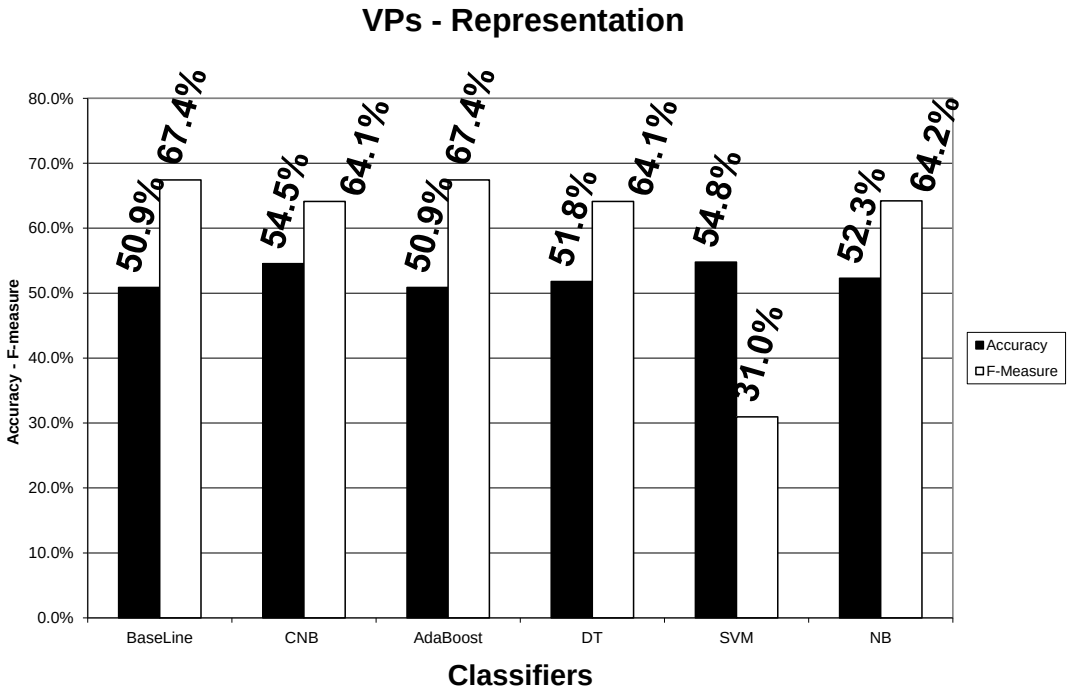

Figure 8.3: Accuracy and F-measure results when using verb phrases as features for task 1.

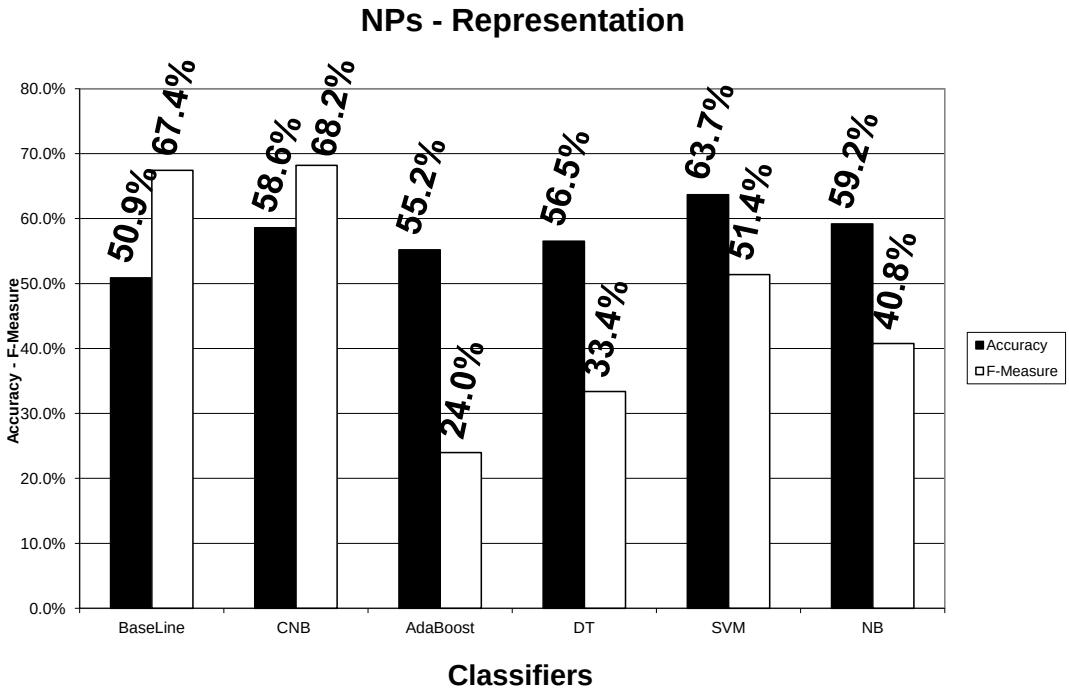

Figure 8.4: Accuracy and F-measure results when using noun phrases as features for task 1.

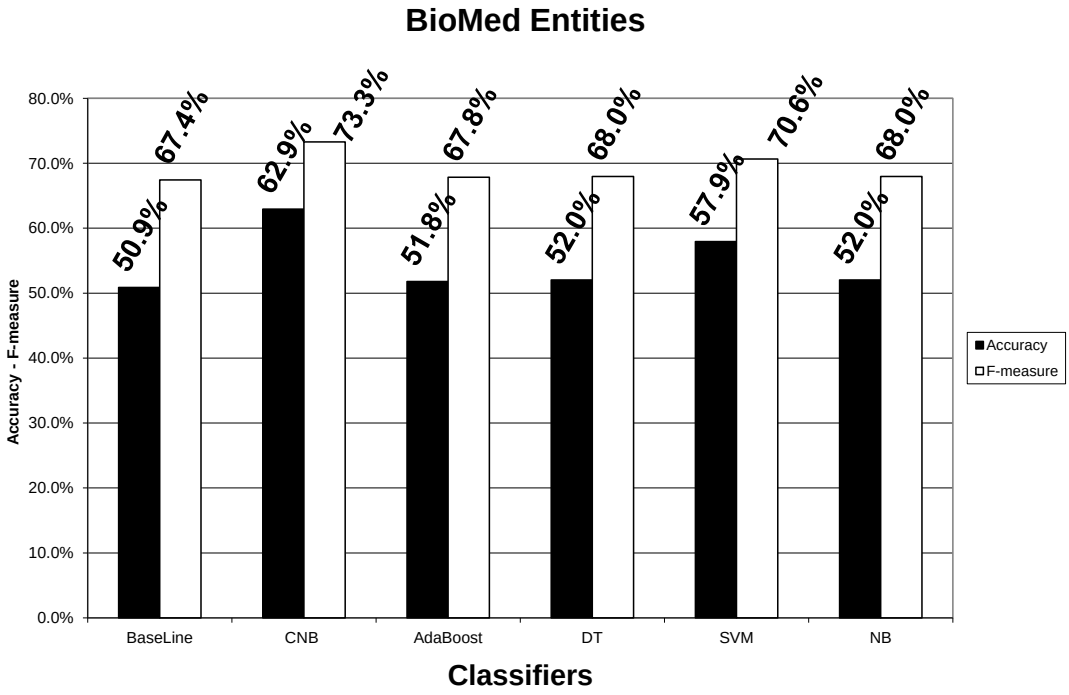

Figure 8.5: Accuracy and F-measure results when using biomedical concepts as features for task 1.

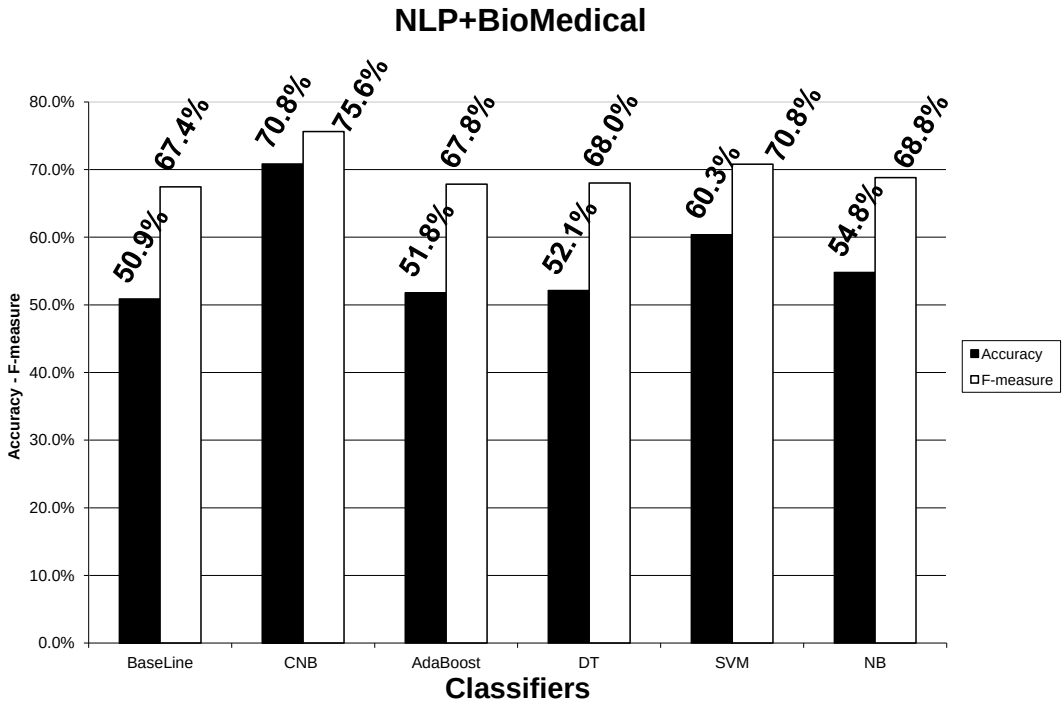

Figure 8.6: Accuracy and F-measure results when using NLP and biomedical concepts as features for task 1.

specific concepts identified by the MetaMap tool. Compared to all previous results, this representation technique for the CNB classifier obtains the best results so far, with an increase of almost 4 percentage points. I believe that this is because these concepts are more general than the biomedical concepts and the feature space is not that sparse. Classifiers tend to obtain better results when the feature space is well represented for each instance. This observation is what I believe happened with the results presented in Figure 4.1.

A representation based on noun phrases, verb phrases, biomedical concepts, and UMLS concepts brings again an increase in results compared to the ones presented in Figure 4.1. This increase of 2 percentage points can be observed in Figure 8.8. As stated before, the trend of increase in results is due to the representation that captures more information.

The bag-of-words representation technique is known in the literature to be one that is hard to beat. Even though is not a very sophisticated method — it contains only the words in context — it is one that often obtains good results. In these experiments, the BOW representation, Figure 8.9, obtains the best results among all the representation techniques mentioned in this section so far. An increase of almost 2 percentage points is obtained compared to the results in Figure 8.8.

Figure 8.10 presents the results for a BOW plus UMLS concepts representation. Even though the BOW representation is one that gives good results, when used in combinations with other types of representations, the performance improves. This observation can be drawn from Figure 8.10, where a 2 percentage points improvement is obtained for CNB, compared to the BOW representation; an improvement of 10 percentage points is obtained when comparing to the UMLS representation. For all the other classifiers, an increase in results can be observed as well.

The results obtained by only a BOW representation can be further improved when these features are combined with the noun phrases, verb phrases, and biomedical concepts. Figure 8.11 presents results for this representation technique with all the

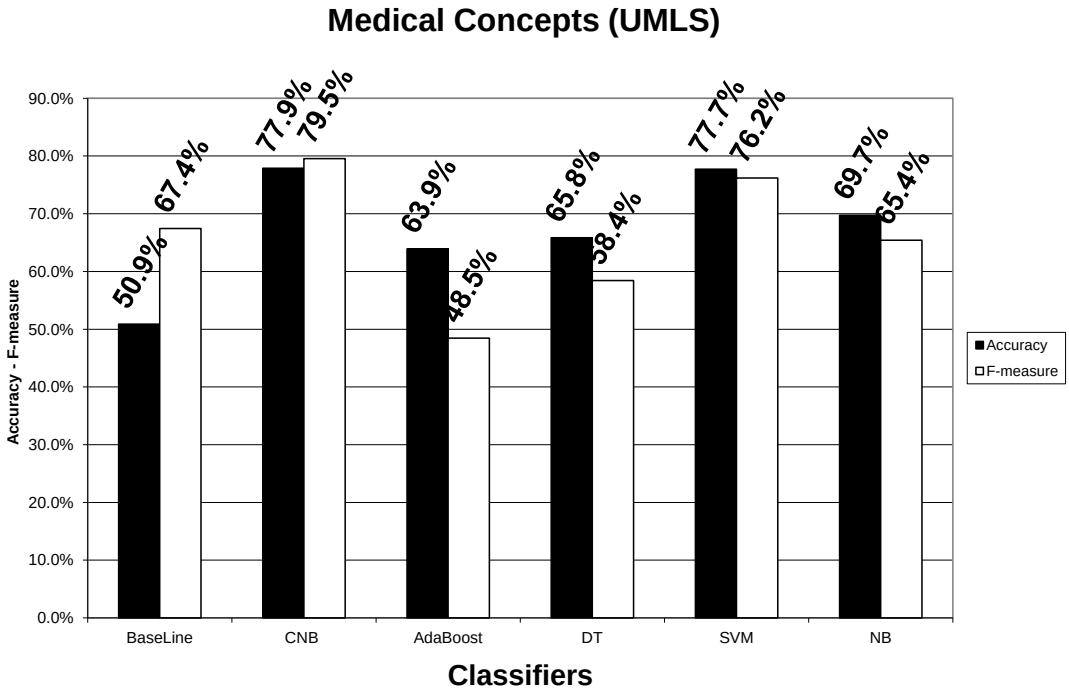

Figure 8.7: Accuracy and F-measure results when using UMLS concepts as features for task 1.

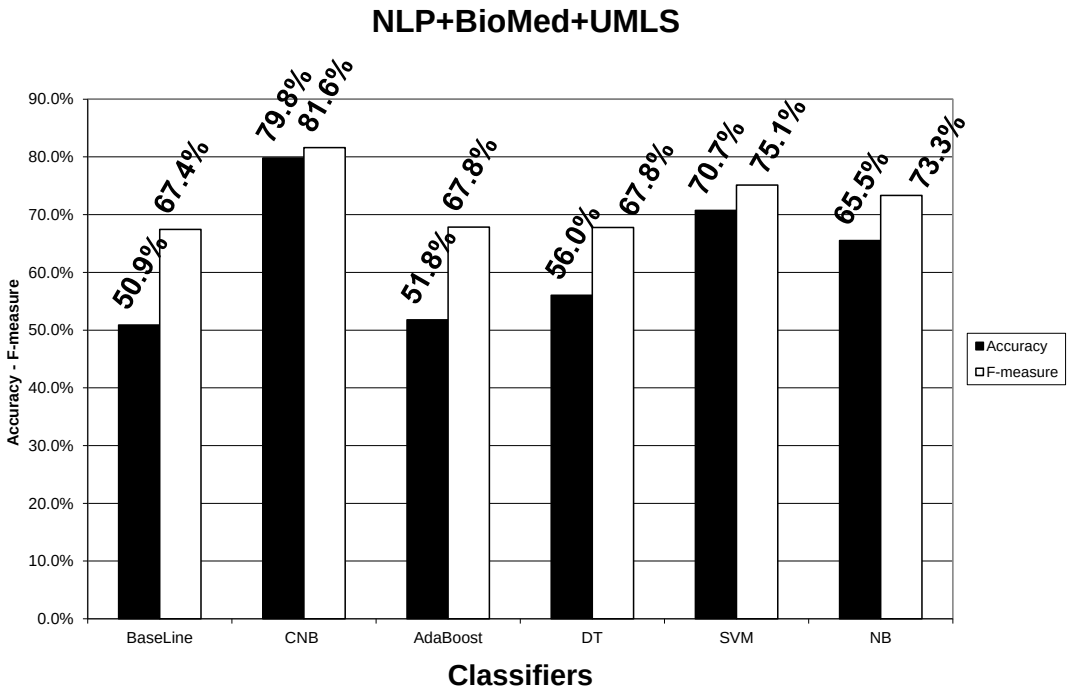

Figure 8.8: Accuracy and F-measure results when using NLP, biomedical and UMLS concepts as features, task 1.

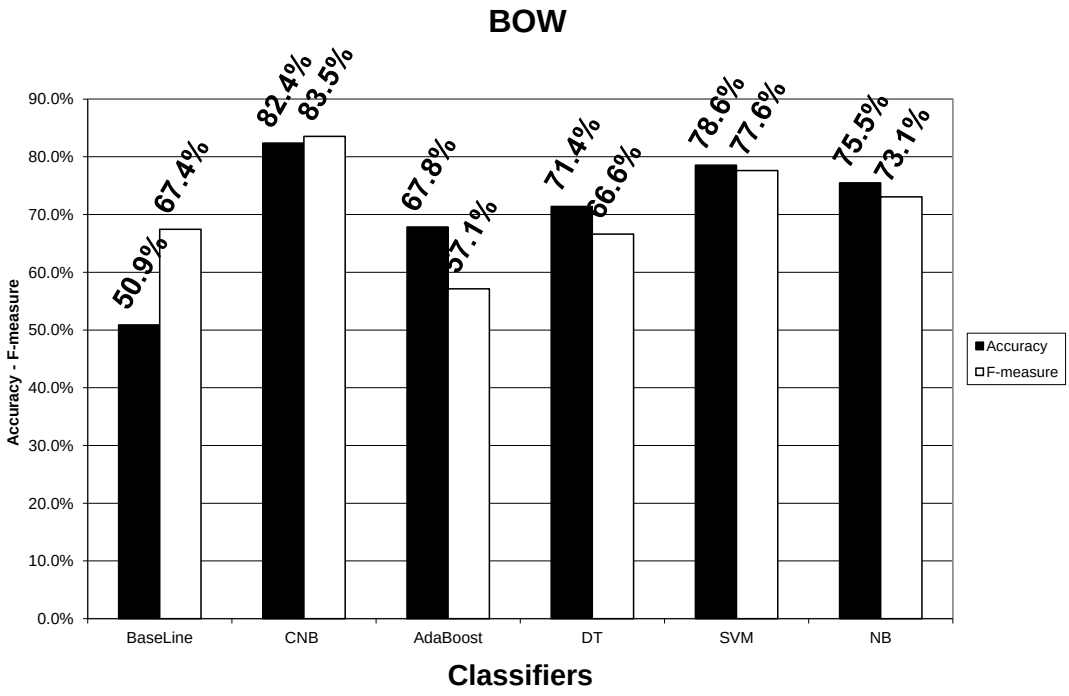

Figure 8.9: Accuracy and F-measure results when using BOW features for task 1.

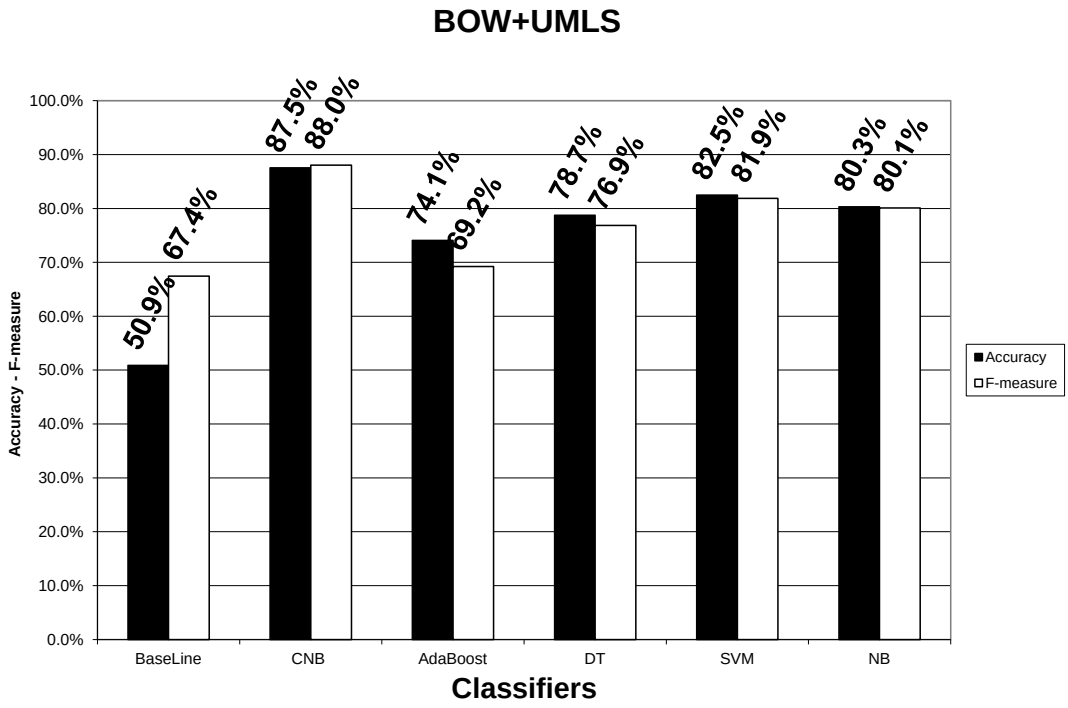

Figure 8.10: Accuracy and F-measure results when using BOW and UMLS concepts as features for task 1.

classification algorithms. The best result is improved with 1 percentage point compared to the one in Figure 8.10.

For this current task, identifying which sentences from the abstracts of Medline articles that contain informative sentences for diseases and treatments, the best results obtained are the one presented in Figure 8.12. The representation technique that uses BOW features, UMLS concepts, noun and verb phrases, and biomedical concepts with the CNB classifier obtain a 90.72% F-measure and 90.36% accuracy. These increases in results are due to the fact that all the various types of features create a rich and predictive feature space for the classifiers.

Even if some of the features alone are not the ones that obtain good performance, e.g., verb phrases, when combined with other types of features form a representation model capable to predict with 90% accuracy if a sentence is informative or not. The fact that the best results for this task are obtained when all features are put together shows that a good representation technique is a crucial aspect of a classification task. In order to statistically support the conclusions, I run t-tests measures for the CNB classifier with all the representation techniques for both accuracy and F-measure. Even though some of the differences are small, they are significantly different, except the one between Figure 8.9 and Figure 8.8. Based on the experiments, I can conclude and suggest as future guidelines for similar tasks that the richer and more informative the representation technique is, the better the performance results. For a good performance level, I suggest a combination of all the features.

Results for the task of identifying semantic relations (task 2)

The focus for the second task is to automatically identify which sentences contain information for the three semantic relations: *Cure*, *Prevent*, or *Side Effect*. The reported results are based on similar settings to the ones used for the previous task.

Since imbalanced data sets are used for this task, the evaluation measure that I am going to report is the F-measure. I consider as baseline a classifier that always predicts

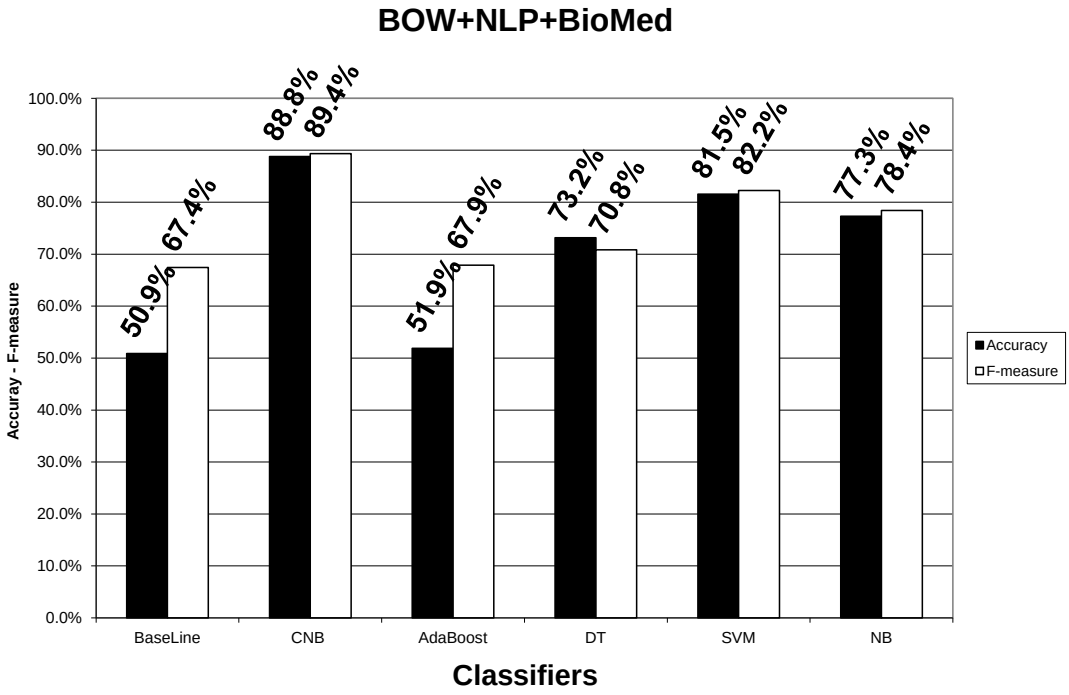

Figure 8.11: Accuracy and F-measure results when using BOW, NLP and biomedical features for task 1.

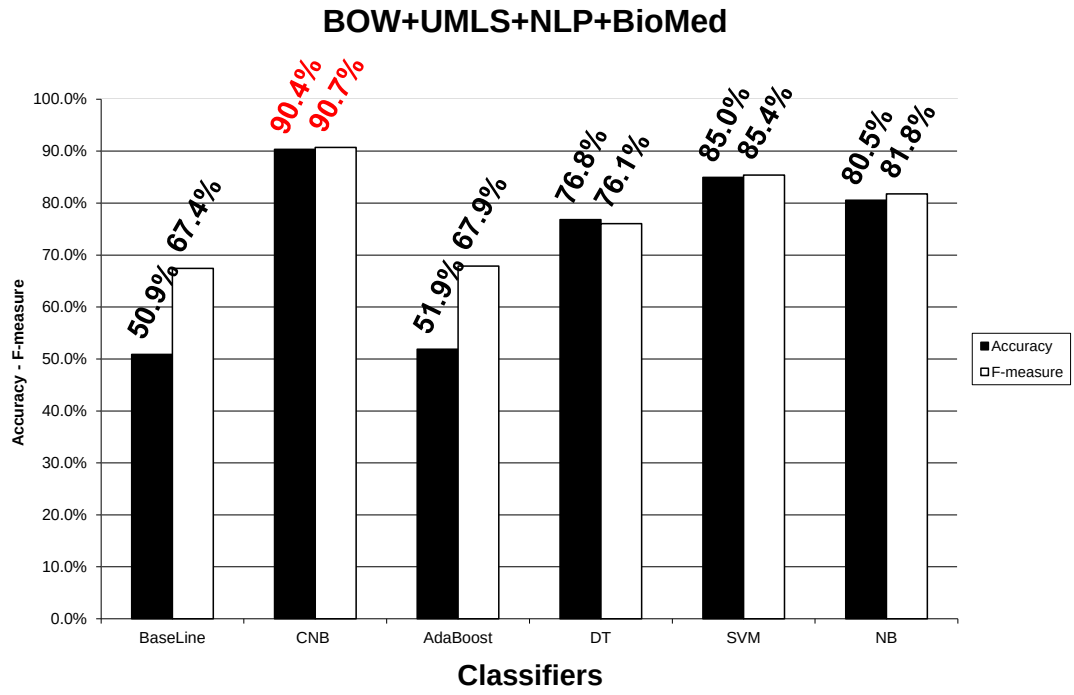

Figure 8.12: Accuracy and F-measure results when using BOW, NLP, biomedical and UMLS concepts features for task 1.

the majority class. For the relation Cure the F-measure baseline is 66.51%, for Prevent and Side Effect it is 0%.

I am going to present the best results obtained for all the three settings. The best results are chosen from all the representation techniques and all classification algorithms that I also used for the first task. The labels on the x-axis stand for the name of the semantic relation, the representation technique, and the classification algorithm used.

In Figure 8.13, I present the results when using Setting 1, described in section 8.1, as the set-up for the experiment. On the x-axis, I present for each relation the best F-measure result, the representation technique, and the classifier that obtained the result. For example, for the Cure relation, the combination of BOW features, noun phrases and verb phrases, biomedical and UMLS concepts, with SVM as a classifier, obtained the 87.10% result for F-measure. SVM and NB with rich feature representations are the set-ups that obtained the best results.

In Setting 2, the results are better than in the previous setting, showing that the neutral examples used in the previous experiments confused the algorithms and were not appropriate. These results, Figure 8.14, validate the fact that the previous setting was not the best one for the task.

Figure 8.15 presents the best results that I obtained for the second task, a level of almost 100% F-measure for the Cure relation, 100% F-measure for Prevent relation, and 75% F-measure for Side Effect. For this setting, I trained a model for all three relations in the same time, and I distinguished sentences between these three relations.

For Setting 3, the NB classifier with combinations of various representation features is the one that obtained the best results for all the relations. The improvement over the other settings can be due to the fact that the combination of classifier and features has a good predicting value for a model trained on the three relations. Each of the relations can be well-defined and predicted when using the model that I propose in Setting 3. The fact that I achieve close to perfection prediction suggests that the choice of classifier and representation technique are key factors for a supervised classification task, even for

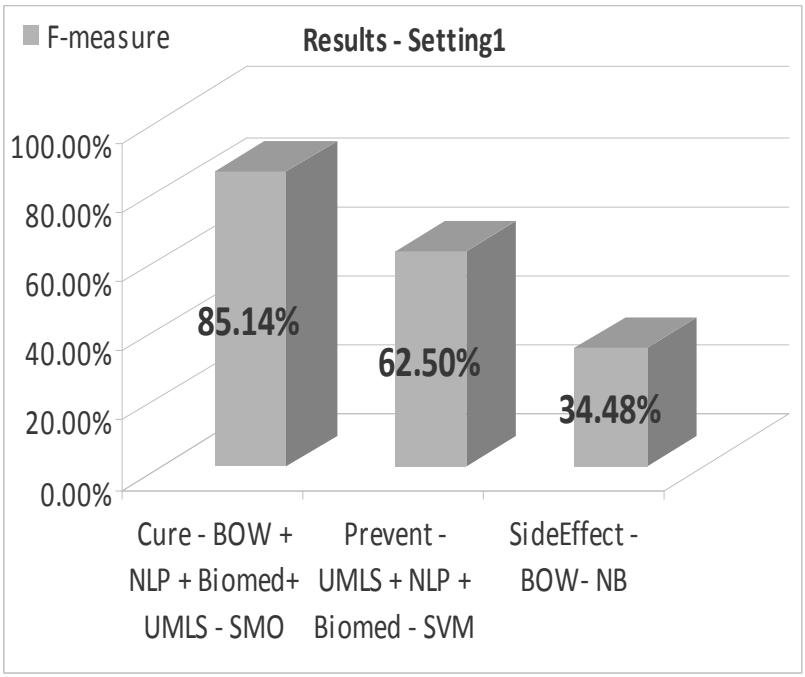

Figure 8.13: Best results for Setting 1 for clinical data.

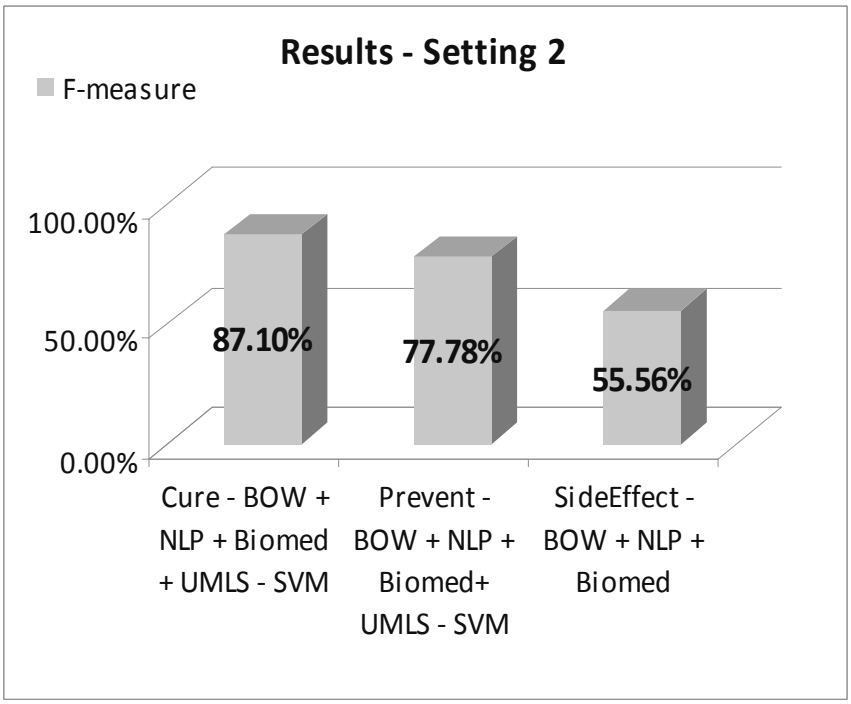

Figure 8.14: Best results for Setting 2 for clinical data.

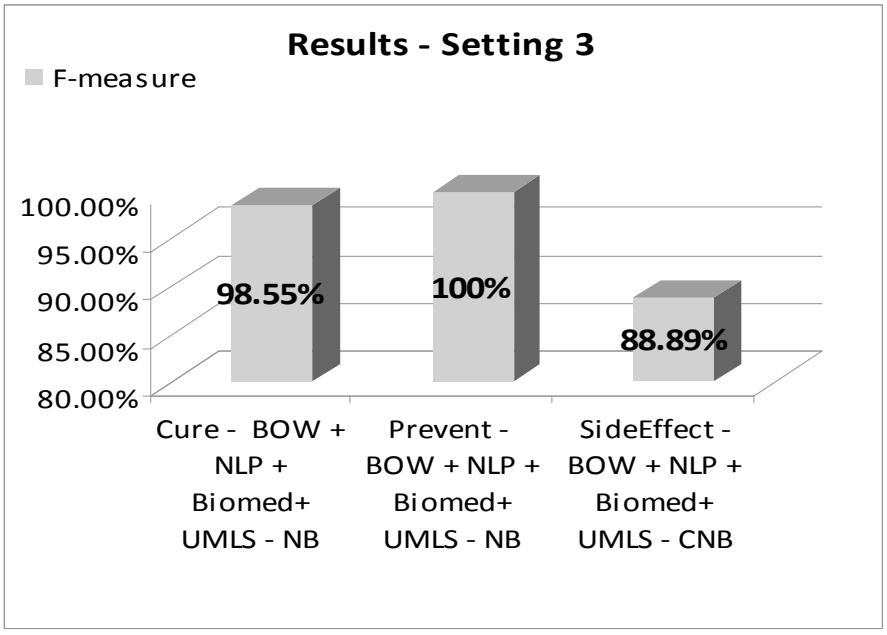

Figure 8.15: Best results for Setting 3 for clinical data.

semantically-charged tasks like mine. The good performance results that I obtain with the second setting also suggest that a prior triage of sentences, informative versus non-informative, can be crucial for a finer grained classification of relations between entities. Setting 1 uses all the sentences, including those that do not contain information about the three relations of interests, while in Setting 3 I used as training data only sentences that I knew a priori to contain one of the three relations. This observation for the results of the third setting also validates the choice of proposing the first task, identify which sentences are informative. For good performance level in the relation classification task, the non-informative sentences need to be weeded out.

Results for the pipeline - task 1 followed by task 2

In this subsection, I present the evaluation results for the two tasks in the hierarchical approach. When looking at the results that I obtain for the second task, the best setting was the one in which I classify sentences already known to be informative, Setting 3. This observation lets us believe that a pipeline of the two tasks is a viable solution for the goal.

To show that a pipeline of results is better as a solution for identifying semantic relations in informative sentences, I need to compare its results to the results of a model that classifies sentences into 4-classes directly: the 3 semantic relations *Cure*, *Prevent*, or *Side Effect* and the class for sentences that are uninformative. For a better comparison between the pipeline and 4-way method, I will use the same classification model for both task 1 and task 2 and the 4-way classification.

The results for the pipeline of tasks are obtained by classifying the three relations only on sentences identified as relevant by the first task. To be consistent, I report the F-measure results. For a better comparison between the pipeline and the flat classification approach, I used the same classification model, CNB classifier using a combination of all types of features (Figure 8.12). The overall results of the pipeline of tasks (task 1 and task 2 with the same classification model) along with the ones for the 4-way classification

are presented in Table 8.7.

Semantic Relation	**4-way Classification**	**F-measure Task1**	**F-measure Task2**	**F-measure Pipeline**
Cure	90.70%	90.72%	95.90%	91.07%
Prevent	71.80%	90.72%	97.56%	52.17%
SideEffect	28.60%	90.72%	88.89%	57.14%
AVG_Rel	*63.70%*	90.72%	94.11%	**66.79%**

Table 8.7: F-measure results for the 4-way classification and pipeline - task 1 followed by task 2 on technical data.

The pipeline of tasks outperforms the 4-way classification scenario by 3 percentage points. An increase of less than one percentage point is obtained for the *Cure* class, and 28 percentage points for the *SideEffect* class. For the class *Prevent* class, the 4-way methodology was superior with 19 percentage points. The reason why the results of the *Prevent* class are lower is the fact that the majority of false positive sentences from the first task, classified as relevant while they were irrelevant, were classified with the *Prevent* class in the second task. The slight improvement obtained by the pipeline for the *Cure* class could be due to the fact that distribution of this class is well represented in the data set, and in the end, it has a higher chance to be correctly classified in almost any ML scenario.

The important type of error that can occur in a 4-way classification task is the high rate of false negatives for the 3 classes of interest. Some sentences that belong to one of the classes of interest get classified in the 4th class that contains uninformative sentences. In a pipeline of tasks scenario, if the result of the first task is high and uninformative sentences are removed, then informative sentences have higher chances to be classified into the right class, especially when the choices are reduced by one. Usually, the amount of data that is classified as being non-relevant is higher than the one relevant to a particular relation. In a 4-way classification task, the majority class

overwhelms the underrepresented ones, while in a pipeline of tasks the balance between that relevant data and non-relevant one is higher and the classifiers have better chances of distinguishing between them. The fact that for the most underrepresented classes, I obtain a high increase in results suggests that a pipeline of tasks has the potential to be superior to a 4-way classification task. Also, if I am to compare the accuracy results of the pipeline, 92.84% and the one of the 4-way classification technique, 91.62%, I can see that the pipeline is slightly better.

8.3.2 Results on the clinical data

Results for the task of concept-pairs selection (task 1)

The aim of this task is to identify which pairs of concepts co-exist in a meaningful relation and which do not. From all possible pairs of concepts identified in a sentence, this first task becomes one that tries to weed out the pairs that do not belong to any relation of interest. During this stage, the focus is not on identifying which is the exact relation, but rather on identifying if there is a strong evidence that the pair of concepts is related. The following classification task is aimed to identify which is the exact relation each of the pairs of concepts initially identified through task 1.

The classification models deployed to solve this first task use a combination of the features mentioned earlier in this section with various classification algorithms. Other classification algorithms than the ones mentioned in this section have been tried in a 10-fold cross validation setting on the training data set with no further improvements.

For some experiments, based on observations obtained on early results, additional lexical features are added: the number of commas that exists in the context of the pair and the window size (the number of tokens between the pair of concepts). The number of commas helped in the classification task, especially for detecting the non-related pairs. If the context between the pair of concepts has a higher number of commas, it represented a good indication that the pair of concepts are part of an enumeration of concepts that

most of the times are not in a particular relation. The window size also helped with this task, since pairs of concepts that are too far apart in a sentence tend not to be related.

Figure 8.16 presents representative F-measure and accuracy results for the task of concept-pairs selection. The baseline represents a classifier that will always predict the majority class.

The x-axis describes the feature set and the classification algorithm. In the captions, *Lexical* features represent the BOW features, number of commas, and window size; *ConType* the types of the pair of concepts; *SV* the union of the top 300 most representative terms for each relation (these features were extracting using the Semantic vectors tool); *DSV* the cosine distance between the pair of concepts and each of the relations; *Genia* features contain the number of noun phrases, verb phrases, and bio-medical entities extracted by the Genia tagger; *ConText* are the features that capture the information extracted by the Context tool.

Although the difference in results between the different classification models presented in Figure 8.16 is not significant, similar to the observation made on the technical data, a richer feature representation technique is associated with better results. This observation is also supported by the fact that feature selection techniques did not improve the classification results.

This task is more difficult than the similar first task on the technical data. In this clinical data, a sentence can contain more than one relation and more than one type of relation can exist at a sentence level. Also, one other challenge is the very short contexts available for the decision-making process.

Some of the features improved the results with the general observation that more diverse features helped the classification task. Due to the sparseness of the data, a limit in the number of features might suggest a threshold that should be used.

The probabilistic classifiers Naïve Bayes (NB) and Complement Naïve Bayes did not perform well on this task. One reason could be the fact that the short sentence context is shared by more than one pair of concepts, creating a high-feature overlap and making

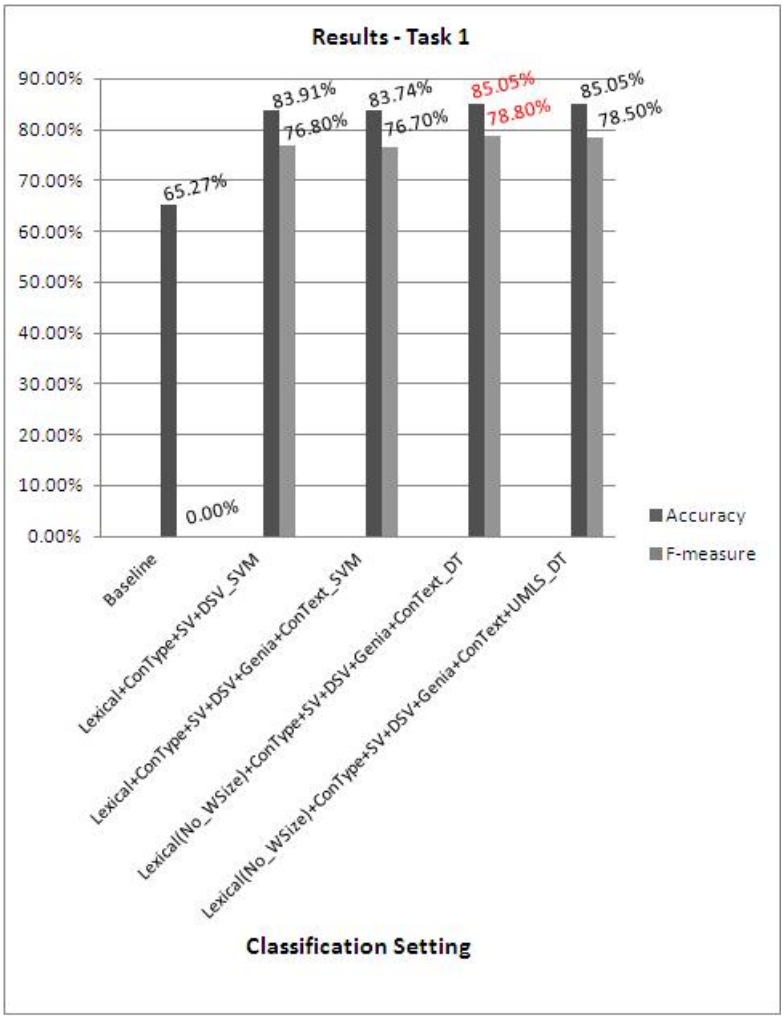

Figure 8.16: Accuracy and F-measure results for the task1, concept-pairs selection on clinical data.

the classifiers unable to predict the right class.

AdaBoost performed slightly better than the probabilistic models. Its improvement in results over the probabilistic models might be caused by the power of the boosting methods to autocorrect some of the errors.

SVM and decision trees are the classifiers that performed the best on this initial task. Decision trees were slightly superior, 3 percentage points, due to their rule-induction power and the ability to focus on features that are more significant for classification. SVM performed well in all experiments due to its power to identify reliable support vectors that can guide the classification and its ability to avoid the overfitting issue.

Results for the task of relation identification (task 2)

In this subsection, I present results for the task of identifying the relation of each pair of concepts. In this experiment, I use as training and test data only instances that have pairs of concepts that co-exist in a relation, I do not use any of the pair of concepts that are not annotated with a relation. We can think as this task as one in which we have all non-relation pairs of concepts removed through task 1. The rationale that stands behind this task is two-fold: we need to identify how well classifiers are able to discriminate between the 8 relations in question and to identifying the contribution of task 1 in the overall performance of the task by comparing a 9-class flat classifier with the pipeline of tasks 1 and 2.

The reported results are based on similar combination of representation and classification techniques as the ones used for task 1. Besides that representation features that have been used in the first task, in this second task, I focus on the verbs that exist in each sentence. The intuition behind this decision is the fact that verbs should be strong indicators for the existing relations. Besides, the annotation guidelines that were given to the human judges were strongly based on the use of certain verbs, *e.g., "Treatment improves medical problem (TrIP). This includes mentions where the treatment cures the problem."* In order to identify verbs in each of the sentence, I used the Genia tagger tool.

In addition to identifying the existing verbs in a sentence, I created groups of verbs associated with each relation. For each of the verbs mentioned in the annotation description of a relation, I identified a set of synonyms for those verbs and considered them as additional features. I created 8 such possible features, plus an additional feature that contain the verbs that were not part of any of these sets of verbs. For example for the relation *treatment improves problem* (Trip), the set of verbs is as follows: *advance, ameliorate, amend, augment, better, boost, civilize, come around, convalesce, correct, cultivate, develop, doctor up, edit, elevate, emend, enhance, gain ground, help, increase, lift, look up, make strides, meliorate, mend, perk up, pick up, polish, progress, promote, purify, raise, rally, recover, rectify, recuperate, refine, reform, revamp, revise, rise, set right, shape up, sharpen, skyrocket, straighten out, take off, touch up, turn the corner, update, upgrade, alleviate, attend, cold turkey, doctor, dose, dress, dry out, help, improve, kick, kick the habit, make better, make healthy, make whole, medicate, mend, minister to, nurse, palliate, quit cold, redress, rehabilitate, relieve, remedy, repair, restore, restore to health, right, shake, sweat it out, treat, heal.*

From the set of different experiments that were tried, Figure 8.3.2 presents representative results for the relation classification on clinical data.

Compared to the first ask, the decision tree classifier was inferior to the SVM classifier. One reason could be the ambiguity of contextual cues in the decision tree due to the limited available context. From previous experience on using decision tree classifiers on clinical data, Chapter 5, the decision trees tend to work well on multiple class classification tasks when the available context is large, *e.g.,* entire discharge summary, abstracts, etc. and not too well on very short textual contexts.

In general, the accuracy of the models that I used to solve task 2 on clinical data lie around 80%. The lowest accuracy results for this task are achieved with a simple BOW representation, 71% accuracy. In terms of F-measure results, the best macro-averaged F-measure is 60.06%.

Based on these results, clearly improved over the first task, I can conclude that the

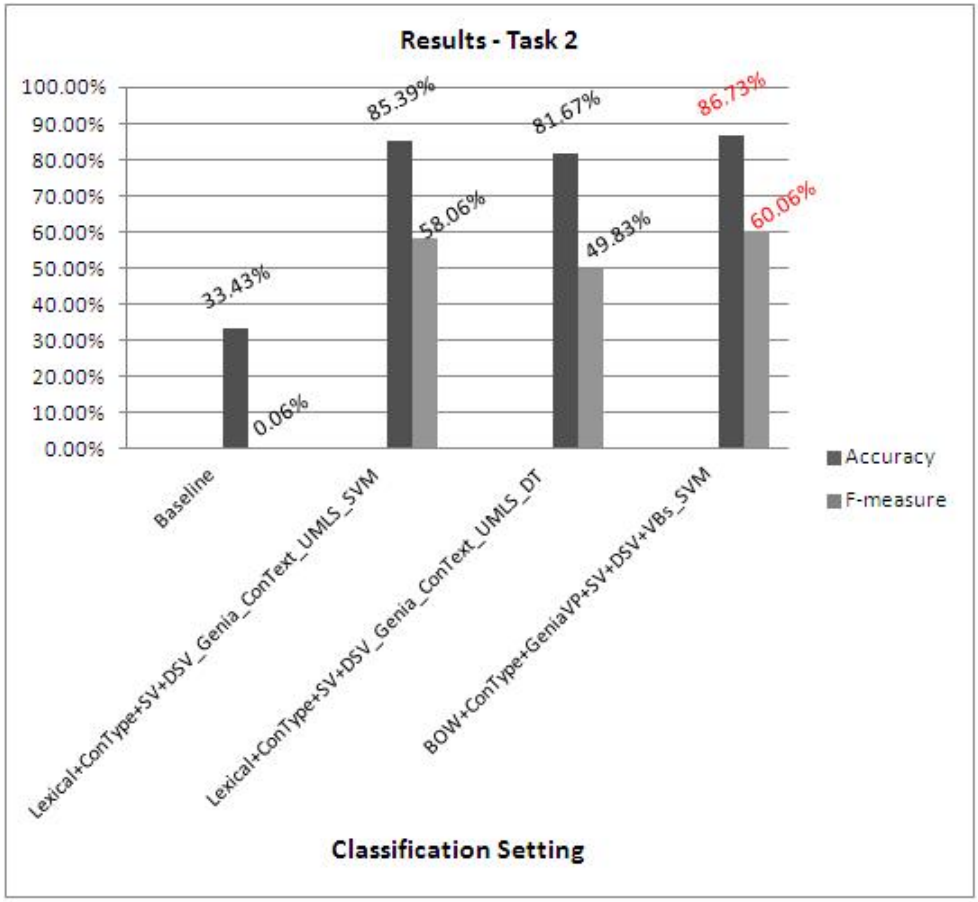

Figure 8.17: Accuracy and F-measure results for task 2, relation classification on clinical data.

initial step of identifying related concepts, task 1, is important in the overall performance of the task of identifying medical relations. Classifying pairs of concepts that are known a-priori as being in a medical relation, task 2, becomes an easier task if the initial step, task 1, is performed. Results to support this claim are presented in the next section.

Results for the pipeline - task 1 followed by task 2

This section presents results for solving the task of identifying and classifying 8 possible medical relations between pairs of medical concepts. Up to this point, in the previous two sections, these two tasks were addressed separately with the scope of identifying which representation and classification technique yields the best results. The global task is to identify which sentence-level annotated medical concepts are related and what is their relation. The methodology that I suggest is based on a hierarchical approach, the results are presented in Figure 8.1. The initial step identifies which pairs of concepts are related while in a second step the actual relation is identified.

In order to show that this approach is successful on clinical data, the same way it was successful on technical data in subsection 8.3.1, I have to show that the results of this pipeline approach are superior to the results of a flat classifier that tackles both tasks in one step. For a better comparison between the pipeline and flat method, I use the same classification model, (Figure 8.16). I used a similar approach to compare the flat classification and the pipeline method when I used technical data, subsection 8.3.1.

Table 8.8 presents comparative results between the pipeline approach and the flat-classification approach. To be consistent with all other experiments, I report the macro-averaged F-measure results. Detailed F-measure results are reported for each class.

The difference in macro-averaged F-measure between the best result of the 9-way classification and the pipeline is of 12.76 percentage points. Other classifiers yielded similar differences with an overall average of 5 percentage points between the method[5].

[5]If I am to report results for the difference between the best flat-classification model and the pipeline approach using the best model for each task, the difference in result is 6.54 percentage points. The best

Semantic Relation	**9-way Classification**	**F-measure Task1**	**F-measure Task2**	**F-measure Pipeline**
PIP	46.00%	78.80%	100.00%	48.16%
TeCP	43.80%	78.80%	43.20%	51.63%
TeRP	83.10%	78.80%	90.08%	88.38%
TrAP	72.50%	78.80%	81.50%	83.33%
TrCP	45.90%	78.80%	43.50%	59.75%
TrIP	33.90%	78.80%	19.40%	45.99%
TrNAP	19.40%	78.80%	12.20%	26.50%
TrWP	12.80%	78.80%	8.40%	9.15%
AVG_Rel	*38.45%*	78.80%	42.67%	**51.21%**

Table 8.8: F-measure results for the 9-way classification and pipeline - task 1 followed by task 2 on clinical data.

The pipeline of tasks outperforms the best 9-way classification technique for 7 out of 8 classes. Differences vary from 13.85 percentage points for relation *TrCP* to 2.16 percentage points for relation *PIP*. The only relation that obtained a decrease in result in the pipeline approach is *TrWP*, 3.65 percentage points.

The decrease in performance for the class *TrWP* could be due to some mistakes from the first step of the pipeline and to the fact that this class is very poorly represented in comparison with the other classes. For all the other classes the improvement is significant, with the hierarchical approach being superior to the flat classification technique.

One type of possible errors that can occur in the flat classification approach is due to the high rate of false negatives - the negative class is larger than any of the other classes. In a pipeline of tasks scenario, if the result of the first task is high and non-related pairs are removed, related pairs have higher changes to be classified into the right class

F-measure result for the flat classification on clinical data is 44.67% obtained by the SVM classifier that uses a combination of lexical, concept type, and distributional similarity features.

later on, especially when the number of classes is reduced by one. In a flat classification approach, the majority class overwhelms the underrepresented ones, while in a pipeline of tasks the balance between that relevant and non relevant instances is higher and the classifiers have better chances of distinguishing between them.

Also, if I am to compare the accuracy results of the pipeline, 81.29%, and the one of the 9-way classification technique, 79.46% we can observe that the pipeline is slightly better.

8.4 Combining technical and clinical data

One contribution this thesis brings represents the fact that it is one of the few studies that combines technical and clinical data for solving the same task of identifying and classifying relations between medical concepts. The goal of this experiment is to improve the overall performance of identifying and classifying medical relations in both technical and clinical data. The fact that technical data can be easily accessed could represent an asset for similar tasks performed on clinical data. In the research literature, Xu et al. (2009) suggests that there is little overlap between the UMLS terms associated with medical treatment concepts and the same type of terms used in randomized clinical trials. These earlier discoveries suggest that important knowledge can be missed if systems are not adapted to the type of data they are used for. It also shows that, in order to have a reliable and vast knowledge about a problem, information coming from multiple sources needs to be taken into account.

In order to combine the two data sets, I first have to map the annotated relations from the technical data with the annotated relations from the clinical data. One of the major problems relation classification tasks face is the fact that some of the relations are significantly underrepresenteted in comparison with the others, for example the *TrWP* relation has 49 instances in comparison with the *TeRP* relation that has 1,305.

Based on the annotation description of each relation from the clinical and the technical

data, I can map data for the following three relations: the *Cure* relation from technical data with the *treatment improves problem (TrIP)* from clinical data, the *Prevent* relation with *treatment is administered for medical problem (TrAP)*, and the *SideEffect* relation with *treatment causes medical problem (TrCP)*.

After "similar" technical and clinical data are identified, the next step is to combine these data and try to improve the overall results of identifying and classifying relations in both technical and clinical data.

8.4.1 Combining data for testing on technical data

The first set of experiments tries to combine technical and clinical training data to improve the results on the technical data. In order to perform this experiment, I used as training data both the instances from the original technical data for the classes: *Cure*, *Prevent*, and *SideEffect*, and added the new instances from the clinical data that map these relations. As test data, I used the same technical test set that I used in the experiments described in Section 8.1, subsection 8.1.1. I focused only on task 2 on technical data since very few instances will be added as training examples to the *informative* class used in task 1.

The rationale behind this experiment is to try to identify if a standard supervised machine-learning approach of combining these types of data improves the results on the technical data. Also, I would like to test if complementary knowledge from the clinical data can be added in a simple manner to the knowledge extracted from the technical data, for benefiting the task of relation classification.

As a classification method, I used the same representation and classification algorithms as the ones that obtained the best results on technical data. A combination of BOW, UMLS, and all Genia-type relations in combination with the SVM classifier.

Table 8.9 presents the data sets and the results obtained in this experiment for all three relations of interest annotated in the technical data.

Besides the experiment when I combined the data for the second task of the pipeline, I

Relation	**Train** (technical + clinical)	**Test** (technical)	**F-measure**
Cure	554 + 107	276	82.00%
Prevent	42 + 1,423	21	80.50%
SideEffect	20 + 296	10	69.30%

Table 8.9: Combining technical and clinical data for testing on technical data.

also performed a similar experiment when I combined the data for the flat-classification approach. Both results obtained after combining the technical and the clinical data sets did not improve the best results of the model that is trained and tested only on technical data. Possible reasons for the decreased performance could be the fact that the technical training data is overwhelmed by the clinical data that is added. From Table 8.9, we can observe that the number of instances added from the clinical data for two out of the three classes is significantly larger. Another reason could be the difference in vocabulary between these types of data, it is known in the medical literature that significant differences in vocabulary exist even between the same type of data coming from different medical departments, different sources of clinical data, *e.g.,* clinical notes, discharge summaries, radiology reports, etc. In addition, the particular clinical data that I use has a major difference: while in the technical data one single relation is identified at a sentence level, in the clinical data more than one relation can be annotated at a single sentence level.

The overall conclusion for this experiment is that better results are obtained when a classifier is trained on the same type of data that will be used for classification. Other ways of combining these two types of data could improve the results, but fine details regarding the distribution of the classes, data type, classification representation and algorithms represent important variables that need to be taken into consideration.

8.4.2 Combining data for testing on clinical data

Similar to the previous set of experiments, the next set of experiments trains a classifier on both technical and clinical data and tests its performance on the clinical test set data. Similar to the previous experiments, I combine only the instances that are similar in the way they were annotated. The three relations from the technical data will be added to the clinical training data for the three mapped relations. Compared to the original training data, now three out of the 8 classes will have a larger number of instances.

Because the identification and classification task on clinical data has been shown to be more difficult than on the technical data, both task 1 and task 2, I decided to run two sets of experiments. Each set is focused on one of the tasks. For both experiments, I use similar classification methods as I used on clinical data.

For task 1 on clinical data, identification of related pairs of concepts, I use a training data set that represents a combination of technical and clinical instances, the original training clinical data plus the mapped technical instances. The new created training data represents a new *Related* class while the *Non-related* class remains the same as in the original clinical data experiments. These experiments showed to have insignificant effect on the performance of the classifier compared to the original results. A possible reason could be the fact that little technical data is added in this task and because the two classes are well represented; the few hundred instances added from the technical data did not change the results.

For the second task on clinical data, I use a similar methodology as I used for technical data but in addition, I add new instances that were used for task 2 on clinical data. Table 8.10 presents representative results for the classifiers trained on a combination of technical and clinical data for task 2, for all eight relations of interest.

Similar to the results obtained on the technical data, the new results of combining the data to test on the clinical data did not improve the relation classification task. I also performed experiments when I used the flat-classification technique on the combination of data.

Relation	**Train** (clinical + technical)	**Test** (technical)	**F-measure**
PIP	1,239	1,989	100.00%
TeCP	303	588	60.30%
TeRP	1,734	3,033	93.30%
TrAP	1,423 + 42	2,487	84.50%
TrCP	296 + 20	444	55.60%
TrIP	107 + 554	198	25.30%
TrNAP	106	191	26.40%
TrWP	56	143	8.50%

Table 8.10: Combining technical and clinical data for testing on clinical data.

The decreased performance of this experiment can be explained by the same reasons mentioned earlier. The class *TrCP* obtained a 1.60 percentage points improvement when the combination of data is used. For the other two classes the decrease in performance could be associated with the large number of instances added to the class or the overall effect of introducing a different type of data.

8.5 Discussion and Conclusions

8.5.1 Discussion

Technical data In the following paragraphs, I discuss the results that I obtained for the tasks of relation identification and classification on technical data, while later on in this section I address the results on the clinical data.

For the first task, the one of identifying informative sentences, the results show that probabilistic models based on the Naïve Bayes formula obtain good results. The fact that the SVM classifier performs well shows that the current discoveries are in line with the

literature. These two classifiers have been shown to perform well on text classification tasks. Even though the independence of features is violated when using Naïve Bayes classifiers, they still perform very well. The AdaBoost classifier was outperformed by the other classifiers, which is a little surprising taking into account the fact that it is designed to focus on hard-to-learn concepts. In previous work, it was shown to perform well on medical domain texts with imbalanced classes (Frunza and Inkpen, 2008b). One reason why the AdaBoost classifier did not perform well might be that fact that in previous experiments I used the entire abstracts as source of information while in this current study I use sentences.

In the NLP and the ML community, BOW is a representation technique, that even though it is simplistic, most of the times it is really hard to outperform. As shown in Figure 8.9, the results obtained with this representation are among the best ones, but for both tasks, it is outperformed when combining it with more structured information, such as medical and biomedical concepts.

One of the major contributions of this work is the fact that the current experiments show that additional information in the representation settings brings improvements for the task of identifying informative sentences. The task itself is a knowledge-charged task; the labeling process involves a human-intensive annotation process, since relations between entities need to be manually identified. The experiments designed for the automatic task aim to show that classifiers perform better when richer information is provided. In the first task, the CNB classifier using all the representation techniques obtains the best result of 90% F-measure, which is significantly better than the other representations. The classifier is specially designed for imbalanced data, the fact that it proved to be one of the best in text classification tasks, even on short texts, was somewhat a foreseeable result.

The results obtained for the second task suggest that when the focus of a task is to obtain good reliable results, extra analysis is required. The best results are obtained with Setting 3 when a model is built and trained on a data set that contains all three data

sets for the three relations. The representation and the classification algorithms were able to make the distinction between the relations and to obtain the best results for this task. Similar observations as the ones obtained for the first task are valid: probabilistic models combined with a more informative feature representation bring the best results. The best results obtained are: 98% F-measure for the class *Cure*, 100% F-measure for the class *Prevent*, and 75% F-measure for the *SideEffect* class.

The fact that I obtain the best results when using Setting 3 also validates the proposed methodology for a pipeline of tasks in order to better classify relevant information in the three semantic relations. It is more efficient to solve the second task when using data that is known a-priori to contain information about the relations in question, rather than identifying which sentences are uninformative as well. In order to better validate the choices made in terms of representation and classification algorithms and to directly compare with the previous work, additional experiments for all 8 semantic relations originally annotated on the data set were performed. These experiments are addressing exactly the same task as the previous work (Rosario and Hearst, 2004) and are evaluated with the same evaluation measure, accuracy. I will report in Figure 8.18 only the best results with the algorithms and representations that I used for this task.

The first bars of results are obtained with the best model for each of the 8 the relations (e.g., for *Cure* the representation that obtains the best results is reported, a representation that can be different from the one for another relation; the label of each set of bars describes the representation); the second bars of results report the model that obtains the best accuracy over all relations (one representation and one classification algorithm are reported for all relations - CNB with BOW+NLP+Biomed features), and the third bar of results represent the previous results obtained by Rosario and Hearst (2004).

The accuracy measure is reported, since it is the measure that was reported in the previous work. As depicted in the figure, the results obtained in this study outperform the previous ones. In one case, the same low results are obtained; for the *No_Cure* class

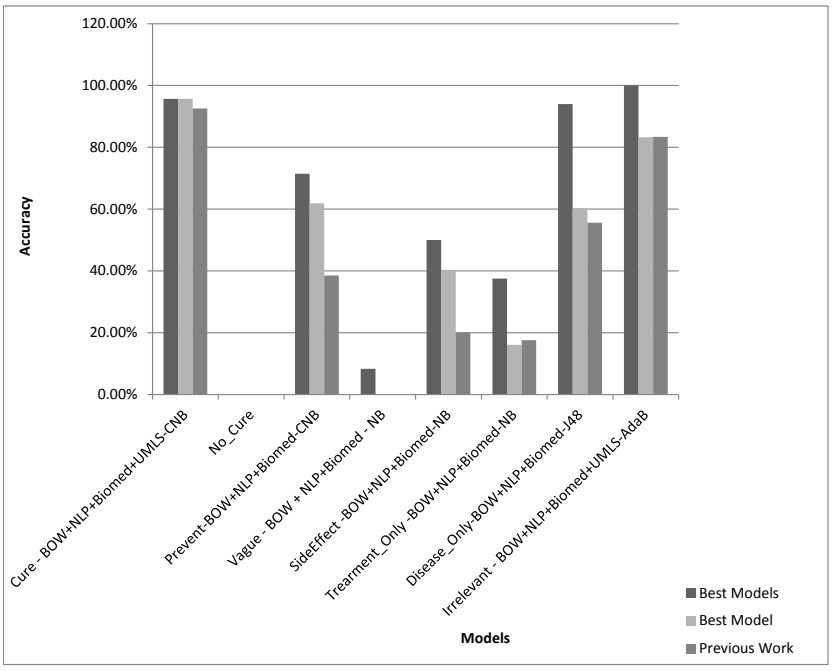

Figure 8.18: Comparison in results for all annotated relations in the technical data set.

the low results are due to the fact that this class is underrepresented in the data set, by only 4 examples in total.

The class *Vague* obtains similar results when one model is used for all relations, but it outperforms previous results when the best model is chosen for this class. For the other relations, the results are better with either the same model for all relations, or for the best one for each.

Clinical data This section is dedicated to discuss the results obtained for the tasks of relation identification and classification using clinical data. One of the main contributions this thesis brings is that addresses important tasks that use clinical data. The research literature is not rich enough in conclusions and analysis on this type of data, as it is on others.

The next paragraphs present conclusions for both tasks when using clinical data. In the hierarchical approach that I propose, the first task identifies related pairs of concepts that need to be further classified in one of the 8 relations of interest. In this task different representation and classification techniques are used. Similar to the overall methodology that I use in this thesis, features that capture lexical, syntactic, and semantic information with emphasis on the domain of the task showed to be the best way of representing data, but with improved results obtained when subtle observations of the data are incorporated into the representation. As classification techniques, state-of-the art classifiers are used. Emphasis is put on classifiers that are believed to perform well and fit the description of the tasks.

The initial task is a 2-class classification problem where non-related pairs of concepts are weeded out. The best model that is able to perform this task is a decision tree classifier that uses a combination of all types of features. The SVM classifier performed similarly with no significant difference. Similar to the task on technical data, a rich representation improved the results. The only features that did not show a clear contribution represent the UMLS features. Similar to the literature observations, UMLS features get mixed

results due to their broad coverage and false positive errors that are introduced, *e.g.,* non-medical tokens that get classified as medical concepts. The models that I present in this thesis are able to perform this task with an accuracy of 85% and an macro-averaged F-measure result of 78.5%. The results are slightly lower than the similar task on technical data because of the fact that the clinical data is harder to classify, due to possible multiple-class annotations in a single sentence.

Besides the straight-forward machine learning approach of solving this first task, I have also tried other ways of identifying related pairs of concepts in clinical data.

An alternative technique to the straight-forward ML approach follows the lines of a cascading classification method. I initially classified instances in the two classes, then I used the results obtained from this classifier and I constructed another classifier using a set of three features: the initial classifier's prediction (related versus non-related), the difference of the semantic distribution of the pairs of concepts in the related and non-related corpus, and the window size between the pairs of concepts. The rationale behind this experiment was to try to correct the original classifier and try to improve the classification results. I considered this few features because the initial classifiers is using the majority of all other features. While this experiment did not improve the results, its results are close to the value of 76.2% F-measure achieved by one of the best models for task 1.

Another set of experiments performed for the first task consists in a voting technique. I considered 8 binary-classifiers obtained from the original 9 classes, each classifier is focused on one of the relations of interest plus the non-related class. I classified each instances with these eight classifiers and considered a majority vote as being the final class of an instance. The negative majority class competed with the sum of the non-negative classes assigned by each classifier. This experiment follows similar conclusions as the one above, while the results did not improve (the difference was not significant).

The third type of experiments focus on using unlabeled data that was supplied with the annotated clinical data in the i2b2 competition. In an initial step, I identified only

potentially-good candidate sentences to use as unlabeled data. I used one of the already trained classifiers for distinguishing related versus non-related instances. This selected unannotated data represents additional training data for an SVM implementation, UniverSVM[6] that learns the classification task using unlabeled instances. These results were slightly lower than the previous experiments, suggesting that future investigation in this direction could represent a reliable solution to this task, especially since large amounts of unlabeled data are easily accessible.

Alternative experiments for the second task on clinical data looked into dividing the 8 classes of interest in three types of classes, using the type of concepts involved in a relation. The 8 relations fall into three categories: relations between medical problems and treatments, tests and problems, and medical problems and other medical problems. Based on this, I used three classifiers, each focused only on one type of relation. Overall, these three types of classifiers were not superior to the classifier that best discriminated between the 8 relations of interest and the 60% macro-averaged F-measure results result were not outperformed.

The real-life scenario of identifying meaningful relations between pairs of concepts in clinical data, and not only, naturally follows the two tasks that I address: identification of related pairs followed by a relation classification task. Another way of solving these two tasks could be by using a single classifier that classifies each pair of concepts in the relation of interest plus a non-related class. In this thesis I compared both this approaches and show that the former method is significantly better than the latter one.

A contribution of this current work is the fact that I show that on both technical and clinical data, the pipeline approach significantly improves the results over a one-step classification approach. On technical data, the difference between these approaches is of 3 percentage points while on the clinical data the difference is of 6.50 percentage points in favor of the pipeline technique.

The best reported results on the clinical data were a micro-averaged F-measure result

[6]http://mloss.org/software/view/19/

of 73.65% obtained in the i2b2 competition. The best micro-averaged F-measure result that I obtain with the pipeline approach is 71.91%, less than 2 percentage points lower. In the competition, the mean results were 59.58%, the median 66.45%, with a standard deviation of 18.88%. My results clearly show that the methodology that I propose is a competitive one for addressing the task of relation identification and classification in clinical and technical data.

Besides individual contributions on technical and clinical data, my research study presents one of the first works on combining technical and clinical data to solve the same task. Even though the results of combining these types of data for this particular task did not improve individual results, interesting insights and conclusions obtained can benefit the research community at large.

The straight-forward machine-learning approach of bringing together instances from the two types of data showed not to be the desired way of combining knowledge coming from technical and clinical data. The medical domain, especially the clinical one, maybe more than any other domain, faces important issues that hinder many of the computational linguistics and machine-learning techniques. Thorough preprocessing steps need to be performed in order to identify the best representation and classification techniques to overcome problems like vocabulary variations, writing style, misspelling errors, ungrammatical texts, etc. While technical data has important issues that need to be taken into account when using a machine learning problem-solving approach, clinical data is even harder to deal with.

I am still confident that knowledge coming from one type of data can be used and complemented with the other type of data. Since the technical data tends to be larger and much easily to be accessed, it could be used as a type of background knowledge from which statistics and observations can be drawn and used on the clinical data. Results obtained when combining technical and clinical data in the straight-forward manner that I describe in this chapter show that subtle differences in the characteristics of the data need to be taken into account. Rarely, a simple plug-and-play approach yields

improved results. Most often, general observations and guidelines coming from previous experiences and related literature need to be complemented by thorough analysis and optimization techniques for obtaining improved results.

Even though when combining the two types of data, for some of the classes a large number of instances were added, the final result did not improve. Sometimes more is not better, and, even in a supervised machine-learning approach quality can substantially beat quantity.

Similar to the general methodology that I follow in addressing the major goal of personalized medicine by focusing on four individual problems, the divide-and-conquer approach has shown to work well for computational linguistics and machine learning problem-solving approaches.

8.5.2 Conclusions and future work

The conclusions of the study suggest that domain-specific knowledge improves the results. Probabilistic models and SVM-based classifiers are stable and reliable for tasks performed on short technical and clinical texts. The representation techniques influence the results of the ML algorithms, but more informative representations are the ones that consistently obtain the best results.

This chapter addresses the overall task of relation identification and classification between pairs of medical concepts identified both in technical and clinical data. I address this general problem by using a hierarchical approach in which I first identify related pairs then I classify them. Results obtained on both technical and clinical data suggest that this significantly improves the results over a one-step methodology approach.

The differences in the type of data, technical versus clinical, suggest that more subtle ways of combining knowledge coming from these two types of data is required. In the same time, experimental results suggest that classification models focused on single data-types are superior to hybrid-data trained classifiers.

The first task addressed on technical data has applications in information retrieval,

information extraction, and text summarization. I identify potential improvements in results when more information is brought in the representation technique for the task of classifying short medical texts. I show that the simple BOW approach, well known to give reliable results on text classification tasks, can be significantly outperformed when adding more complex and structured information from various ontologies.

Similarly, the first task on clinical data aimed to identify related pairs of concepts has potential to be used for further more elaborated tasks. Representation techniques that are successful for this task are the ones that use richer, domain-specific, and task specific information. While the decision tree classifier insignificantly improved the results over the SVM-based classifier, the general conclusion is that the SVM classifier is a reliable algorithm to use.

The second task on both technical and clinical data, relation classification, can be viewed as a task that could benefit from solving the first task first. In this study, on technical data, I have focused on three semantic relations between diseases and treatments. This current work shows that the best results are obtained when the classifier is not overwhelmed by sentences that are not related to the task. Also, to perform a triage of the sentences (task 1) for a relation classification task is an important step. In one classification model, I included the sentences that did not contain any of the three relations in question and the results were lower than when I used models trained only on sentences containing the three relations of interest. These discoveries validate the fact that it is crucial to have the first step to weed out uninformative sentences, before looking deeper into classifying them. Similar findings and conclusions can be made for the representation and classification techniques for task 2.

The similar task on clinical data is focused on 8 medical relations with an overlap of three relations between the two types of data. The best results for this task are obtained by an SVM classifier that focuses on verbs extracted from sentences. The identification and classification tasks on clinical data have shown to be more difficult than the ones on technical data due to the nature of the annotations, single versus multiple relations in a

single sentence, and due to the higher number of classes. Never-the-less, the results that I obtain outperform and compete with other results on the same data sets obtained in related work.

The above overall observations support the pipeline of tasks that I propose in this work. The 28 percentage points improvement in results that I obtain for the most underrepresented class in the technical data shows that a framework in which task 1 and task 2 are used in pipeline is superior to the one step classification approach. Similar observations are drawn from the clinical data experiment, the pipeline of tasks improves the overall averaged single classifier by 6.5 percentage points with class differences varying from 12.90% percentage points to 2.50% percentage points.

On technical data probabilistic models combined with a rich representation technique bring the best results, while on clinical data the SVM classifier has shown to be a reliable solution.

The straight-forward combination of technical and clinical data did not show to be beneficial for the tasks of relation identification and classification due to differences in the annotations, type of data and vocabulary variations. Better results are obtained when each data-type classifier is optimized individually.

As future work, I would like to extend the experimental methodology, to use additional sources of information as representation techniques, and to focus more on ways to integrate the research discoveries in a framework that could be deployed to consumers. In addition, I would like to focus on different ways of combining the types of data and in bringing in additional information coming from the Web. Identifying and classifying medical-related information on the Web is a challenge that can bring valuable insights to the research community and to the end user. I also consider as potential future work ways in which the framework's capabilities can be used in a commercial recommender system and in integration in a new EHR system.

Chapter 9

Conclusions and Future Directions

9.1 Conclusions

In this section, I present important conclusions that are drawn from the research presented in this thesis. The main objective of this thesis is to build automatic methods that facilitate the identification and extraction of valuable medical knowledge from medical texts to achieve personalized medicine. Relevant research has shown that such tools can contribute to the improvement of the medical practice.

In this work, I presented models that can identify disease presence in clinical data, Chapter 5. These models use various representation techniques based on lexical information (BOW representation), syntactic information (noun-phrase representation), and semantic information extracted from controlled vocabularies (UMLS concepts). These representation techniques are used in combination with suitable classifiers. The model that obtained the best results is using a BOW representation technique in combination with an adaptive classifier. The results suggest that, at times, simple lexical representations in combination with the right algorithm can achieve good results.

In a possible real-life scenario, after the disease is automatically identified using the above mentioned tool, evidence-based medical information can be automatically identified and extracted to further assist the healthcare provider in the assessment and

diagnosis process. In Chapter 6 I present two methodologies intended to automatically identify published abstracts to be relevant or not to a specific medical problem. The first methodology is based on a straight-forward machine learning approach, while the second one uses specific information from the protocol of building systematic reviews. The results suggest that it is beneficial to use additional specific knowledge to obtain improved results. The combination of lexical and medical concepts along with the CNB classifier performed the best. The experimental results show that a good recall performance can be obtained on a real data set. The results also show that the best results are obtained when the classifier's output is combined with the judgments from one of the human experts.

Genetic information represents one of the most important pieces of information in a personalized medical practice. Functional genomics, in particular gene regulation, is the task that is presented in Chapter 7. The way this task is tackled, by only using textual information (Medline abstracts with gene mentions) represents a major contribution. For this task, the classifier that obtained the best results is the adaptive classifier when using BOW features; UMLS phrases also represented a good choice for features. Binary feature values clearly outperformed frequency values, mostly due to the highly-imbalanced data set.

One of the most important steps in assessing a medical case stands in the ability to prescribe the right therapy for a certain medical case. This step is addressed in this thesis by means of identifying semantic relations that exist between medical entities in technical and clinical data. Identifying and mining relations between medical entities represents valuable information for healthcare providers and consumers. Because of its potential value, a significant part of this thesis is dedicated to this task. Experiments are performed using technical texts (Medline abstracts) and clinical data, as described in Chapter 8. The proposed solutions for identifying semantic relation between medical concepts represent a hierarchical technique. Relevant instances that contain related pairs of concepts are identified and subsequently classified by their semantic relation.

Lexical, syntactic, semantic, and domain specific information is used to represent the data. Several classification algorithms are tested. The results suggest that a richer representation technique in combination with probabilistic classifiers, CNB and NB, obtain the best results on technical data. When working with clinical data, the best classification model are the DT and SVM with a combination of features that take into account the distributional semantics of the relations, verbs associated with each relation and the lexical features extracted from the training data. Similar to the observations on the technical data, a richer representation technique helped the classification task when using short contexts.

Table 9.1 presents an overview of the methodology that is being used in this thesis for all tasks.

As classification algorithms, CNB, AdaBoost, and SVM performed well. CNB showed to work well both on long and short texts, with improved results on shorter texts and on imbalanced data sets. CNB, even though it does not follow the constraint of the attribute independence, often works very well on texts. The obtained results proved that the classification algorithm works well when the feature space manages to capture the specifics of the classes, without much overlap between features belonging to different classes and when the features are well . From a different perspective, the CNB classifier showed to work well with relatively short texts.

For the disease identification task, where an instance could be of few pages-long and for the gene function prediction task where an instance could have many abstracts associated, the AdaBoost classifier performed the best. AdaBoost also proved to be the best learning model for the tasks that use longer texts with highly-imbalanced class distribution. As it is intended to be, this algorithm manages to autocorrect itself and focuses on the correct classification of instances belonging to the underrepresented classes.

The SVM classifier showed to be one of the most reliable classification models for the task of relation identification and classification on the clinical data set. The results obtained by this classifier proved to be the best ones when different representation

techniques were used. For the same task, the DT classifier showed a slight improvement in the first step of the hierarchical methodology, the relation identification part, while in the second step, the SVM classifier proved to be a better solution.

As a representation methodology, for some tasks, simple BOW features outperformed the more elaborated, deeper syntactic and semantic features. The diversity of the BOW features, e.g., tense, number, inflections, etc., managed to capture a diversity in vocabulary that, in the end, translated into better classification results. This conclusion characterizes the disease identification task and the functional genomics tasks. Also, the BOW binary representation performed well on long texts, while for short texts additional features showed to help the classification tasks.

In this thesis, the sparsity problem is alleviated through features that map the lexical tokens to their semantic meanings. In order to achieve this, specific tools and ontologies are used. Each of the four tasks that are addressed in the thesis has shown to need additional tuning in the feature representation for achieving a higher performance. In the relation identification and discrimination task, the use of verbs and verb phrases showed to substantially improve the results. When using short contexts for classification, additional features, e.g., medical concepts, noun and verb phrases that captured small differences between classes, were the best representation techniques. Feature engineering techniques like feature selection and tf-idf feature values showed to be ineffective for these tasks. These types of techniques tend to work well with relatively long contexts. Some of the representations techniques took into account the negation aspect, but for the relation identification task, this type of features did not show to be of significant importance.

A large part of this thesis is devoted to research performed on clinical data. Besides the individual contributions that are brought with each task, special attention has been devoted to combining the clinical and technical data. In the research literature, there are not many attempts to transfer medical knowledge between these types of data. The conclusions that can be drawn from these current experiments suggest that a straightforward way of combining these two types of knowledge does not represent a successful

way. More data does not necessary means better results, even if we are in a supervised-learning setting. A possible way of successfully combine these two types of data could represent a hybrid system that trains classifiers on each type of data and that are later combined through a voting or bootstrapping manner.

The same conclusion, more data does not mean better results, can be drawn from the experiments performed when using unlabeled data. Results suggest that careful engineering needs to be performed in order to improve results when unlabeled instances are used as additional training data.

To conclude this section, the next few remarks, highlight general observations carried out by the current research:

- general straight-forward techniques need to be adapted to the specifics of the task and domain of the problems for improved results;

- often, diverse representations techniques need to be complemented by different methodologies for achieving the best results. The hierarchical methodology that is proposed in Chapter 8 clearly outperformed the general machine-learning technique for the relation identification and classification tasks. The per-question technique, that uses the specific human workflow of building systematic reviews, described in Chapter 6, outperforms the straight-forward machine learning method. A hybrid combination of the human effort with the automatic tools achieved the best results for this task;

- for certain tasks, simple representation techniques outperform more elaborated representations, e.g., for the disease identification task and functional genomics task. But, in general, richer features are the ones that best capture the task and better represent the data suggesting that thorough investigations need to be performed for achieving desired results;

- textual information has shown to be a powerful source of knowledge for biology-specific tasks. The methodological setting that uses only textual data improved

earlier relevant research that used additional biology-specific knowledge.

- more data does not necessary translates in better results. Careful analysis and fine-crafted techniques are required for successful integration of diverse and unlabeled data;

- short contexts with poor discriminative features can be successfully complemented by additional syntactic and semantic attributes. Verb-type features showed to be beneficial for the relation discrimination task;

- a divide-and-conquer approach was shown to be successful in the analysis, data representation, classification, and addressing this thesis' goal.

9.2 Future directions

A possible future research direction for the current work represents the integration of the four objectives. The automatic diagnosis process can be augmented with knowledge coming from systematic reviews, the gene-function information can be used for inferring relations between diseases, treatments, dosage prescription, etc. Inferring automatic medical relations between medical concepts can facilitate the implementation of a better targeted preventive care, and also identify patients that are better suited to be part of a particular clinical trial.

The results of the integration of the systems that I presented can be used to create automatic summaries of important information about a particular patient or group of patients. The textual information could also be complemented by other types of medical data, e.g., signal, image, sound, etc., for a better overall medical assessment. I believe that geographical information is another type of data that can be successfully used in the medical domain. Geo-medicine is a new and emerging way of making use of geo-location data to use and infer medical-related information.

Another interesting direction of research of my current work represents the integration of the time-line dimension in the medical diagnosis process. The questions that should be asked and answered are: can medical providers assisted by intelligent tools better identify and foresee what will the patient experience next time she/he will encounter the medical provider? Is there enough information at time t to take the right decision path without waiting, running more tests, etc., till time *t+1*?

I consider patient outcome prediction as being another important research direction that can be built on my current research. With the right information extracted from the medical records of a patient or group of patients, we could infer important medical relations that could determine the right treatment to be prescribed to the right patient. For this particular task, I believe that the genetic information represents a crucial type of data.

From the point of view the methodologies that I used, possible future work that could extend this current work stands in further investigation on different representation techniques, e.g., deeper syntactic analysis and more domain-specific knowledge. The negation and hedging information represents valuable linguistic information for the medical domain, it can play an important role in the overall performance. I believe that further investigation into using this type of information in the representation techniques will be beneficial.

Different learning algorithms could also represent future directions for the current work. The conditional random fields (CRF) (Sutton and Mccallum, 2006) algorithm proved to work well on textual data and also on relational learning, not only on labeling tasks, e.g., part-of-speech tagging. I would also like to use the SemRep[1] tool, provided by NLM for the relation identification task. This tool makes use of the UMLS concepts and it is able to automatically identify relations between UMLS concepts.

While the current work provides a step forward in several areas of interest related to PM, it is clear that it only makes a small step towards the full use of self-ruling

[1] http://skr.nlm.nih.gov/papers/index.shtml#SemRep

automatic solutions that are ready to be used in clinical settings. Future investigation on transferring knowledge from technical to clinical data and vice versa can only bring further improvements to the current results. The same type of investigations can improve the use of unlabeled data for the under-represented classes to boost the models' performance.

Task	**Disease Id.**	**Building SRs**	**Gene Function**	**Relation Id.**	
				Tech.	**Clinical**
Data sets					
Balanced					
Imbalanced	√	√	√	√	√
Long	√		√		
Short		√	√	√	√
System					
Supervised	√	√	√	√	√
Semi-supervised					√
Representation					
Lexical [BOW]	√	√ +	√	√ +	√ +
Syntactic [NPs/VPs]	NPs			NPs VPs **+NPs+VPs**	NPs VPs NPs+VPs
Semantic [UMLS/Genia/Vectors]	UMLS	**+UMLS**	UMLS	UMLS Genia **+UMLS+Genia**	UMLS Genia UMLS+Genia **+Vectors**
Verbs					+ √
Evaluation Measure	F-measure	Recall-F-measure	AUC	F-measure	F-measure
Classification Algorithms	CNB **AdaBoost**	**CNB**	CNB **AdaBoost** SVM	**CNB** AdaBoost SVM DT	CNB AdaBoost **SVM** **DT**

Table 9.1: In the table, *Long* stands for textual data longer than abstracts, *Short* stands for abstracts and sentences; where "+" appears it shows that the representation is a unification of the marked feature types; the information marked in bold represents the best results.

Bibliography

Abrahamas, Edwards (2009), "The case for personalized medicine." *Journal of diabetes science and technology*, 3, 680–684.

Agichtein, Eugene and Luis Gravano (2000), "Snow-ball: Extracting relations from large plain-text collections." In *Proccidings of the 5th ACM Internal Conference on Digital Libraries (ACMDL'00)*, 85–94.

Amarasingham, Ruben, Laura Plantinga, Marie Diener-West, Darrell J. Gaskin, and Neil R. Powe (2009), "Clinical information technologies and inpatient outcomes: A multiple hospital study." *Archives of Internal Medicine*, 2, 108–114.

Ambert, KH and AM Cohen (2009), "A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection." *Journal of American Medical Informatics Association*, 16, 590–595.

Ananiadou, Sophia and John McNaught (2006), *Text Mining for Biology and Biomedicine*. Artech House.

Aphinyanaphongs, Y. and C Aliferis (2005), "Text categorization models for retrieval of high quality articles." *Journal of the American Medical Informatics Association*, 12, 207–216.

Aronson, Alan R. and François-Michel Lang (2010), "An overview of MetaMap: historical perspective and recent advances." *Journal of the American Medical Informatics*

Association (JAMIA), 17, 229–236, URL `http://dx.doi.org/10.1136/jamia.2009.002733`.

Blaschke, C., M. Andrade, C. Ouzounis, and A. Valencia (1999), "Automatic extraction of biological information from scientic text: Protein-protein interactions." In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB99)*, 60–67, AAAI Press.

Brady, Scott and Hagit Shatkay (2008), "Epiloc: A (working) text-based system for predicting protein subcellular location." *Pacific Symposium on Biocomputing*, 13, 580–591.

Buchan, William (1784), *Domestic medicine or a treatise on the prevention and cure of diseases by regimen and simple medicines*. Dublin: H Chamberlain and others.

Bundschus, Markus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel (2008), "Extraction of semantic biomedical relations from text using conditional random fields." *BMC Bioinformatics*, 9, 207, URL `http://www.biomedcentral.com/1471-2105/9/207`.

Bunescu, R. and R. Mooney (2005), "Shortest path dependency kernel for relation extraction." In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 724–731.

Bunescu, R., R. Mooney, Y. Weiss, B. Schölkopf, and J. Platt (2006), "Subsequence kernels for relation extraction." *Advances in Neural Information Processing Systems*, 18, 171–178.

Cai, YueHong and XianYi Cheng (2009), "Biomedical named entity recognition with tri-training learning." In *Biomedical Engineering and Informatics, 2009. BMEI '09. 2nd International Conference on*, 1 –5.

Caropreso, Maria Fernanda (2008), *Syntactically-Informed Representation for Sentence Selection*. Ph.D. thesis, School of Information Technology and Engineering, Faculty of Engineering, University of Ottawa.

Castro, Daniel (2007), "Improving health care: Why a dose of it may be just what the doctor ordered." *The Information Technology and Innovation Foundation*.

Cebul, Randall D., Thomas E. Love, Anil K. Jain, and Christopher J. Hebert (2011), "Electronic health records and quality of diabetes care." *New England Journal of Medicine*, 365, 825–833, URL `http://www.nejm.org/doi/full/10.1056/NEJMsa1102519`.

Chapman, WW, D Chu, and JN Dowling (2007), "Context: an algorithm for identifying contextual features from clinical text." In *ACL-07 workshop on Biological, translational, and clinical language processing (BioNLP-07)*, 81–88.

Chapman, WW, M Fizman, BE Chapman, and PJ Haug (2001), "A comparison of classification algorithms to automatically identify chest x-ray reports that support pneumonia." *Journal of Biomedical Informatics*, 34, 4–14.

Chen, Elizabeth S., George Hripcsak, Hua Xu, Marianthi Markatou, and Carol Friedman (2008), "Automated acquisition of disease-drug knowledge from biomedical and clinical documents: An initial study." *Journal of the American Medical Informatics Association*, 15, 87 – 98.

Childs, LC, RJ Taylor, and L Simonsen (2009), "Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data." *Journal of American Medical Informatics Association*, 16, 571–575.

Cohen, A and W. Hersch (2005), "A survey of current work in biomedical text mining." *Briefings in Bioinformatics*, 6, 57–71.

Cohen, AM and WR Hersh (2006), "The TREC 2004 genomics track categorization task: classifying full text biomedical documents." *Journal of biomedical discoveries and collaboration*, 1.

Cohen, AM, WR Hersh, and RT Bhupatiraju (2004), "Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage." In *Thirteenth Text Retrieval Conference TREC*, Gaithersburg, MD: National Institute of Standards and Technology.

Cohen, AM, WR Hersh, K Peterson, and PY Yen (2006), "Reducing workload in systematic review preparation using automated citation classification." *Journal of the American Medical Informatics Assoc*, 13, 206–219.

Cohen, J (1968), "Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit." *Psychological Bulletin*, 70, 213–220, URL `http://www.ncbi.nlm.nih.gov/pubmed/19586159`.

Colas, F. and P. Brazdil (2006), "Comparison of SVM and some older classification algorithms." In *Text Classification Tasks: Artificial Intelligence in Theory and Practice*, 169–178, Springer, Boston.

Craven, M. (1999), "Learning to extract relations from Medline." In *AAAI-99 Workshop on Machine Learning for Information Extraction*, 25–30.

Culotta, Aron and Jeffrey Sorensen (2004), "Dependency tree kernels for relation extraction." In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 423, Association for Computational Linguistics, Morristown, NJ, USA.

de Bruijn, Berry, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu (2010), "Nrc at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features."

Demner-Fushman, Dina, Wendy W. Chapman, and Clement J. McDonald (2009), "Methodological review: What can natural language processing do for clinical decision support?" *Journal of Biomedical Informatics*, 42, 760–772.

DeShazo, JP and AM Turner (2008), "Hands-on NLP: An interactive and user-centered system to classify discharge summaries for obesity and related co-morbidities." In *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.

Dietterich, T. (1997), "Machine-learning research: Four current directions." *Artificial Intelligence Magazine*, 18, 97–136.

Divita, Guy, Qing Z. Treitler, Doug. Redd, Adi Gundlapalli, and Sasikiran Kandula (2010), "Automatically detecting concepts, assertions, and relations in clinical text documents."

Donaldson, I., J. Martin, B. de Bruijn, C. Wolting, V. Lay, and B. Tuekam (2003), "Prebind and textomy: Mining the biomedical literature for protein-protein interactions using a support vector machine." *BMC Bioinformatics*, 4, 11–24.

Drummond, C. and RC. Holte (2003), "C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling." In *ICML'2003 Workshop on Learning from Imbalanced Datasets II*.

Elkin, P.L., D. Froehling, D. Wahner-Roedler, B. Trusko, G. Welsh, and H. Ma (2008), "Nlp-based identification of pneumonia cases from free-text radiological reports." 172–176.

Erhardt, Ramón A-A., Reinhard Schneider, and Christian Blaschke (2006), "Status of text-mining techniques applied to biomedical text." *Drug Discovery Today*, 11, 315–325.

Feldman, R., Y. Regev, M. Finkelstein-Landau, E. Hurvitz, and B. Kogan (2002), "Mining biomedical literature using information extraction." *Current Drug Discovery*, 19–23.

Forman, G. (2002), "Choose your words carefully: An empirical study of feature selection metrics for text classification." In *Proceedings of the 13th European Conference on Machine Learning and the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, 150–162, Springer-Verlag.

Frank, E. and R.R. Bouckaert (2006), "Bayes for text classification with unbalanced classes." In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 503–510, Springer, Berlin.

Friedlin, J. and C.J.A. McDonald (2006), "Natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports." In *AMIA Annual Symposium Proccidings*, 269–273.

Friedman, C., P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky (2001), "Genies: A natural language processing system for the extraction of molecular pathways from journal articles." *Bioinformatics*, 17, 74–82.

Friedman, C, L Shagina, Y Lussier, and G Hripcsak (2004), "Automated encoding of clinical documents based on natural language processing." volume 11, 392–402.

Frunza, Oana and Diana Inkpen (2008a), "Representation and classification techniques for clinical data focused on obesity and its co-morbidities." In *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data.*

Frunza, Oana and Diana Inkpen (2008b), "Textual information in predicting functional properties of the genes." In *BioNLP '08: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, 110–111, Association for Computational Linguistics.

Frunza, Oana and Diana Inkpen (2010a), "Extraction of disease-treatment semantic relations from biomedical sentences." In *BioNLP Workshop in conjunction with ACL-2010*.

Frunza, Oana and Diana Inkpen (2010b), "Identifying and classifying semantic relations between medical concepts in clinical data (i2b2 challenge)." In *i2b2 challenge Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.

Frunza, Oana and Diana Inkpen (2011), "Extracting relations between diseases, treatments, and tests from clinical data." *The 24th Canadian Conference on Artificial Intelligence (Canadian AI)*.

Frunza, Oana and Diana Inkpen (2012), "Natural language processing and machine learning techniques help achieve a better medical practice." In *Medical Applications of Intelligent Data Analysis: Research advancements* (J. Guerrero J. Gmez-Sanchis R. Magdalena, E. Soria and A.J. Serrano, eds.), IGI Global.

Frunza, Oana, Diana Inkpen, and Stan Matwin (2010), "Building systematic reviews using automatic text classification techniques." In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 303–311, Beijing, China.

Frunza, Oana, Diana Inkpen, Stan Matwin, William Klement, and Peter O'Blenis (2011a), "Exploiting the systematic review protocol for classification of medical abstracts." *Artificial Intelligence in Medicine*, 51, 17–25.

Frunza, Oana, Diana Inkpen, and Thomas Tran (2011b), "A machine learning approach for identifying disease-treatment relations in short texts." *IEEE Transactions on Knowledge and Data Engineering*, 23, 801–814.

Gaizauskas, R., G. Demetriou, P.J. Artymiuk, and P. Willett (2003), "Protein structures

and information extraction from biological texts: the pasta system." *Bioinformatics*, 19, 135–143.

Ginsberg, J., M. H. Mohebbi, S. P. Rajan, Lynnette Brammer, Mark S. Smolinski, and L. Brilliant (2009), "Detecting influenza epidemics using search engine query data." *Nature*, 457, 1012–1014.

Giuliano, Claudio, Alberto Lavelli, and Lorenza Romano (2006), "Exploiting shallow linguistic information for relation extraction from biomedical literature." In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 401–409.

Glasziou, Paul, Amanda Burls, and Ruth Gilbert (2008), "Evidence based medicine and the medical curriculum." *Brithis Medical Journal*, 337, 704–705.

Goadrich, M., L. Oliphant, and J. Shavlik (2004), "Learning ensembles of first-order clauses for recall-precision curves: A case study in biomedical information extraction." In *Proceedings of the Fourteenth International Conference on Inductive Logic Programming*, 98–115.

Groopman, Jerome (2007), *How Doctors Think*. Houghton Mifflin Company, Boston, New York.

Grouin, Cyril, Asma Ben Abacha, Delphine Bernhard, Bruno Cartoni, Louise Delger, Brigitte Grau, Anne-Laure Ligozat, Anne-Lyse Minard, Sophie Rosset, and Pierre Zweigenbaum (2010), "Caramba: Concept, assertion, and relation annotation using machine-learning based approaches."

Halgrim, Scott, Fei Xia, Imre Solti, Eithon Cadag, and Özlem Uzuner (2010), "Extracting medication information from discharge summaries." In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, 61–67, Association for Computational Linguistics.

Haynes, R.B., N. Wilczynski, K.A. McKibbon, C.J. Walker, and J.C. Sinclair (1994), "Developing optimal search strategies for detecting clinically sound studies in medline." *Journal of the American Medical Informatics Association*, 1, 447–458.

Hersh, William, Chris Buckley, T. J. Leone, and David Hickam (1994), "Ohsumed: an interactive retrieval evaluation and new large test collection for research." In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, 192–201, Springer-Verlag New York, Inc., New York, NY, USA, URL `http://dl.acm.org/citation.cfm?id=188490.188557`.

Hersh, William, Aaron Cohen, Lynn Ruslen, and Phoebe Roberts (2007), "Trec 2007 genomics track overview." In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*.

Hirschman, Lynette, Alexander Yeh, Christian Blaschke, and Alfonso Valencia (2005), "Overview of biocreative: critical assessment of information extraction for biology." *Bioinformatics*, 6.

Hripcsak, G., C. Friedman, P.O. Alderson, W. DuMouchel, S. B. Johnson, and P.D. Clayton (1995), "Unlocking clinical data from narrative reports: a study of natural language processing." *Annals of Internal Medicine*, 122, 681–689.

Hunter, L. and K. B. Cohen (2006), "Language processing: What's beyond PubMed?" *Molecular Cell*, 21, 589–594.

Hunter, L., Z Lu, J. Firby, WA Jr. Baumgartner, HL. Johnson, PV. Ogren, and KB. Cohen (2008), "Opendmap: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression." *BMC Bioinformatics*, 9, 78–89.

Ignizio, James P. (1991), *An Introduction To Expert Systems*. Mcgraw-Hill College.

Jenssen, T.K., A. Laegreid, J. Komorowski, and E. Hovig (2001), "A literature network of human genes for high-throughput analysis of gene expression." *Nature Genetics*, 28, 21–28.

Jiexun, Li, Zhang Zhu, Li Xin, and Chen Hsinchun (2008), "Kernel-based learning for biomedical relation extraction." *JASIST*, 59, 756–769.

Joachims, Thorsten (1998), "Text categorization with support vector machines: Learning with many relevant features."

Kim, J.-D., T. Ohta, Y. Tateisi, and J. Tsujii (2003), "Genia corpusa semantically annotated corpus for bio-textmining." *Bioinformatics*, 19, i180–i182, URL `http://bioinformatics.oxfordjournals.org/content/19/suppl_1/i180.abstract`.

Kim, Jin-Dong, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii (2009a), "Overview of bionlp'09 shared task on event extraction." In *Proceedings of the Workshop on BioNLP: Shared Task*.

Kim, Jin-Dong, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii (2009b), "Overview of bionlp'09 shared task on event extraction." In *Proceedings of the Workshop on BioNLP: Shared Task*, Boulder, Colorado.

Kohavi, R and F. Provost (1998), "Glossary of terms." *Machine Learning*, 30, 271–274.

Kouznetsov, A., S. Matwin, D. Inkpen, A. Razavi, O. Frunza, M. Sehatkar., and L. Seaward (2009), "Classifying biomedical abstracts using committees of classifiers and collective ranking techniques." In *22th Canadian Conference on Artificial Intelligence - Canadian AI*, 224–228.

Kowalczyk, Adam and Bhavani Raskutti (2002), "One class svm for yeast regulation prediction." 4, 99–100.

Krogel, Mark A, Marcus Denecke, Marco Landwehr, and Tobias Scheffer (2002), "Combining data and text mining techniques for yeast gene regulation prediction: a case study." *ACM SIGKDD Explorations Newsletter*, 4, 104–105.

Leroy, G., H.C. Chen, and J.D. Martinez (2003), "A shallow parser based on closed-class words to capture relations in biomedical text." *Journal of Biomedical Informatics*, 36, 145–158.

Liu, Hongfang, Yves A. Lussierb, and Carol Friedman (2001), "Disambiguating ambiguous biomedical terms in biomedical narrative text: An unsupervised method." *Journal of Biomedical Informatics*, 34, 249–261.

Ma, Yimin (2007), *classification on imbalanced data: application to systematic reviews automation*. Master's thesis, University of Ottawa.

Mandraekar, J.N. and S.J. Mandraekar (2005), "Statistical methods in diagnostic medicine using sas software." 30th SAS User Group International Conference.

Matwin, Stan, Alexandre Kouznetsov, Diana Inkpen, Oana Frunza, and Peter OBlenis (2010), "A new algorithm for reducing the workload of experts in performing systematic reviews." *Journal of the American Medical Informatics Association (JAMIA)*, 17, 446–453.

Matwin, Stan, Alexandre Kouznetsov, Diana Inkpen, Oana Frunza, and Peter OBlenis (2011), "Performance of svm and bayesian classifiers on the systematic review classification task." *Journal of the American Medical Informatics Association (JAMIA)*, 18, 104–105.

Meystre, S. M., G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle (2008), "Extracting information from textual documents in the electronic health record: a review of recent research." *Yearbook of medical informatics*, 128–144, URL `http://view.ncbi.nlm.nih.gov/pubmed/18660887`.

Mitchell, Tom M. (1997), *Machine learning*. McGraw-Hill.

Mitsumori, T., M. Murata, Y. Fukuda, K. Doi, and H. Doi (2006), "Extracting protein-protein interaction information from biomedical text with svm." *IEICE Transactions on Information and Systems*, 89, 2464–2466.

Miyao, Yusuke, Kenji Sagae, Rune Sætre, Takuya Matsuzaki, and Jun'ichi Tsujii (2009), "Evaluating contributions of natural language parsers to protein-protein interaction extraction." *Bioinformatics*, 25, 394–400.

Moore, C., J. Wisnivesky, S. Williams, and T. McGinn (2003), "Medical errors related to discontinuity of care from an inpatient to an outpatient setting." *Journal of General Internal Medicine*, 18, 646651.

Ng, S.K. and M. Wong (1999), "Toward routine automatic pathway discovery from on-line scientific text abstracts." In *Genome Information Ser. Workshop Genome Information*, 104–112.

Novichkova, S., S. Egorov, and N. Daraselia (2003), "Medscan, a natural language processing engine for medline abstracts." *Bioinformatics*, 19, 1699–1706.

Ould, M., F. Caropreso, P. Manine, C. Nedellec, and S. Matwin (2003), "Sentence Categorization in Genomics Bibliography: a Naïve Bayes Approach." In *Informatique pour l'analyse du transcriptome*.

Özgür, Arzucan, Thuy Vu, Günes Erkan, and Dragomir R. Radev (2008), "Identifying gene-disease associations using centrality on a literature mined gene-interaction network." *Bioinformatics*, 24, 277–285.

Patrick, J and P Asgari (2008), "A brief summary about the approach and explanation of the attributes of the developed system." In *i2b2 challenge Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.

Patrick, Jon and Min Li (2009), "A cascade approach to extract medication event (i2b2 challenge 2009)." In *Presentation at the Third i2b2 Workshop*.

Patrick, Jon, Yitao Zhang, and Yefeng Wang (2007), "Developing feature types for classifying clinical notes." In *BioNLP '07: Proceedings of the Workshop on BioNLP 2007*, 191–192, Association for Computational Linguistics, Morristown, NJ, USA.

Penz, JF, AB Wilcox, and JF Hurdle (2007), "Automated identification of adverse events related to central venous catheters." *Journal of American Medical Informatics Association*, 174–182.

Pestian, John P., Christopher Brew, Pawel Matykiewicz, D. J. Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch (2007), "A shared task involving multi-label classification of clinical free text." In *BioNLP '07: Proceedings of the Workshop on BioNLP 2007*, 97–104, Association for Computational Linguistics, Morristown, NJ, USA, URL `http://portal.acm.org/citation.cfm?id=1572392.1572411`.

President's Council of Advisors on Science and Technology (2008), "Priorities for personalized medicine." Technical report.

Pustejovsky, J., J. Castaño, J. Zhang, M. Kotecki, and B. Cochran (2002), "Robust relational parsing over biomedical literature: Extracting inhibit relations." *Pacific Symposium on Biocomputing*, 7, 362–373.

Quinlan, J. R. (1986), "Induction on decision trees." *Machine Learning*, 1, 81–106.

Ray, S. and M. Craven (2001), "Representing sentence structure in hidden markov models for information extraction." In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, 1273–1279, Morgan Kaufmann.

Regev, Y., M. Finkelstein-Landau, and R. Feldman (2002), "Rule-based extraction of experimental evidence in the biomedical domain: The kdd cup 2002 (task 1)." 4, 90–92.

Rennie, Jason D. M., Lawrence Shih, Jaime Teevan, and David R. Karger (2003), "Tackling the poor assumptions of naive bayes text classifiers." In *In Proceedings of the Twentieth International Conference on Machine Learning*, 616–623.

Rindflesch, T.C., L. Tanabe, J.N. Weinstein, and L. Hunter (2000), "Edgar: Extraction of drugs, genes, and relations from the biomedical literature." *Pacific Symposium on Biocomputing*, 5, 514–525.

Roberts, Angus, Robert Gaizauskas, and Mark Hepple (2008), "Extracting clinical relationships from patient narratives." In *BioNLP '08: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, 10–18, Association for Computational Linguistics, Morristown, NJ, USA.

Roberts, Kirk, Bryan Rink, and Sanda Harabagiu (2010), "Extraction of medical concepts, assertions, and relations from discharge summaries for the fourth i2b2/va shared task."

Rosario, B. (2005), *Extraction of semantic relations from bioscience text*. Ph.D. thesis, Ph.D. dissertation, Berkeley, CA, USA, 2005.

Rosario, Barbara and Marti A. Hearst (2004), "Classifying semantic relations in bioscience texts." In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 430–438, Association for Computational Linguistics, Morristown, NJ, USA.

Sackett, David L, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson (1996), "Evidence based medicine: what it is and what it isn't." *BMJ*, 312, 71–72.

Sasaki, Yutaka, Yoshimasa Tsuruoka, John McNaught, and Sophia Ananiadou (2008), "How to make the most of ne dictionaries in statistical ner." *BMC Bioinformatics*, 9, S5, URL `http://www.biomedcentral.com/1471-2105/9/S11/S5`.

Scharfe, C, HH Lu, J Neuenburg, EA Allen, GC Li, T Klopstock, T Cowan, GM Enns, and RW Davis (2009), "Mapping gene associations in human mitochondria using clinical disease phenotypes." *Public Library of Science Computational Biology*, 5.

Sibanda, T. (2006), *Was the patient cured? Understanding semantic categories and their relationships in patient records. Master's thesis.* Ph.D. thesis, Massachusetts Institute of Technology: Department of Electrical Engineering and Computer Science.

Smith, Barry, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenber, Susanna-Assunta Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel, and Suzanna Lewis (2007), "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." *Nature Biotechnology*, 25, 1251–1255, URL `http://dx.doi.org/10.1038/nbt1346`.

Smith, L., T. Rindflesch, and W. J. Wilbur (2004), "Medpost: a part-of-speech tagger for biomedical text." *Bioinformatics*, 20, 2320–2321, URL `http://bioinformatics.oxfordjournals.org/content/20/14/2320.abstract`.

Soderland, S., D. Fisher, J. Aseltine, and W. Lehnert (1995), "Crystal: Inducing a conceptual dictionary." In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1314–1319.

Solt, I., D. Tikk, V. Gal, and Z.T. Kardkovacs (2009), "Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier." *Journal of the American Medical Informatics Association*, 16, 580–584.

Solt, Illés, Ferenc P. Szidarovszky, and Domonkos Tikk (2010), "Concept, assertion and relation extraction at the 2010 i2b2 relation extraction challenge using parsing information and dictionaries." In *Fourth i2b2/VA Shared-Task and Workshop Challenges in Natural Language Processing for Clinical Data* (Özlem Uzuner, ed.), 1–6,

i2b2 National Center for Biomedical Computing, i2b2 National Center for Biomedical Computing, Washington, DC, URL `https://www.i2b2.org/NLP/Relations/`.

Spear, Brian B, Margo Heath-Chiozzi, and Jeffrey Huff (2001), "Clinical application of pharmacogenetics." *Trends in Molecular Medicine*, 7, 201 – 204, URL `http://www.sciencedirect.com/science/article/pii/S1471491401019864`.

Srinivasan, P. and T. Rindflesch (2002), "Exploring text mining from medline." In *Proceedings of the AMIA Symposium*, 722–726.

Stapley, B.J. and G. Benoit (2000), "Bibliometrics: Information retrieval visualization from co-occurrences of gene names in medline abstracts." *Pacific Symposium on Biocomputing*, 5, 526–537.

Stephens, M., M. Palakal, S. Mukhopadhyay, and R. Raje (2001), "Detecting gene relations from medline abstracts." In *Paccific Symposium of Biocomputation*, 483–496.

Sutton, Charles and Andrew Mccallum (2006), *Introduction to Conditional Random Fields for Relational Learning*. MIT Press.

Sven, Mika and Rost Burkhard (2004), "Nlprot: extracting protein names and sequences from papers." *Nucleic Acids Research*, 32, W634–W637, URL `http://nar.oxfordjournals.org/content/32/suppl_2/W634.abstract`.

Swensen, Stephen J., Gregg S. Meyer, Eugene C. Nelson, Gordon C. Hunt, David B. Pryor, Jed I. Weissberg, Gary S. Kaplan, Jennifer Daley, Gary R. Yates, Mark R. Chassin, Brent C. James, and Donald M. Berwick (2010), "Cottage industry to postindustrial care - the revolution in health care delivery." *New England Journal of Medicine*, 362, e12, URL `http://www.nejm.org/doi/full/10.1056/NEJMp0911199`.

Szarvas, G, R Farkas, and A Almási (2009), "Semi-automated construction of decision rules to predict morbidities from clinical texts." *Journal of American Medical Informatics Association*, 16, 601–605.

Thomas, J., D. Milward, C. Ouzounis, S. Pulman, and M. Carroll (2000), "Automatic extraction of protein interactions from scientific abstracts." 5, 538–549.

Tsuruoka, Yoshimasa (2005), "Bidirectional inference with the easiest-first strategy for tagging sequence data." In *In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 467–474.

Tsuruoka, Yoshimasa, Jun'ichi Tsujii, and Sophia Ananiadou (2008), "Accelerating the annotation of sparse named entities by dynamic sentence selection." *BMC Bioinformatics*, 9, S8, URL `http://www.biomedcentral.com/1471-2105/9/S11/S8`.

UMLS (????), "Unified medical language system fact sheet." National Library of Medicine. http://www.nlm.nih.gov/pubs/factsheets/umls.html.

Uzuner, O, I Goldstein, Y Luo, and I Kohane (2008), "Identifying patient smoking status from medical discharge records." *Journal of American Medical Informatics Association*, 15, 14–24.

Uzuner, O., X. Zhang, and T. Sibanda (2009), "Machine learning and rule-based approaches to assertion classification." *Journal of the American Medical Informatics Association*, 16, 109–115.

Vogel, David S. and Randy C. Axelrod (2002), "Predicting the effects of gene deletion." 4, 101–103.

Ware, H, CJ Mullet, and V Jagannathan (2009), "Natural language processing framework to assess clinical conditions." *Journal of American Medical Informatics Association*, 16, 585–589.

Widdows, Dominic and Kathleen Ferraro (2008), "Semantic vectors: a scalable open source package and online technology management application." In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)* (Bente Maegaard-Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias Nicoletta Calzolari

(Conference Chair), Khalid Choukri, ed.), European Language Resources Association (ELRA), Marrakech, Morocco. Http://www.lrec-conf.org/proceedings/lrec2008/.

Xu, Rong, Alex Morgan, Amar K Das, and Alan Garber (2009), "Investigation of unsupervised pattern learning techniques for bootstrap construction of a medical treatment lexicon." In *Proceedings of the Workshop on BioNLP*, 63–70.

Yakushiji, A., Y. Tateisi, Y. Miyao, and J. Tsujii (2001), "Event extraction from biomedical papers using a full parser." 6, 408–419.

Yang, H, I Spasic, JA Keane, and G Nenadic (2009), "A text mining approach to the prediction of a disease status from clinical discharge summaries." *Journal of American Medical Informatics Association*, 16, 596–600.

Yang, Y and X Liu (1999), "A re-examination of text categorization methods." In *Proceedings of the ACM SIGIR Conference on Research and Development in Inf Retrieval*, 42–49.

Yang, Y. and J. Pedersen (1997), "A comparative study on feature selection in text categorization." In *The Fourteenth International Conference on Machine Learning*, 412–420.

Yeh, AS, L Hirschman, and AA Morgan (2003), "The evaluation of text data mining for database curation: lessons learned from the kdd challenge cup." *Bioinformatics*, i331–i339.

Zweigenbaum, P, D Demner-Fushman, H Yu, and KB Cohen (2007), "Frontiers of biomedical text mining: current progress." In *Brief Bioinformatics*, 358–75.

Appendix A

Example of Clinical Data

Example of discharge summary

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<report>
<checksum>20070322DS-Q15Mz+lRVivl-848-385376871</checksum>
<subtype>TXS DISCHARGE</subtype>
<type>DS</type>
<chief_complaint>COUGH</chief_complaint>
<admit_diagnosis>786.2</admit_diagnosis>
<discharge_diagnosis>428.0,486,996.81,280.9,272.4,346.90,</discharge_diagnosis>
<year>2007</year>
<downlaod_time>2009-08-18</downlaod_time>
<updatetime/>
<deid>v.6.22.07.0</deid>
<report_text>[Report de-identified (Safe-harbor compliant) by De-ID v.6.22.07.0]



**INSTITUTION
TRANSPLANT SURGERY
DISCHARGE SUMMARY
PATIENT NAME:  **NAME[AAA, BBB M]
ACCOUNT #:  **ID-NUM
**ROOM
ATTENDING PHYSICIAN:  **NAME[YYY M ZZZ]
ADMISSION DATE:  **DATE[Mar 19 2007]
DISCHARGE DATE:  **DATE[Mar 22 2007]
```

```
This is a patient of Dr. **NAME[YYY ZZZ].
DISCHARGE DIAGNOSES:
1. Cadaveric renal transplant in **DATE[Mar 2006] for congenital renal dysplasia.
2. Chronic rejection on rapamycin.
3. Iron-deficiency anemia.
4. Rapamycin pneumonitis.
CONSULTS:
1. Renal Medicine.
2. Pulmonary Medicine.
3. Transplant Infectious Disease.
PROCEDURES:
1. Chest x-ray on **DATE[Mar 19 2007] showing patchy airspace disease on the
right side greater than the left suggesting pneumonia with underlying volume
overload and cardiomegaly.
2. CT scan of the chest on **DATE[Mar 19 2007] showing diffuse interstitial
disease with small areas of consolidated airspace disease likely due to
atypical pneumonia with pulmonary edema or medication side effects was
likely.  Also, an abdominal CT scan was secondary hyperparathyroidism and
hypoattenuating lesions in the liver some of which was due to focal fat, as
well as possible gallstones or gallbladder sludge.
3. Blood cultures x2 and urine culture on **DATE[Mar 20 2007] which found no
growth at 2 days at the time of discharge.
4. Urinary Legionella antigen which was negative.
5. Repeat chest x-ray on **DATE[Mar 22 2007] showed little interval change from
the previous one on **DATE[Mar 19 2007].
HISTORY OF PRESENT ILLNESS:  The patient is a **AGE[in 30s]-year-old female
with past medical history significant for congenital renal dysplasia status post
cadaveric renal transplant in **DATE[Mar 2006].  She has a history of at least 2
documented episodes of chronic rejection in **DATE[May 2006] and
**DATE[Jul 2006] for which she was initially placed on CellCept and then changed
to Rapamune when she had an uncontrollable diarrhea on the CellCept.
She also had IV Solu-Medrol at the time of her rejections.
She presented to the Emergency Department with symptoms of cough and hemoptysis
for the past month.  She states the symptoms __________ was before Christmas
of the past year.  She felt congestion of her chest as well as cough.
She had clear expectoration but recently this had specks of blood.
She denied any fevers or night sweats.
Her nephrologist from home had told her that she would probably need
intravenous iron therapy and thought that her shortness of breath might be
due to anemia.  She had a chest x-ray done on an outpatient basis which
showed a lesion suspicious for tuberculosis and hence was referred to
**INSTITUTION for further workup and management.
```

PAST MEDICAL HISTORY:
1. Cadaveric renal transplant in **DATE[Mar 2006] as above.
2. Chronic rejection x2.
3. Left arm AV fistula.
4. Iron-deficiency anemia.
5. Hypercholesterolemia.
6. Migraine.
MEDICATIONS: At home include:
1. Bactrim.
2. Aspirin.
3. Lasix.
4. Toprol XL.
5. Tacrolimus.
6. Rapamycin.
7. Sodium bicarbonate.
8. Magnesium gluconate.
9. Oral contraceptive pill.
10. Depakote.
PHYSICAL EXAMINATION: The patient was afebrile with stable vital signs. Her blood pressure was slightly elevated at 162/100. She was saturating 98% on room air. Her physical exam was positive for pallor on eye exam. Her chest exam showed clear breath sounds over the right, middle, and lower lobes. Her belly was soft and nontender and her lower extremity were without edema.
At this time, it was felt she should be admitted to 12-North for respiratory isolation, oxygen as needed, and antibiotic management. She was started on cefepime and Zithromax for community-acquired pneumonia and a CT scan of the chest was obtained. The results are as noted above. Pulmonary Medicine was consulted to see her and recommended that she undergo diuresis. They felt that the most likely cause is an atypical microorganism and thought that the cefepime was not necessary and that the Zithromax should be sufficient to control her pneumonia. Additionally, she had rapamycin level drawn that was 21 which is above the upper limit of normal and her rapamycin was withheld, given the thought that this might be rapamycin pneumonitis. She was given a couple of doses of IV Lasix and diuresed briskly in response to this. B-type natriuretic peptide on the time of admission was 1571 further supporting that she may have been somewhat volume overloaded. The Pulmonary consult team did consider that she should undergo transthoracic echocardiogram; however, this can be scheduled on an outpatient basis. They did feel that TB was an unlikely cause of her atypical pneumonia and did take her off the isolation precautions. She never was able to bring up a sputum sample that was sufficient for culture during her hospital stay. They thought another potential likely cause would be CMV and the PCR for this was negative

however. They also recommended to check Legionella urine antigen and that was also negative. Their final recommendation was that she undergo a CT scan of the chest repeated when she was at her driveway. They did not feel that she had any immediate indication for bronchoscopy or bronchial alveolar lavage at the time of her admission.
Renal Medicine was also consulted to see her for her chronic kidney disease as well as her iron-deficiency anemia. An iron level was sent which came back at 27 with a lower limit of normal at 65. She had a ferritin level which was 110, which was within normal limits. They did say that she would require intravenous iron replacement and that this could be done on an outpatient basis so as not to hold up her discharge. They also recommended a PTH level, which turned about to be about 689.
By hospital day #3, she was feeling quite a bit better. Her cough was improving and she had no further dyspnea. At this time, I felt she was stable for discharge to her home. At the time of discharge, she was awake and alert. Her heart was regular and her lungs were clear. Her abdomen was soft, nontender, and nondistended. Her previous incisions were all well healed. Her legs were without edema and her neurologic exam was without deficit.
DISCHARGE INSTRUCTIONS: The patient may be discharged to her home with a diagnosis of atypical pneumonia as well as possible rapamycin pneumonitis.
1. Diet: Low potassium, low phosphorus.
2. Activity: Should be increased as tolerated.
3. ALLERGIES: ARE TO IRON DEXTRAN.
4. She should call her transplant coordinator for any fever, increased cough, or shortness of breath.
5. Labs: She should have a BNP, magnesium, phosphorus, CBC with diff, and FK level on her return to **INSTITUTION.
6. She should have an outpatient transthoracic echocardiogram.
7. She should follow up with **INSTITUTION on Friday **DATE[Mar 26 2007] and she will call for this appointment.
8. She may follow up with her home nephrologist in the next 2 to 3 weeks.
DISCHARGE MEDICATIONS: She should resume the following home medications:
1. Prograf 1 mg p.o. b.i.d.
2. Bactrim Single Strength 1 tablet p.o. Monday, Wednesday, and Friday.
3. Toprol XL 50 mg p.o. daily.
4. Sodium bicarbonate 1950 mg p.o. b.i.d.
5. Magnesium gluconate 1 g p.o. b.i.d.
6. Depakote 500 mg p.o. nightly.
7. Triphasil oral contraceptive pill 1 tablet p.o. daily.
8. Aspirin 81 mg p.o. daily.
9. She should not resume her Rapamune.

```
10. Increase her Lasix to 40 mg p.o. b.i.d. x3 days, then resume 20 mg p.o.
daily.
11. Aranesp 20,000 units subcutaneously every 2 weeks.
She should take the following new medications:
1. Nephrocaps 1 p.o. daily.
2. Zithromax 250 mg p.o. daily x3 more days for her last dose to be on
**DATE[Mar 25 2007].
3. Venofer 300 mg IV x1, this is to be given on her return to the
**INSTITUTION on **DATE[Mar 26 2007].


-----------------------------------
**NAME[YYY M ZZZ]

Dictator:  **NAME[WWW XXX], M.D.
TL/sr
D:  **DATE[Mar 22 2007] 14:43:08
T:  **DATE[Mar 23 2007] 04:39:18
R:  **DATE[Mar 23 2007] 04:39:18/sr
Job ID:  356698/**ID-NUM
Cc:

*** Dictated By: -**NAME[XXX, VVV] ***
Electronically Signed by **NAME[YYY ZZZ]  **DATE[Mar 24 2007] 11:16:16 AM

</report_text>
</report>
```

Example of progress notes

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<report>
<checksum>20071131PGN-6m5pUjiD52vk-848-1554778106</checksum>
<subtype>INFDIS ATTEND</subtype>
<type>PGN</type>
<chief\_complaint>OPEN WOUND RIGHT ANKLE</chief\_complaint>
<admit\_diagnosis>682.6</admit\_diagnosis>
<discharge\_diagnosis>996.67,682.6,041.11,E878.1,V09.0,401.9,272.4,250.00,
730.17,</discharge\_diagnosis>
<year>2007</year>
<downlaod\_time>2009-10-05</downlaod\_time>
```

```
<updatetime/>
<deid>v.6.22.08.0</deid>
<report\_text>[Report de-identified (Safe-harbor compliant) by De-ID v.6.22.08.0]



**INSTITUTION
INFECTIOUS DISEASE
ATTENDING PHYSICIAN PROGRESS NOTE
PATIENT NAME:  **NAME[AAA, BBB M]
ACCOUNT \#:  **ID-NUM
**ROOM
ATTENDING PHYSICIAN:  **NAME[YYY M ZZZ]
ADMISSION DATE:  **DATE[Nov 17 2007]
EVALUATION DATE:  **DATE[Nov 31 07]
Mrs. **NAME[AAA] is doing well today.  She denies any fever, chills, or
diarrhea.  She is getting increased sensation in her right lower extremity
although she still has some anesthesia across her lower abdomen.  She remains
afebrile at 98.5.  Her lungs are clear.  Her abdomen is soft and nontender.
The incision there is grossly clean.  The JP drain has serous fluid.  Her
right lower extremity sensation, movement, and capillary refill are all
intact.  Her white cell count is 5.1 and platelets are 383.
IMPRESSION:
MRSA wound infection and osteomyelitis.  She is doing well clinically and is
tolerating the linezolid.  I would like to continue linezolid 600 mg p.o.
b.i.d. until **DATE[Dec 28 2007].  Please check a CBC with platelets two times
per week and a sedimentation rate and C-reactive protein weekly.  These
results can be faxed to my office at **PHONE.
----------------------------------
**NAME[WWW XXX], M.D.
Assistant Professor of **NAME[CCC]
Dictator:  **NAME[WWW XXX], M.D.
**INITIALS
D:  **DATE[Nov 31 2007] 12:24:58
T:  **DATE[Nov 31 2007] 14:47:54
R:  **DATE[Nov 31 2007] 14:47:54/ksc
Job ID:  369472/**ID-NUM

*** Dictated By: -**NAME[XXX, VVV] ***
Electronically Signed by **NAME[WWW XXX]  **DATE[Dec 12 2007] 06:01:54 PM
</report\_text>
</report>
```

Example of radiology reports

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<report>
<checksum>20071131PGN-6m5pUjiD52vk-848-1554778106</checksum>
<subtype>INFDIS ATTEND</subtype>
<type>PGN</type>
<chief_complaint>OPEN WOUND RIGHT ANKLE</chief_complaint>
<admit\_diagnosis>786.50</admit_diagnosis>
<discharge\_diagnosis>493.22,414.00,V45.81,V45.73,327.23,401.9,272.4,
724.5,311,250.00,</discharge_diagnosis>
<year>2007</year>
<downlaod\_time>2009-08-18</downlaod_time>
<updatetime/>
<deid>v.6.22.07.0</deid>
<report\_text>[Report de-identified (Safe-harbor compliant) by De-ID v.6.22.07.0]

**INSTITUTION
History \&amp; Physical
Name: **NAME[AAA, BBB]
Acct \#: **ID-NUM
MRN: **ID-NUM
Admitted: **DATE[Jul 18 2007]
Dict: **NAME[XXX, WWW]
Attend: **NAME[ZZZ, YYY]
CHIEF COMPLAINT:  Shortness of breath and cough.
HISTORY OF PRESENT ILLNESS:  The patient is a **AGE[in 50s]-year-old female with
a past medical history significant for COPD and CAD, status post coronary artery
bypass graft surgery who presents from home with a several day history of
cough and shortness of breath.  The patient had been on a short course of
oral prednisone without relief.  The patient also complains of some
left-sided chest pain as she believes it is related to her cough.  She
received one sublingual nitroglycerin without relief in the ER.  She states
that her cough is productive of clear sputum.  She denies any substernal
chest pain at this time or any other associated symptoms.
ALLERGIES:  VASOTEC AND ETODOLAC.
CURRENT MEDICATIONS:
1.  Coreg 6.25 mg twice a day.
2.  Ecotrin 325 mg daily.
3.  Flonase 2 sprays each nostril daily.
4.  Flovent inhaler.
5.  Vicodin 7.5/750, 2 to 3 times a day as needed.
```

6. Insulin 70/30, 30 units in the morning, 25 units at night.
7. Lasix 20 mg daily.
8. Lipitor 80 mg daily.
9. Lorazepam 0.5 mg as needed.
10. OxyContin 40 mg twice a day.
11. Oxygen at bedtime.
12. Proventil 2 puffs twice a day as needed.
13. Quinapril 20 mg daily.
14. Senna 2 tablets daily.
15. Zoloft 100 mg daily.
16. Trazodone 300 mg at bedtime.
17. Stool softener twice a day.

PAST MEDICAL HISTORY: Significant for coronary artery disease, status post coronary artery bypass graft surgery x3 vessel disease in 1997, nephrectomy in 1967 secondary to cyst removal, and then destruction of the kidney, obstructive sleep apnea, hypertension, insulin dependent diabetes mellitus, chronic back and leg pain, hyperlipidemia, depression, and COPD.

SOCIAL HISTORY: The patient lives in Southside by herself. She quit tobacco 4 years ago, but had smoked at least 1 pack per day since the age of **AGE[in teens].

She has an occasional beverage.

FAMILY HISTORY: Coronary artery disease in her mother and brother. Mother died in her 70s of an MI. Brother died in his 50s of an MI. Father with diabetes mellitus.

REVIEW OF SYSTEMS: In general, the patient denies fevers, chills, or night sweats. Denies fatigue or lightheadedness. HEENT: No headache. No change in vision or hearing. Cardiovascular: No substernal chest pain. She does have left-sided chest pain which she believes is related to her cough. Pulmonary: Shortness of breath as described, exertional dyspnea, and cough productive of a clear to brownish sputum. GI: No nausea, vomiting, diarrhea. Positive for constipation. She is on chronic narcotic pain medications. GU: No dysuria or pyuria. Musculoskeletal: No new muscle aches or pain. She does have chronic leg pain. Extremities: Slight increase in lower extremity edema. Neurologically: No deficits described. Her rest of the review of systems was negative.

PHYSICAL EXAMINATION: In general, the patient is awake, alert, and oriented. She is not in any apparent distress. She answers questions appropriately. Vital signs showed a temperature of 36.6, heart rate 81, respirations 20, oxygen saturation 95% on 2 L, and blood pressure 113/64. HEENT: Mucous membranes are moist without evidence of erythema or exudate. Pupils are equal, round, and reactive to light. Extraocular muscles are intact. Sclerae are nonicteric. Neck: Impossible to assess for JVD secondary to

body habitus. No carotid bruits. Cardiovascular: Regular rate and rhythm with distant heart sounds without murmur, rub, or gallop. Pulmonary shows bilateral wheeze throughout both lung fields. GI: Morbidly obese, soft, nontender, nondistended with positive bowel sounds. Extremities shows trace lower extremity edema. Neurologically. The patient is awake, alert, and oriented. Cranial nerves 2 through 12 appear to be normal. Sensation is intact throughout. She is able to move all 4 limbs against gravity.
Chest x-ray shows minimal left basilar atelectasis. No focal areas of consolidation. CBC with differential shows white blood cell count of 6.8, hemoglobin 11.2, hematocrit 34.4, and platelets 260. Neutrophils 54, lymphocytes 29, and monocytes 13. Troponin I is less than 0.10. BNP is 30. Basic Metabolic Panel: Sodium 142, potassium 4.2, chloride 100, bicarbonate 30, BUN 17, creatinine 1.2, and glucose 102. EKG shows sinus rhythm without ST change. Urine is leukocyte esterase negative, otherwise negative.
ASSESSMENT AND PLAN: The patient is a **AGE[in 50s]-year-old former heavy smoker with a history of chronic obstructive pulmonary disease, also coronary artery disease, status post coronary artery bypass graft surgery who presents with shortness of breath, wheeze, and left-sided chest pain.
1. Pulmonary: Chronic obstructive pulmonary disease exacerbation. O2 to maintain saturations, drug aerosol treatments, IV Solu-Medrol 40 mg IV q.12 h. as well as IV antibiotics with azithromycin 500 mg IV q. day and Rocephin 1 g IV q. day. We will also check her sputum culture.
2. Cardiovascular: History of coronary artery disease. Chest pain is reproducible with palpations, did not respond to p.o. nitroglycerin. Rule out myocardial infarction with serial troponin I and keep on the monitor. Continue beta blocker, aspirin, and Lipitor.
3. Diabetes mellitus. Continue home 70/30 and Regular insulin sliding scale with q.i.d. Accu-Cheks.
4. Chronic pain. Continue home medications.
5. Gastrointestinal and deep venous thrombosis prophylaxis.
Call placed for covering physician for Dr. **NAME[ZZZ], no call back at this time.

**NAME[WWW XXX], MD
HS Job \# 816063 / 34606 /
SHY \# **ID-NUM
D: **DATE[Jul 18 2007] 21:22
T: **DATE[Jul 19 2007] 05:21
**CARBON-COPY
</report_text>
</report>