LINKING CLINICAL RECORDS TO THE BIOMEDICAL LITERATURE

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2016

By

Noha Alnazzawi

School of Computer Science

Contents

Word Count: 59887

List of Tables

List of Figures

Abstract

Narrative information in Electronic Health Records (EHRs) contains a wealth of clinical information about treatments, diagnosis, medication and family history. In addition, the scientific literature represents a rich source of information that summarises the latest results and new research findings relevant to different diseases. These two textual sources often contain different types of valuable phenotypic information that may be complementary to each other. Combining details from each source thus has the potential to be useful in uncovering new disease-phenotypic associations. In turn, these associations can help to identify patients with high risk factors, and they can be useful in developing solutions to control the causes responsible for the development of different diseases. However, clinicians at the point of care have limited time to review the large volume of potentially useful information that is locked away in unstructured text format. This in turn limits the utility of this "raw" information to clinical practitioners and computerised applications. Accordingly, the provision of automated and efficient means to extract, combine and present phenotype information that may be scattered amongst a large number of different textual sources in an easily digestible format is a prerequisite to the effective use and comprehensive understanding of details contained within both the records and the literature. The development of such facilities can in turn help in deriving information about disease correlations and supporting clinical decisions.

This thesis is the first comprehensive study focussing on extracting and integrating phenotypic information from two different biomedical sources using Text Mining (TM) techniques. In this research, we describe our work on (1) extracting phenotypic information from both EHRs and the biomedical literature; (2) extracting the relations between phenotypic information and distilling them from EHRs using an event-based approach; and (3) using normalisation methods to link the phenotypic information found in EHRs with associated mentions found in the literature as a first step towards the automatic integration of information from these heterogeneous sources.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

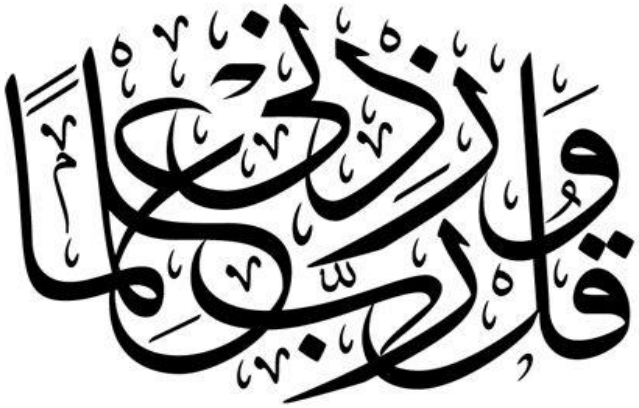
I. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including or administrative purposes.

II. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

III. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

IV. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's policy on presentation of Theses.

Dedication



“Say: My Lord, increase me in knowledge”

IN THE NAME OF ALLAH

And with His blessing

The All-Knowing, The Most-Wise

This work is dedicated to the memory of my late father Abdulkareem Alnazzawi. I am totally indebted to the tremendous inspiration he gave me throughout his life.

Additionally, I want to dedicate this thesis to my lovely mother Khadija. I always enjoyed her care and guidance. This thesis is also dedicated to my lovely son (Ibrahim) and to my brothers and sisters.

Acknowledgements

By coming to the end of this scientific journey; first, all praises are due to ALLAH for his merciful guidance throughout my life and during my stay in Manchester.

I am deeply indebted (quite literally) to my supervisor Prof. Sophia Ananiadou not only for the supervision, research guidance and constructive criticism she has provided me as her student, but most especially for her understanding and encouragement during more challenging times.

My thanks are also extended to William Black for his invaluable help, support in explaining the CAFETIERE system.

I would like to express my gratitude and deepest respect to my colleagues at the National Centre for Text Mining (NaCTeM). Special mention goes to Claudiu, George, Matt, Paul and Riza, who have constantly given me their encouragement and who have all been like family to me.

In particular, I would like to especially thank my collaborator and mentor, Paul, who has generously provided me with his invaluable research input and support. I would like also to show my sincere appreciation to Ms. Riza, for all help and assistance she offered to me throughout the years.

My genuine appreciations go to Jubail University College and Royal commission for Jubail and Yanbu in Saudi Arabia for funding and supporting my study at the University of Manchester.

It definitely would not have been even possible for me to pursue this research without my family. I have no words to express how grateful I am to my ever-supportive mother, brothers and sisters (Maha and Nada) who encouraged me to pursue my dreams. Words cannot describe how thankful I am to Mohammed and Ibrahim for their tremendous patience and understanding, and for sharing this entire journey with me.

Publications

Part of the work presented in this thesis has been published in the following peer-reviewed conference and journal articles:

1. Alnazzawi, N., Thompson, P., and Ananiadou , S. Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature. in *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL*. 2014.
2. Alnazzawi, N., Thompson, P., and Ananiadou , S. Using text mining techniques to extract phenotypic information from the PhenoCHF corpus. BMC Medical Informatics and Decision Making, 2015. **15**(Suppl 2): p. S3.
3. Mihaila, C., Batista-Navarro, R., Alnazzawi, N., Kontonatsios, G., Korkontzelos, I., Rak, R., Thompson, P., Ananiadou, S. Mining the Biomedical Literature. *Healthcare Data Analytics,* 2015. **36**: p. 251.
4. Alnazzawi, N. and S. Ananiadou.Using text mining techniques to extract relations between phenotype information in electronic health records and the literature. *Poster presented at the 8th Saudi Conference*. Imperial College London. 2015.
5. Alnazzawi, N., Thompson, P., and Ananiadou., S. *Mapping Phenotypic Information in Heterogeneous Textual Sources to a Domain-Specific Terminological Resource.* Submitted to PLOS ONE journal.

Abbreviations

AMBIT	Acquiring Medical and Biological Information from Text
BeCAS	Biomedical Concept Annotation System
BIO	Begin-Inside-Outside
BioNLP	Biomedical Natural Language Processing
BMI	Body Mass Index
BOW	Bag Of Words
BRAT	Brat Rapid Annotation Tool
CAFETIERE	Conceptual Annotations for Facts, Events, Terms, Individual Entities and Relations
CDSS	Clinical Decision Support System
CG	Cancer Genetics
CHF	Congestive Heart Failure
CKD	Chronic Kidney Disease
CLEF	Clinical E-Science Framework
COPD	Chronic Obstructive Pulmonary Disease
CRAFT	The Colorado Richly Annotated Full Text Corpus
CRFs	Conditional Random Fields
cTAKES	Clinical Text Analysis and Knowledge Extraction System
CUIs	Concept Unique Identifiers
EE	Event Extraction
EHRs	Electronic Health Records
EPI	Post-translational modifications
GATE	General Architecture for Text Engineering
GDep	GENIA Dependency
GE	GENIA Event
HD	Heart Disease
HIPAA	Health Insurance Portability and Accountability
HITEx	Health Information Text Extraction
HMMs	Hidden Markov Models
HPO	Human Phenotype Ontology

ICD-9	International Classification of Diseases version 9
ID	Infectious Disease
IE	Information Extraction
IR	Information Retrieval
IT	Information Technology
LBD	Literature-Based Discovery
MEDLINE	Medical Literature Analysis and Retrieval System Online
MedLEE	Medical Language Extraction and Encoding System
MEMM	Maximum Entropy Markov Model
MeSH	Medical Subject Headings
MIMIC II	Multiparameter Intelligent Monitoring in Intensive Care
ML	Machine Learning
NER	Named Entity Recognition
NLM	National Library of Medicine
NLP	Natural Language Processing
ODIE	Development and Information Extraction
OMIM	Online Mendelian Inheritance in Man
ONC	National Coordinator
PAS	predicate-argument structure
PASTA	Protein Active Site Template Acquisition
PATO	Phenotype and Trait Ontology
PC	Pathway Curation
PHI	Personal Health Information
POS	parts-of-speech
PPIs	Protein-Protein Interactions
RE	Relation Extraction
SNOMED CT	Systematized Nomenclature of Medicine-Clinical Terms
SSVMs	Structural Support Vector Machines
TF-IDF	Term Frequency-Inverse Document Frequency
TM	Text Mining
UBERON	Uber Anatomy Ontology
UIMA	Unstructured Information Management Architecture
UMLS	Unified Medical Language System

WC	Waist Circumference
WHO	World Health Organisation

Chapter 1 Introduction

Advances in health care are dependent upon the use, integration and organisation of massive amounts of genomic, phenotypic, pharmacological, biological and clinical information. The ability to rapidly access and integrate information gathered by biomedical scientists from multiple fields constitutes the first step towards keeping researchers aware of the latest advances and innovations in their field of interest [1].

Human phenotypic information constitutes the observable traits of human beings (e.g., height, eye colour, etc.) resulting from genetic make-up and environmental influences. A more contemporary definition of phenotypes includes the measurable biological, behavioural or cognitive markers that distinguish individuals with specific medical conditions from the general population [2]. More precisely in this research phenotypic information refers to the causes, risk factors, signs or symptoms of a given disease.

A greater understanding of phenotype-disease correlations is essential to improve disease prevention measures. Such an understanding can help to allow a better evaluation of an individual's risk of developing a disease, and can assist in controlling the disease's causes and risk factors [3]. Acquiring a detailed knowledge of phenotypic information for complex diseases such as cancer, asthma and cardiovascular disease can help to provide patients with personalised medical treatments, by allowing an individual's therapeutic response to be determined and by offering better treatment for patients, based on their phenotypic profile [3, 4].

The primary source of biomedical scientific information which describes new findings and results of experimental studies is text [5]. Various biomedical information is published in different biomedical resources including the biomedical literature (e.g., original reports and summaries in journals, books etc.), biological databases (e.g., annotations in genes, protein, disease databases), web pages and EHRs (e.g., clinical narrative reports).

In this introductory chapter we provide an overview of the primary textual sources for phenotypic information i.e., biomedical literature and EHRs. We also provide an overview of the motivation for this research. This is followed by an outline of the research aims, hypotheses, objectives and contributions.

1.1 Biomedical literature

The biomedical literature constitutes a major and reliable repository of knowledge, which has been constructed by thousands of scientists over decades of experimentation, analysis and discovery. Human phenotypes comprise a very important part of this knowledge [6]. The goal of biomedical research is to discover new knowledge. Clinicians use findings from research to improve their methods of disease diagnosis, treatment and prevention to deliver optimum clinical care to patients [7]. Medical Literature Analysis and Retrieval System Online (MEDLINE) is the U.S. National Library of Medicine's (NLM) database of bibliographic references and it is well known to bioinformaticians. MEDLINE currently contains over 23 million citations which focus on biomedicine [8]. This abundance of biomedical information is constantly expanding, in line with the rapid creation and publication. With such exponential growth in published information, it is extremely challenging for clinicians to keep abreast of all the new findings and discoveries within their own specialist fields. This is particularly the case, given that the majority of new knowledge is locked away in the unstructured text of scientific articles, which makes the information difficult to use and integrate with other sources (e.g. biological databases) [6, 9].

1.2 Electronic Health Records

The last decade has seen a remarkable uptake of the adoption of Information Technology (IT) in health care settings. An increase in the awareness of the utility of transforming medical records from paper into fully digitised records has resulted in a rapid shift towards the use of EHRs in healthcare settings [10, 11].

EHRs are written by clinicians and describe patients' personal, social and medical histories. EHRs contain structured (name-value pairs) and unstructured (narrative) information. Unstructured narrative texts are written using natural language. Examples include: progress notes, discharge summaries as well as test reports from radiology, electroencephalography (EEG) and pathology. In contrast to structured data, narrative reports allow clinicians to freely express information about a patient's conditions, without constraints of space or vocabulary. They can explain why drugs

were given or discontinued, describe the results of physical examinations or provide other information important for patient care [12]. Furthermore narrative text usually includes valuable predictive information, which is usually not available in structured parts of EHRs [13]. Examples of information found in free text are family history, risk factors and signs or symptoms (e.g., ejection fraction is a strong indicator of congestive heart failure) [14]. This kind of clinical data constitutes a rich source of information that is used in the formulation of appropriate healthcare plans. For example, it allows clinicians to compare the characteristics of patients with similar medical histories. Such information is also useful for clinical research. For example, combining and comparing information from different clinical records in large repositories can help researchers to find answers to questions such as: "How many patients with stage 2 adenocarcinoma who were treated with tamoxifen were symptom-free after 5 years?". Such information forms the basis for generating new hypotheses, which can subsequently be explored and validated in clinical trials [12].

1.3 Problem definition

Clinicians at the point of care face an information overload problem when dealing with the overwhelming volume of narrative information in EHRs and searching for relevant information in the biomedical literature. The vast amount of information locked away within biomedical literature repositories can make it virtually impossible to make connections between pieces of biomedical knowledge dispersed within many different articles. However, such connections are vital to facilitate advances in clinical care [15]. As a consequence, there is evidence of a 13−17 year gap between the time when research findings are reported in the literature and when they are put into practice in the context of clinical care [16]. Encoding the clinical information contained within free text resources in a structured format increases the feasibility of developing a wide range of clinical applications that will become invaluable tools for clinicians and researchers [7].

According to the above, there has been a surge of interest in developing methods that can aid clinicians in making more efficient use of existing biomedical knowledge and in helping them to make practical usage of this knowledge. While manual curation and indexing can assist clinicians in searching for and locating

appropriate literature, MEDLINE indexing and Medical Subject Headings (MeSH) vocabulary cannot represent all the concepts of interest for the clinicians, mainly because MeSH is a manually curated resource that is not expressive enough to capture all biomedical concepts [7]. Furthermore, the full-text of the biomedical literature contains a wealth of information that is neither mentioned in abstracts nor encoded in MeSH terms, and thus may not be captured by curators [7]. This can mean that a large amount of relevant information is overlooked by clinical researchers.

In response, there has been a large amount of research in the Natural Language Processing (NLP) field to develop systems that analyse biomedical free text, in order to extract and structure relevant information [17]. This collaboration between the NLP and biomedical communities forms a research area known as Biomedical Natural Language Processing (BioNLP). NLP tools can be combined together into pipelines to carry out different text mining (TM) tasks, which involve the processes of discovering and extracting knowledge from unstructured textual data. TM comprises three major tasks: 1) Information Retrieval (IR), to collect relevant documents; 2) Information Extraction (IE), to extract key information from these documents and convert it into structured knowledge; and 3) data mining, to find associations between entities extracted from different texts, and thus aid in the discovery of new knowledge [17, 18]. TM systems help clinicians to manage the vast amount of information available in the literature. Well-established information retrieval techniques allow the search space to be reduced by an initial query. TM methods move a step further by subsequently applying IE techniques to identify and structure specific types of information contained *within* these documents [6, 19]. IE includes tasks such as Named Entity Recognition (NER), which corresponds to the automatic recognition and semantic categorisation of named entities, such as gene, protein, disease and drug names in free text, and Relation Extraction (RE), which attempts to extract binary associations between named entities, for example, recognising Protein-Protein Interactions (PPIs) and the relations between drugs, genes and cells. Recently, there has been a shift in biomedical IE from identifying binary relations, to the more ambitious task of Event Extraction (EE), which deals with extracting and semantically categorising complex *n*-ary and nested relations, such as gene expression and regulation and protein binding [20].

Over the past few years, biomedical TM has seen dramatic progress in the development of increasingly complex high-performance NLP techniques that are able to better address the information needs of biomedical researchers than traditional IR techniques [17]. A major driving force behind the progression of the field has been the organisation of shared tasks. Some of these tasks have focussed on processing biomedical literature, while others have been concerned with the extraction of information from medical records. As a result, current BioNLP tools and resources can be broadly categorised according to whether they handle biomedical or clinical text. The focus of biomedical shared tasks has been the extraction of named entities such as gene, protein, diseases and drugs names, classifying binary relations between two entities (e.g., extracting the relations between genes and diseases) and extracting complex *n*-ary relations or events. These shared tasks facilitate the benchmarking of methods proposed by different research groups. As a result of these shared tasks, a large number of robust IE tools have been developed, e.g., NER tools such as ABNER [21] and NEMine [22], relation extractors such as AkanePPI [23] and OpenDMAP [24] and event extractors such as the Turku Event Extraction System [25] and EventMine [26]. This is because PPIs and other molecular-level relationships are a central theme of modern translational and genomic research, which are frequently described in the biomedical literature. Thus the automatic extraction of such information is a priority for biomedical TM researchers.

However, clinical shared tasks focussing on IE have appeared more recently and have been less numerous than the biomedical shared tasks. The first clinically focussed IE shared task took place in 2006, as part of the Information Biology and Bedside (i2b2) [27] challenge. This was followed by ShARe/CLEF in 2013 [28] and SemEval in 2014 [29]. Most of the available NLP tools and resources for the clinical domain have been developed as a result of these shared tasks. Examples of these resources include corpora and systems to determine different health problems at the document level (e.g. smoking status [30], obesity and its co-morbidities [31], etc.), recognising disease names in clinical notes and normalising them by mapping each mention to a clinical concept in Unified Medical Language System (UMLS), a large scale terminological resource of biomedical terminology [28], and identifying/classifying relations between clinical concepts (e.g., between medical problems, treatments and tests) [32].

Less attention has been given to the application of TM techniques to recognise phenotypic information. This is mainly because there is no available annotated corpus for phenotypes. Furthermore, there is no comprehensive dictionary covering phenotype names. For example, although the UMLS Metathesaurus [33] is an extremely comprehensive biomedical resource, which integrates more than 100 terminologies and ontologies, it does not contain a semantic type that corresponds directly to phenotypic information. Other resources, such as the Online Mendelian Inheritance in Man (OMIM) [34] and Human Phenotype Ontology (HPO) [35] are more specifically focussed on phenotypes. However, they are manually constructed which makes them difficult to update and maintain. Although HPO is specialised for disease phenotypes, it only covers a subset of human diseases. Furthermore, its coverage of synonyms is not exhaustive. For example, it includes *endocrine abnormality*, but it does not include the synonym *endocrine disorder*. Additionally, phenotypic information is highly expressive. There are often different ways to describe the same phenotypic concept. For example, adjectives and other modifiers are added to phenotype names to give more information about them (e.g., *right ventricular enlargement* is in the HPO, but an automatic system would not suggest that *mild to moderate right ventricular enlargement* is a phenotype simply by searching in the HPO). However, HPO is not sufficiently comprehensive to cover all of these expressions [36].

Narrative information in EHRs and literature articles often include detailed phenotypic information for specific diseases. Since the use of TM tools has been previously shown to provide an efficient automated means to extract and integrate vital information hidden within the vast volumes of biomedical text, it is important to develop tools and resources that can better support the application of TM techniques to extract phenotypic information for textual resources. Extracting phenotypic information from heterogeneous sources, i.e., EHRs and literature articles constitutes a first step towards the automatic integration of complementary information. This integration can help to improve healthcare applications, including Clinical Decision Support Systems (CDSS) and Evidence-based Medicine.

Our research focusses specifically on Congestive Heart Failure (CHF). We chose CHF as it is a life-threatening disease and according to the World Health Organisation (WHO), cardiovascular diseases represent the highest cause of death globally [37, 38]. In the United Kingdom, for example, about one in six men and one

in ten women die from heart disease [39]. Having access to detailed CHF phenotype information can identify low and high-risk CHF patients, save the life of many others and advance patient recruitment for clinical trials and case control studies. Our research focus is additionally motivated by the fact that more than one third of patients with chronic kidney disease (CKD) develop symptoms of heart failure. CHF is also a common contributor to the progression of CKD. Thus, a vicious circle exists between these two diseases [40]. Therefore, renal failure or renal insufficiency may be more than a marker for heart failure severity and instead may play a causative role in the progression of heart failure. Understanding and managing the interaction between these two diseases is an evolving challenge for clinicians [41].

1.4 Research Aims, Hypotheses and Objectives

1.4.1 Aims

Our overall research aims are defined as follows:

A1 To extract phenotypic entities from heterogeneous biomedical sources (EHRs and literature articles).

A2 To extract *n*-ary relations between phenotypic entities in EHRs.

A3 To study the variability between the two biomedical text types.

A4 To investigate challenges arising from the integration of phenotypic information from the literature and EHRs.

1.4.2 Hypotheses

In line with the four fundamental research aims, we formulate the following hypotheses:

H1 Existing text mining techniques can be adapted to extract phenotypic information from the overwhelming volume of information in the literature and EHRs and to discover hidden knowledge and associations that may occur across texts of different types.

H2 *N*-ary relations between phenotypic entities can be cast as events, and they can be extracted using an event-based approach.

H3 Normalising various types of phenotypic information that appear in both EHRs and literature articles can act as a first step towards the automatic integration of knowledge that is dispersed within these two text types.

1.4.3 Objectives

Based on the proposed hypotheses, we establish 6 research objectives:

O1 To conduct a comprehensive review of existing resources, annotated corpora and approaches for clinical NER.

O2 To apply NER techniques at a large scale to extract phenotypic information from both EHRs and literature articles.

O3 To conduct a comprehensive review of existing corpora and approaches for clinical relation extraction.

O4 To adapt TM tools currently used to extract relations and events from full papers and abstracts and make them suitable for extracting the relations between phenotypic entities in EHRs.

O5 To develop a novel method to normalise phenotypic concept mentions from heterogeneous textual sources (i.e., EHRs and literature articles) and to map them to UMLS concepts.

O6 To integrate information extracted from both text types using the normalisation approach.

1.5 Contributions of this thesis

The contributions of this research are summarised in the following points:

C1 Extraction of phenotypic information from narrative text within EHRs, integrated with information extracted from the vast amounts of available literature.

C2 Construction of a gold standard, semantically annotated corpus (PhenoCHF) for the purposes of training and testing various TM techniques on clinical texts dealing with the CHF disease. The corpus consists of texts drawn from two sources: discharge summaries from EHRs and scientific articles. The annotation in PhenoCHF is concerned with the identification of phenotype information, i.e. entities and relationships between them.

C3 Adaptation of machine learning NER algorithms to the clinical domain to extract phenotypic information from EHRs and the literature.

C4 Adaptation of an event extraction algorithm to the clinical domain to extract relations between phenotypic entities.

C5 Development of normalisation methods to link the phenotypic information found in EHRs with associated mentions found in the literature.

Chapter 2 Literature review

Clinical knowledge is growing constantly as new discoveries are made. Advancements in health care are dependent on the use, integration and organisation of massive amounts of genomic, phenotypic, pharmacological and clinical information. This important information is usually expressed as free text within a number of sources, including articles from the scientific literature and clinical narrative reports form EHRs [1]. The high rate of publication in the biomedical literature has made it impossible for researchers and clinicians to keep abreast of the new findings on their own field of interest, since there are simply too many articles to read. However, knowledge about different disease characteristics from disparate textual sources can be automatically extracted and integrated into a structured format (e.g. a database), which can help to facilitate a greater understanding of the various characteristics of diseases (e.g. treatment and symptoms) that may be overlooked in unseen literature articles. Having ready access to up-to-date information about diseases (e.g., detailed disease profiles) can be valuable for a variety of applications, including decision support (e.g., recommending treatments), quality assurance (e.g., inter- and intra-institutional review), clinical information needs (e.g., answering clinical questions), information retrieval (e.g., finding relevant documents), and data mining (e.g., hypothesis discovery).

In this chapter, we present a review of the state-of-the-art in clinical IE. Since the focus of this research is on IE methods for documents from heterogeneous biomedical sources (i.e., biomedical literature and clinical records), we review notable resources and techniques that have been applied to both text types.

2.1 Biomedical text mining

Biomedical TM accelerates knowledge discovery by automatically extracting knowledge that is hidden in text and presenting this knowledge to researchers in a concise and easily understandable format. Therefore, the results of TM analysis can provide efficient access to required facts and the associations between them [1]. Biomedical TM is partly inspired by the work of Swanson [42], in which it was

demonstrated that extracting and linking facts from different literature articles can lead to the generation of new scientific hypotheses [43].

Over the last decade, there has been an impressive increase in biomedical TM research, leading to significant advances. Research in the area includes work on producing increasingly high-performance tools that are able to carry out the fundamental IE tasks introduced in chapter 1 as well as more complex tasks such as automatic summarisation and question answering. Developing TM solutions to address such complex tasks is becoming increasingly straightforward, thanks to the development of TM workflow platforms, such as U-Compare [44], Argo [45] and General Architecture for Text Engineering (GATE), which allow heterogeneous text processing tools to be flexibly combined into different processing pipelines. Further important research outcomes include evaluation methodologies and an increasing availability of resources that are important both for developing and evaluating tools. However, a number of unsolved problems and challenges remain. These continue to present themselves to the biomedical text mining community and provide great potential for interesting research [9].

Most biomedical TM systems are broadly aimed at handling text belonging to one of two main types, i.e. biomedical scientific text or clinical text (e.g. narrative reports from EHRs) [43]. As mentioned previously in Section 1.3, the majority of IE tools that have been developed for the biomedical domain focus on identifying bio-entities (such as genes and proteins) and relationships (such as proteins and their binding sites) among them [9]. Examples of such tools include ABNER [21], BANNER [46] and the GENIA Tagger [47]. Despite the advances in the automatic processing of biomedical literature, progress on processing clinical data has been relatively slow. There are only a small number of research teams working on clinical TM, which is partly due to the greater challenges in accessing and processing this type of text compared to biomedical literature [20].

2.1.1 Challenges of processing clinical reports

EHRs are one of the most important healthcare innovations of the last decade [48]. The introduction of EHRs is particularly promising, since they facilitate increased accessibility of clinical information, which in turn can lead to improvements in clinical research [1]. EHRs are written by clinicians and describe patients' personal, social and medical histories. Information within these records has the potential to

improve health care outcomes. Possible secondary uses of the data include tracking performance of drugs, optimising resources, appraising treatments and alerting the community about potential post-marketing adverse drug effects. However, a barrier to the effective use of data from EHRs in clinical research and computerised applications is that most information in EHRs is in the form of narrative text [1]. Only by structuring the narrative clinical information it can be used to aid in the development of a wide range of clinical applications that will be invaluable for clinicians and researchers [1]. Manual encoding/structuring of all important clinical information contained within the records is infeasible, because it is costly and time consuming.

Automatically adapting and extending existing NLP techniques to identify, extract and structure relevant information in EHRs can significantly increase the potential to carry out better, novel, clinical studies [1].

A major hurdle to be overcome is the fact that information contained within EHRs is confidential. In order to make such information available for research use, personal identifying information (e.g., names, addresses, telephone numbers, etc.) must be removed from the records to comply with laws protecting patient confidentiality. The automatic detection of personal information is a difficult task that often requires manual review and sometimes even after removing personal information, it is still possible to identify patients according to rare characteristics [27]. Even after the removal of Personal Health Information (PHI), approval to access to the records must still be obtained from an Institutional Review Board and from institutional administrators.

The HIPAA (Health Insurance Portability and Accountability Act) and the European Union Data Protection Directive protect the confidentiality of patient data by requiring the consent of the patient and the approval of the Institutional Review Board in order for patient data to be used for research purposes. However, these requirements may be waived if PHI is de-identified, i.e., all personal identifying information is removed.

Another challenge is that there is no standardised format for writing clinical reports. Rather, the nature of these reports can be highly heterogeneous; the text often does not conform to grammatical conventions and can be full of domain specific idiosyncrasies, acronyms and abbreviations, as well as spelling and other typographical errors. Indeed the rate of misspelling in medical records (around 10%)

is very high compared to other text genres [43]. Furthermore, the vocabulary used in clinical reports can be very wide-ranging, while sentences can be very long and highly complex [49]. Punctuation is often missing and new lines may be used instead of full stops to indicate the end of the sentences. A further characteristic is the frequent use of highly ambiguous abbreviations, e.g. "*PE"* may refer to '*physical examination*', '*pleural effusion*' or '*pulmonary embolism*', amongst others [1]. Another example of ambiguous abbreviation is "*RA"* which may refer to '*right atrium*', '*rheumatoid arthritis'*, '*refractory anemia'*, '*renal artery'* or one of several other concepts.

2.1.2 Semantic analysis of biomedical text

Three major subtasks of information extraction are particularly relevant for processing biomedical text: 1) NER, 2) relation extraction and 3) event extraction.

Although each of these subtasks is distinct in the type of information it aims to extract, they each achieve their goals by employing similar methods, which include machine learning, statistical analysis and other techniques of NLP. Challenges and approaches to the subtasks of biomedical information extraction are discussed below.

2.1.2.1 Named entity recognition (NER)

a) Dictionary-based approaches

The first methods aimed at extracting medical concepts from clinical notes relied solely on the application of string matching techniques to administrative billing codes such as those contained within the International Classification of Diseases version 9 (ICD-9). However, ICD-9 coding is aimed specifically at billing purposes and therefore, the ICD-9 classification system cannot accurately capture the nuances of phenotypic characteristics, such as family history, signs and symptoms or known risk factors for a disease that are typically embedded in narrative text. Thus, a number of studies have shown that the use of ICD-9 coding alone has performance limitations in terms of sensitivity, and thus is not sufficient to reliably identify patients suffering with a particular disease or having specific risk factors [50-52].

Over the last two decades, there have been many efforts to apply NLP technologies to clinical text. The Linguistic String Project [53, 54] and the Medical Language Extraction and Encoding System (MedLEE) are a few of the earliest NLP

systems developed for application to text within the clinical domain. UMLS is the world's largest medical knowledge source and it has been widely used as a dictionary for the identification of medical named entities in clinical reports. Advances in natural language and semantic processing techniques have contributed towards the development of many dictionary-based methods to extract medical concepts that are typically mentioned in narrative text. Systems such as Clinical Text Analysis and Knowledge Extraction System **(**cTAKES), the Health Information Text Extraction (HITEx) system and MetaMap use a variety of NLP tools (e.g., part-of-speech POS taggers and shallow parser) to identify all the noun phrases in a given text and map them to medical concepts of various types within knowledge resources (e.g., UMLS) [55].

In the following section, we provide an overview of some of the major clinical NLP systems.

1) cTAKES

cTAKES is a natural language processing system developed at the Mayo clinic to process and extract information from the free text in EHRs [56]. cTAKES is a modular, pipelined system of NLP components, built using the OpenNLP[1] toolkit. The employment of the Unstructured Information Management Architecture (UIMA) framework [57] in constructing the pipeline ensures it can be reconfigured and/or extended with additional tools in a straightforward manner. The components within the cTAKES pipeline are specifically adapted for application to clinical texts. These components create rich linguistic and semantic annotations that can be utilised by clinical decision support systems and in clinical research. The current pipeline of components consists of a sentence boundary detector, tokeniser, normaliser, POS tagger, shallow parser, named entity recogniser, co-reference resolver, temporal relation detector, semantic role labeller and clinical question answerer.

The NER component implements a dictionary look-up algorithm within a noun phrase window. Each named entity identified through dictionary lookup is mapped to a concept in the terminology (a subset of UMLS which includes the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) and RxNorm vocabularies). Each term in the terminology belongs to one of the following semantic

[1] http://opennlp.sourceforge.net/projects.html

types: disorders/diseases (with a separate group for signs/symptoms), procedures, anatomy and drugs. Each recognised named entity has attributes including: 1) a span attribute, corresponding to the text span associated with the named entity, 2) a concept attribute, which denotes the UMLS identifier to which the named entity maps, 3) a negation attribute, indicating whether or not the named entity is negated and 4) a status attribute, corresponding to one of the following values: *current, history of, family history of, possible*. This attribute aims to capture whether the mentioned disorder is considered a present condition of the patient under discussion, whether it is found in the context of the patient's personal or family medical history, or whether there is speculation about the disorder.

However, the NER component in cTAKES does not resolve ambiguities that result from the identification of multiple terms in the same text span. A further limitation of cTAKES is its inability to correctly handle the interpretation of coordinated structures. For example, the phrase "ovarian and breast cancers" should be interpreted as "ovarian cancer" and "breast cancer". However, cTAKES currently incorrectly recognises this as "ovarian" and "breast cancer".

cTAKES has been used by many research groups to process and extract information from EHRs. For example, it was used by Savova et al. [58] to classify radiology notes as positive, negative, probable and unknown cases for peripheral artery disease. The overall accuracy of cTAKES, compared to the gold standard was 0.93. cTAKES was also extended to participate in the first and second i2b2 NLP challenges, to classify smoking status [59] and recognise obesity and its co-morbidities [59]. It achieved F-scores of 0.60 and 0.73 for the tasks of classifying smoking status and recognising obesity, respectively.

2) HITEx

HITEx was developed in response to the i2b2 project on extracting factors contributing to asthma exacerbation and hospitalisation project [60]. HITEx uses the GATE as the development platform to adapt different NLP modules which are then assembled into pipelines to carry out different tasks [61]. GATE is an open source NLP framework that contains a Collection of REusable Objects for Language Engineering (CREOLE). CREOLE consists of a set of NLP modules that perform common NLP tasks, such as tokenising, POS tagging and syntactic chunking. The GATE framework can be viewed as a backplane for plugging in CREOLE

components, and it provides various services to the components such as bootstrapping, loading and reloading, management and visualisation of data structures and process execution.

HITEx consists of the following 11 components: a section splitter, section filter (selects the subset of sections based on the selection criteria, e.g., category name, section name, etc.), sentence splitter, tokeniser, POS tagger, noun phrase finder, UMLS concept mapper, negation finder, n-gram tool and classifier (to determine the smoking status of a patient). These modules are assembled into different pipelines according to the task to be undertaken. For example, a pipeline to extract diagnoses from clinical notes is constructed through the sequential combination of the following: section splitter, section filter, sentence splitter, sentence tokeniser, POS tagger, noun phrase finder, UMLS concept mapper, and negation finder [62].

HITEx has been used to extract principal diagnoses, co-morbidities and smoking status information from discharge summaries for patients with a known history of asthma or Chronic Obstructive Pulmonary Disease (COPD) [60]. The accuracy of HITEx was 0.82 for extracting principal diagnoses, 0.87 for co-morbidities and 0.90 for smoking status**.**

3) MetaMap

MetaMap is a general biomedical NLP system developed by Aronson et al. [63] at NLM. Although MetaMap was originally developed to map entities of interest (e.g., diseases, drugs, etc.) mentioned in the biomedical literature to concepts in UMLS [64, 65], it has also been widely used by many researchers to extract information from clinical text within EHRs. For example, Meystre and Haug [66] evaluated the ability of MetaMap to extract medical problems from EHRs and recorded a recall of 0.74 and a precision of 0.76.

MetaMap uses a minimal commitment parser, the UMLS SPECIALIST lexicon and a POS tagger, all developed at the NLM. Firstly, MetaMap finds all noun phrases and for each phrase, a set of lexical variants is generated. The candidate set of all UMLS terms containing at least one of the variants is retrieved. Each candidate UMLS term is assigned a score that is a measure of how strongly the actual term is mapped to the UMLS vocabulary. Complete mappings are constructed by combining candidates involved in disjoint parts of the phrase, and the strength of the complete mappings is computed just as for candidate mappings. The complete mapping with

the highest score represents the best MetaMap match for the original phrase. The process also incorporates a word-sense disambiguation mechanism, recently enhanced with a statistical context-sensitive method.

4) MedLEE

MedLEE was developed by Friedman et al. [67] at Columbia University, initially to process and encode information in radiology reports. The system was subsequently expanded to handle and process information in reports from a variety of clinical sub-domains, including pathology reports and discharge summaries. It has been applied to many clinical information extraction problems and has showed promising results. Examples of the diverse range of tasks that it has been used to address include adverse event detection, automated trend discovery and acquisition of disease-drug associations.

MedLEE consists of several modular components with different functions (i.e., pre-processor, parser, phrase regulariser and encoder) [68]. The pre-processor is the first component. It separates complete reports into sentences, and then identifies phrases within sentences. Subsequently, it assigns words and multiword phrases to semantic categories. The second component is the parser, which uses grammar rules that combine semantic and syntactic patterns to recognise relevant clinical information and modifier information (i.e., negation, temporal information, family history etc.), and to generate target forms. For example, the output of the parser for the sentence "enlarged heart noted" would be [problem, enlarged heart, [certainty, 'high certainty']]. The phrase regulariser formulates a multiword phrase after the parsing stage if the sentence contains a non-contiguous phrase (e.g., "heart was enlarged"). For example, "enlarged heart" is defined in the lexicon as a phrase. If the words in the phrase are non-contiguous, the output of the parser would be composed of the individual words and not the complete phrase (i.e., [problem, enlarged, [bodyloc, heart], [certainty, 'high certainty']]). The purpose of the phrase regulariser is thus to "reconstruct" non-contiguous phrases into single units. This is achieved by using a compositional table containing the phrases and their corresponding compositional structures, such that both contiguous and non-contiguous phrases have the same final structure (i.e., [problem, enlarged heart, [certainty, 'high certainty']]).

The flexible design of MedLEE means that it is straightforward to adapt for use in different domains by changing the underlying knowledge sources and grammar

patterns. For example, MedLEE has been adapted to extract phenotypic information (e.g., cellular body functions and model organism anatomy) from 300 randomly chosen journal articles from the biomedical literature by changing the semantic and syntactic grammar. On this task, the system achieved 64% precision and 77.1% recall [69]. The GENIES system, which extracts biomolecular interactions from the biomedical literature was also adapted from MedLEE, by replacing the lexicon with one relevant for the biological domain and by utilising a new set of grammar patterns [70].

Presented in Table 2.1 is a summary of the features of the clinical dictionary-based systems discussed above.

A primary problem with the use of purely dictionary-based methods arises from the fact that it is impossible to cover the names of *all* clinical entities in a single resource, or even in a combination of resources, due to the frequent appearance in text of new clinical concepts and/or new variant forms describing existing concepts. Although the more generalised patterns encoded by rules can recognise a potentially wider range of entities, formulating such patterns by hand is very time-consuming, and it would still be impossible for a hand-crafted set of rules to account for all possible entity-indicating patterns that may occur in text [71].

Another weakness, particularly of dictionary-based approaches, is their inability to handle ambiguous surface forms. The ambiguity between domain-specific concepts and general English words can lead to the identification of many false positives, especially if the context of terms is not taken into account. For example, the abbreviation "*BE*" for "*Bacterial Endocarditis*", could be confused with the verb "be" [72]. A further example is the abbreviation "*HD*" could refer to various different conditions, e.g., '*Heart Disease*', '*Hansen Disease*', '*Hodgkin disease*', '*Huntington Disease*', or it may refer to the temporal expression '*Hospital Day*' [73].

A further drawback of dictionary-based approaches is that they cannot account for the fact that, especially within clinical records, certain entity types, such as sign or symptoms, may include long descriptive phrases, e.g. "*subtle decrease flow signal within the sylvian branches",* may incorporate modifiers, e.g. "*normal blood pressure*" or may be expressed using various structures, e.g. clauses, rather than the more typical noun phrases, e.g. "*blood pressure was normal*" [73].

Table 2.1 Comparison of Clinical NLP Systems

System	Document Type	Knowledge source	Specific Features	Availability
MedLEE [68]	EHRs	MedLEE lexicon +UMLS	Abbreviation resolution, Disambiguation component	currently unavailable
cTAKES [56]	EHRs	Subset of UMLS (i.e. SNOMED CT and RxNORM)	Family history and negation detection	public
MetaMap[63]	Biomedical literature + EHRs	UMLS	Disambiguation component, Negation detection	public
HITEx [60]	EHRs	UMLS	Family history and negation detection	licensed

b) Rule-based approaches

In order to overcome the problems of low recall that are inherent in a purely dictionary-based approach, a possible solution is to couple basic dictionary lookup with other methods, such as hand-crafted rules. Compared to the purely dictionary-based approach, rules can be advantageous in that they can recognise false negatives [55], can account for context to help to reduce ambiguity problems, and can make it easier to recognise entities having different structures or incorporating descriptive information. However, a drawback of rule-based methods is that they frequently over-fit the corpora used for their development, because the linguistic patterns encoded in the rules are often highly sensitive to the corpus features. Thus, while the rules may work well on documents within this development corpus, they are likely to perform poorly when applied to other corpora or different text types. Since the construction of rules is time consuming and requires significant human effort, they do not constitute an ideal approach for NER [74].

Examples of rule-based approaches include the Acquiring Medical and Biological Information from Text (AMBIT) system [75], which is adapted from the Protein Active Site Template Acquisition (PASTA) system [76] to extract diseases, drugs and genes. AMBIT has been used to mine radiology reports for lung cancer signs (e.g., *mass, collapse*), locations in the lung (e.g., *upper lobe, basal region*) and the relationships between these signs and locations. AMBIT was evaluated on a gold

standard of 83 radiology reports to extract lung cancer signs and locations. It achieved a precision and recall of 0.69 and 0.83, respectively [75].

Another example of a purely rule-based system was developed by Yang [77]. The rules made use of several different linguistic features (e.g., lexical, orthographic, morphological) to extract medication-related information (e.g., drug names, dosages, modes of administration, frequencies, durations, reasons). After terms had been extracted using the rules, a further set of context-based rules was employed to filter out terms that were not directly related to the medications of the patient under discussion (e.g., by detecting negated contexts and information relating to family members). Their system was evaluated on 547 discharge summaries, and obtained an encouraging performance (a micro-averaged F-score of 0.80 [77, 78]).

Childs et al. [79] used ClinREAD, a proprietary healthcare-domain oriented, rule-based NLP system (Lockheed Martin, Bethesda, Maryland, USA) to recognise clinical records mentioning obesity and its co-morbidities. The rules consist of patterns generated by medical experts, which make use of combinations of different types of keywords (e.g. disease names, their synonyms and symptoms). They weighted and combined the evidence for each class of each disease. The system achieved a micro-averaged F-score of 0.95 [31, 79].

c) Machine learning approaches

The rapid growth in the biomedical literature makes the process of NER more difficult compared to other domains. Clinical text also poses a challenge for NER techniques that have been designed to operate on formal styles of text, largely due to the fact that clinical records often employ a more informal linguistic style [80]. On one hand the above issues mean that dictionary-based methods are frequently inadequate to recognise the wide-ranging and sometimes novel variants of concepts that may appear in both the biomedical literature and in clinical text. On the other hand, hand-crafted rules are usually highly sensitive to the features of the text to which they are to be applied. Accordingly, it is difficult to create sets of rules that can be applied robustly to the recognition of concepts in clinical texts, whose linguistic features can vary considerably.

Recent interest has moved away from dictionary and rule-based methods to machine learning algorithms, such as statistical or Machine Learning (ML) techniques. ML techniques recognise instances of specific classes of interest by

learning automatically from annotated corpora. They exploit the distinctive features associated with positive and negative examples to automatically recognise entities in previously unseen text [81].

The performance of ML methods is highly dependent on the features employed, e.g., orthographic, morphological, lexical and semantic features. It has been established in previous work that the selection of the most appropriate features is equally as important as choosing the best algorithm [82].

Supervised ML methods learn how to recognise entities of different types through evidence in annotated corpora. Such corpora consist of samples of documents in which the target information (e.g., entities of interest) has been marked-up manually by domain experts. ML algorithms are able to use this annotated evidence to learn patterns related to different types of entities and their contexts. Once trained, the ML systems can predict the occurrence of entities in previously unseen texts. Supervised ML methods have become popular, owing to the encouraging levels of performance that they have demonstrated. Most previous work on developing ML-based NERs has utilised either hidden Markov models (HMMs) [83, 84], support vector machines (SVMs) [85], or conditional random fields (CRFs) [46, 86, 87], the latter of which has been shown to achieve particularly reliable performance when applied to texts in the biomedical domain.

de Bruijn et al. [84] developed a discriminative semi-Markov HMM using high-dimensional bags of features, derived from both the text and external sources. From the training data, a combination of features was extracted, including token features (e.g., word shape, character n-gram), context features (e.g., token features from four tokens before to four tokens after the word), sentence features (e.g., sentence length) as well as syntactic and semantic features obtained from external tools and sources (i.e., UMLS and cTAKES). They observed that the utilisation of external sources to generate semantic and syntactic features is beneficial, and improves the F-score by 1.5 percentage points (i.e., from 0.836 to 0.852).

Tang et al. [88] developed the Structural Support Vector Machines (SSVMs) model to extract disorder entities from clinical records. Their model is based on a wide range of features generated from both training texts and external knowledge sources (e.g., bag-of-words, POS, type of notes, section information and semantic categories from UMLS). The ML model achieved an F-score of 0.750.

Pathak et al. [87] proposed a CRF-based model to extract clinical concepts (i.e., problems, tests and treatments) from discharge summaries and progress notes. The model utilised domain knowledge features (i.e., standard section headers such as *chief complaints*, *past history*, *lab data*, *medicines* and *current diagnosis*), morphological features, orthographic features (e.g., capitalisation) and linguistic features (e.g., POS, chunks, NP head). They observed that using domain knowledge in the form of features is very effective towards achieving high performance (an F-score of 0.84).

The drawback of supervised ML methods, however, is the need for manually-annotated corpora which can be expensive and time-consuming to produce.

d) Hybrid approaches

Some proposed approaches to clinical NER are based on a combination of different methods [20].

Jiang et al. [89] proposed a hybrid clinical entity extraction system by combining an ML-based named entity recogniser with post-processing rules. Their system utilised a CRF model with various features including: orthographic information (e.g, prefixes and suffixes), syntactic (e.g., POS tags), lexical and semantic information obtained from NLP systems such as MedLEE and knowledgeMap [90] (e.g., UMLS semantic types). The post-processing rules use heuristic patterns to correct possible errors (e.g., false negatives) and further improve the performance. The system achieved an F-score of 0.83 and the study revealed that semantic features derived from existing medical knowledge bases can significantly enhance the performance of clinical NER.

Wang and Akella [91] proposed a hybrid approach to disorder extraction, which integrates supervised ML (SVM), rule-based annotation, and dictionary-based (MetaMap) methods. SVMs utilised rich sets of features, including bag-of-words (BOW), orthographic, morphological, syntactic (POS) and semantic features (e.g., semantic type obtained from SNOMED CT and other semantically related term features obtained from parents and/or children nodes in the ontology). The SVM model integrating all types of features achieved a 0.64 F-score. By incorporating a rule-based annotator, the system performance increased to 0.76 F-score. Finally, with the assistance of MetaMap, the final integrated system achieved an F-score of 0.77.

Comparison

In Table 2.2, we summarise the details of some of the various reported approaches to clinical NER that have been introduced above, categorised into dictionary-based, rule-based, ML based and hybrid types.

It should be noted that the approaches listed in Table 2.2 are not directly comparable in terms of their performance, since they were evaluated on different data sets or corpora; thus, the tabulation of the different performance levels should be treated only as an indication of the general performance trends when different methods are applied to the problem of clinical NER. In chapter 5, however, we provide a direct comparison of the performance of different NER approaches, with the aid of our novel domain-specific annotated corpus.

Table 2.2 Comparison of systems to clinical NER

Approach	System	Key idea	Eval. Corpus	Precision	Recall	F1
Dictionary –based	Gundlapalli et al.[50]	string matching against ICD-9	76,500 clinical records	0.50	0.27	0.35
		MedLEE		0.35	0.77	0.48
Rule-based	AMBIT [75]	lexical, morphological and syntactic features	83 radiology reports annotated for lung cancer signs and locations	0.69	0.83	0.75
	Yang [77]	lexical, orthographic and morphological features	i2b2 medication extraction challenge test set	0.89	0.74	0.81
	Childs et al.[79]	keywords (disease names) their synonyms, symptoms and patterns generated by medical experts	i2b2 recognising obesity challenge	0.97	0.97	0.97
ML	de Bruijn et al.[84]	HMM: token, context and sentence features, as well as features from knowledge sources (UMLS and cTAKES)	i2b2 concept extraction challenge test data set	0.86	0.83	0.85
	Tang et al.[88]	SSVM: BOW, POS, type of notes, section information, semantic categories from UMLS	ShARe/CLEF 2013 test data set	0.80	0.70	0.75
	Pathak et al.[87]	CRF: domain knowledge features , morphological features and orthographic features (e.g., capitalisation) , linguistic features	i2b2 concept extraction challenge test data set	0.88	0.81	0.84

Comparison of systems to clinical NER (Continued from previous page)

Approach	System	Key idea	Eval. Corpus	Precision	Recall	F1
Hybrid	Jiang et al.[89]	CRF coupled with post processing rules	i2b2 concept extraction challenge test data set	0.86	0.81	0.83
	Wang and Akella [91]	SVM coupled with rules and MetaMap	ShARe/CLEF 2013 test data set	0.81	0.74	0.77

2.1.2.2 Relation extraction

Many IE tasks go beyond simply identifying entities; in addition they involve detecting predefined relations between identified entities [92]. In their simplest form, associations between biomedical entities are binary; the detection of such associations forms the focus of this section. However, biomedical relations can involve more than two entities. These more complex associations, often referred to as *events*, are discussed below in the event extraction section.

The goal of relation extraction is to automatically recognise occurrences of relations holding between pairs of given entities. Much work has been carried out on extracting relations from the biomedical literature. Examples of associations of interest include gene-gene interactions [93], interactions between proteins and point mutations [94], proteins and their binding sites [95], genes and diseases [96-98], and genes and their phenotypic context [99]. In the current genomic era, a rich body of work has focussed on automatically extracting relations between genes and proteins. In particular, the detection of Protein-Protein Interactions (PPIs) and gene-disease associations [97, 98] have been the most widely researched topics in biomedical IE [100], due to their critical roles in understanding biological processes.

The extraction of relationships from clinical notes has so far received much less attention [101]. One focussed effort to extract relationships from clinical text was facilitated by the i2b2 NLP shared task, whose goal was to extract and classify relationships between medical problems, treatments and tests, e.g., to determine relations between pairs of medical problems, or to determine that a test can be used to diagnose a medical problem etc. [32]. The Clinical E-Science Framework (CLEF) project aimed to extract clinically significant entities such as drugs, devices and medical conditions, along with relationships between them, such as medical conditions indicated by a device or treated by a drug [102, 103].

In a similar way to NER, biomedical relation extraction faces many challenges. Once again, the difficulties in producing high-quality annotated corpora for relations can inhibit the training and evaluation of ML-based relation extraction systems.

Several methods have been employed to extract such relations from biomedical text. These include the following:

a) Co-occurrence statistics approaches

The simplest method of extracting relations between entity pairs is through the use of co-occurrence statistics. If two entities are repeatedly mentioned together (e.g., in the same sentence or document), then there is a high probability that they are related [20]. Various statistical measures are used to decide whether or not the two co-cited entities are in a relation (e.g., Chi-Square and Log-Likelihood Ratio) [104].

Wang et al. [105] used co-occurrence association to extract two types of relations (diseases-symptoms and diseases-adverse drug) from 25,074 discharge summaries. They apply a contextual filter to certain sections of the reports to improve performance and reduce the amount of potentially confounding information. The contextual filter allows relations to be disregarded if they occur in certain textual contexts and/or sections of the reports. Examples include co-occurrence of the related entities with modifiers corresponding to particular *certainty* values (e.g., negation, low certainty) or those that denote *past* events. Certain sections, such as *family history*, are avoided to try to ensure that the extracted relations are concerned only with the specific patient under discussion, and which describe novel (current) information, rather than information about the patient's past history. Subsequently, extracted disease-symptom and disease-adverse drug event pairs were ranked based on their frequency of occurrence with the disease. A subset of 11 discharge summaries was used as a gold standard to evaluate their method. The evaluation indicated that applying the contextual filters improved recall for disease-symptom relations from 0.85 to 0.90 and that of disease-adverse drug events from 0.43 to 0.75. The filters improved precision from 0.82 to 0.92 for disease-symptom relations and from 0.16 to 0.31 for disease-adverse drug events.

Co-occurrence metrics usually identify statistically significant associations without distinguishing between different types of semantic relations. Since they only depend on statistical strength of association, the relationships identified may not be of medical significance or importance. For example, associations between diseases and symptoms could represent any of the following relations: direct manifestation relations, in which the symptom is a manifestation of the disease (e.g., heart attack, chest pain), or indirect manifestation relations, in which the symptom is manifestation of another disease that is highly associated with the disease of interest (e.g., there is a manifestation relation between 'angina' and 'diabetes', because heart

disease and diabetes are highly associated). As a further example, relations between drugs and symptoms may represent *treat* or *cause* relations, each of which has very different semantics. In a *treat* relation, the drug is used to treat the symptom, whereas in a *cause* relation, the drug causes the symptom [105].

b) Rule-based approaches

Rule-based approaches attempt to go a step further than simple co-occurrence, by patterns which indicate that particular relation types are being described. In a similar way to NER rules, relation extraction rules are usually developed manually by domain experts [106]. An important advance in RE methods is the use of information obtained from dependency parsing trees as features. Dependency parse trees represent the grammatical relations between phrases and words. In dependency parsing, syntactic analysis is commonly enriched with semantic role labelling, which encodes the semantic contributions of words or phrases in the sentence and represents this information within the Predicate Argument Structure (PAS).

The RelEx [107] system used three rules based upon dependency parse trees to identify associations between genes and proteins from a large set of one million MEDLINE abstracts. Their system achieved a performance of 0.79 precision and 0.78 recall based on small hand curated benchmark sets. Miyao et al. [108], described a relation extraction system that utilises deep parsing and term recognition tools to annotate MEDLINE abstracts for PAS and ontological identifiers. The system performs structural matching, i.e., it exploits the PAS to detect relationships between the identified entities.

SemRep [109, 110] is a rule-based NLP system, which extracts semantic predications (i.e., subject-relation-object triples) from biomedical text, supported by domain knowledge obtained from the UMLS Metathesaurus. The system uses MetaMap [63] to extract UMLS concepts from text and defines linguistic and semantic rules specific to each relation. Each semantic relation extracted by SemRep is based on an ontological predication contained within the UMLS Semantic Network, whose arguments are UMLS semantic types. Examples of relations that SemRep is able to extract are DIAGNOSES, CAUSES, LOCATION_OF, ISA, TREATS and PREVENTS [111]. SemRep has been used in a wide range of

applications in biomedical informatics, including automatic summarisation and Literature-Based Discovery (LBD) [112, 113].

While SemRep was originally designed for processing documents from the biomedical literature, it has also been applied to process and extract relations from clinical records [114]. For example, Bejan and Denny [114] used SemRep to extract 958 *treatment* relations between *medication or procedure* and *disease* entries from 6,864 discharge summaries, with an F-score of 0.72.

The types of relations that SemRep can recognise are inherently restricted, since the system uses a predefined predicate ontology that is based on the UMLS Semantic Network.

Recently, Nguyen et al. [115] introduced PASMED, which uses PAS patterns (focusing on verb and preposition predicates) to extract a much broader range of semantic relations from the literature, using a single extraction framework. PASMED uses the UMLS Semantic Network as a constraint to filter out relations that are likely to be spurious (i.e., relations are only extracted between pairs of entities that correspond to concepts in UMLS). The performance of PASMED was compared with that of SemRep on the task of extracting relations from a random selection of 500 sentences from MEDLINE. The results were evaluated by two annotators, who determined the number of correct relations extracted by each system. SemRep and PASMED extracted 346 and 781 correct relations, respectively. The reason that SemRep recognises a far lower number of relations is because its relations are limited to those expressed using a fixed set of verbs defined within the Semantic Network. In contrast, although PASMED also relies upon the Semantic Network, it does so in a less restrictive manner.

PASMED patterns were also evaluated in terms of their ability to extract PPIs from well-known corpora in the biomedical domain. Evaluation of PPI extraction using the BioInfer [116] and LLL [117] corpora revealed that PASMED could achieve a precision and recall of 0.51 and 0.44 on the BioInfer corpus, and 0.87 and 0.81 on the LLL corpus. These results are attributed to the fact that NP pairs of PASMED cover over 82% of the entities in the annotated relations in LLL, whereas PASMED patterns only cover a small portion (i.e., 46%) of the BioInfer corpus. PASMED was also applied to the whole MEDLINE corpus and extracted more than 137 million semantic relations. The most frequent type of semantic relations is between "Amino Acid, Peptide or Protein" entities, which count up to 3.4 million.

The range of relations extracted from MEDLINE provides an insight into the kinds of semantic relations that are actually described in the literature, and which are ultimately extracted by type-specific relation extraction systems [115].

Manually defined rules require large amounts of human effort and, given that the textual patterns that they model can vary between domains. Accordingly, rules can be difficult to port to different domains. Additionally, since language use can be highly creative, it is not realistic to expect that an exhaustive set of rules covering all possible ways of expressing important information can be produced.

c) Machine learning approaches

With the growing availability of biomedical corpora annotated with relations, approaches to relation extraction that employ supervised ML techniques are widely used. This task requires the determination of a set of features that can be used to accurately predict whether or not a given pair of entities is semantically related. A wide range of features has been explored, including orthographic and lexical features of the words that occur between the two entity mentions [118, 119]; syntactic features such as POS, other shallow syntactic features of these words [120] and dependency parse information regarding the grammatical relations that hold between the entity mentions [121].

In [122], Rink et al. used a multi-class SVM to discover 8 different relations between medical problems, treatments and tests mentioned in EHRs from the corpus of the i2b2 relation extraction challenge. The relations in this corpus are fine-grained; for example, a number of different relations can hold between a treatment and a medical problem (e.g., a treatment may improve a medical problem, or it may cause a medical problem). The system relies on the use of a wide range of features, including lexical features (e.g., both the surface forms and the lemmas of the words contained within each of the entities involved in the relation, the latter extracted using WordNet), contextual features (e.g., sequences of words, bigrams and phrase chunk types occurring between the two arguments of the relation) and similarity features (Levenshtein distance was used to find other relations with the most similar sequences of POS tags, chunk tags, shortest dependency paths and word lemmas for the span beginning two words before the first relation argument and ending with the second word after the second argument). Other features were derived from external

knowledge sources such as Wikipedia to provide further information about the likely strength of association between the two potentially related concepts. For example, one such feature determines whether links between the Wikipedia articles correspond to the two relation arguments. Their proposed method was evaluated on the i2b2 relation extraction challenge data sets and achieved the highest score of all systems participating in the challenge, i.e., an F-score of 0.73. It was observed that the use of contextual and similarity features had the biggest impact of the overall performance of the system.

Roberts et al. [103] extracted clinical relations from the CLEF corpus that hold between both entities (e.g. condition, drug, result) and modifiers (e.g. negation). There are seven classes of relations and each entity pair can be linked by one relation only. Therefore, the classification task is considered as a binary classification task between a specific type of relation and the *non-relation* class. The classification is performed using SVMs that employ a number of different sets of features including lexical (e.g., the words that constitute the two related entities), syntactic (e.g., the POS tags of the two entities and different types of information contained within the linguistic analysis provided by dependency parser, such as the number of links on the dependency path connecting the two related entities), direction (e.g., is argument 1 before argument 2 and vice versa), contextual information (e.g., the distance between the two related entities, the surrounding 6 tokens on each side of both entities in the pair) and semantic (e.g., semantic types of the two related entities). The method was evaluated using ten-fold cross validation over the CLEF corpus and achieved a recall of 0.74 and precision at 0.71. The contribution of different feature sets towards the overall macro-averaged F-score was investigated, and revealed that the distance and direction features were the most important, in that they were able to improve the macro-averaged F-score by 10%.

SVMs generally have high performance on various classification tasks, and their usage has been a common trend among the most effective relation extraction systems at the i2b2 relation extraction challenge [32] and by many other relation extraction systems in the biomedical domain [103]. The performance of these techniques depends heavily on the types of features that they use. However, they are usually dependent upon a large amount of manually annotated data for training purposes, which can be expensive and time consuming to obtain.

d) Hybrid approaches

Hybrid approaches combine rule-based and ML methods in an attempt to achieve better performance. Abacha and Zweigenbaum [123] described a hybrid approach that uses patterns developed by domain experts as well as SVM classification to extract relations between diseases and treatments (i.e., *cure*, *prevent* and *side effect*) from MEDLINE abstracts. The number of *cure* relations in the training data was 524, but there were much smaller numbers of *prevent* and *side effect* relations (43 and 44, respectively). The very sparse examples of these latter two relation types make it difficult for the SVM classifier to capture and generalise the patterns that characterise these relations sufficiently to recognise unseen examples in the test set. Accordingly, supplementing the SVM classifier with hand-crafted rules was considered to be a promising solution. The performance of the hybrid approach was compared with the performance of applying pattern-based and ML approaches separately. The results show that the hybrid approach achieved an overall 94.07% F-measure, significantly outperforming either of two individual methods when used in isolation. This result provides strong evidence that rules can be used successfully to enhance ML performance when few training examples are available.

Comparison

Presented in Table 2.3 are the details and results of some of the various reported approaches for clinical RE. In the same way as for Table 2.2, it should be noted that the results are not directly comparable to each other, since different evaluation corpora are used in each case. However, the comparison serves to highlight some general performance trends.

Table 2.3 Comparison of approaches to clinical relation extraction

Approach	System	Key idea	Eval. corpus	Precision	Recall	F1
Co-occurrence statistics	Wang et al.[105]	co-occurrence statistics and contextual filter	11 discharge summaries	0.92	0.90	0.90
Rule-based	RelEx [107]	rule-based: using dependency parse trees	one million MEDLINE abstracts	0.79	0.78	0.78
	SemRep [109]	rule-based: using ontological predications contained in the UMLS Semantic Network.	6,864 discharge summaries	0.81	0.65	0.72
	PASMED [115]	PAS patterns applied to identify pairs of related noun phrases	LLL	0.87	0.81	
			BioInfer	0.51	0.44	0.47
ML	Rink et al. [122]	multi-class SVM: lexical, contextual, syntactic, similarity and Wikipedia features	i2b2 relation extraction challenge test data set	0.72	0.75	0.73
	Robert et al.[103]	SVM: lexical, syntactic, contextual and semantic features	CLEF corpus test data set	0.71	0.74	0.72
Hybrid	Abacha and Zweigenbaum [123]	SVM coupled with pattern-based rules	591 sentences from MEDLINE abstracts	0.94	0.94	0.94

2.1.2.3 Event extraction

Recently, there has been a shift from extracting simple binary relations to more complex relations, known as *events*. Events differ from simple binary relations in a number of ways. Firstly, a more detailed semantic interpretation is associated with the relation, by assigning *semantic roles* to the arguments, to characterise their contributions towards the overall description of the relation. Secondly, it is possible for events to have an arbitrary number of arguments, rather than the two required in binary relations. Thirdly, event arguments could also occur outside the sentence containing the event trigger. Finally, arguments of events are not restricted to being simple entities; an argument can also consist of an "embedded" event. The automatic extraction of events is important, given the frequent occurrence of complex information in text. The growing interest in event extraction research has largely been facilitated by the increasing availability of semantically annotated corpora containing the detailed annotations necessary for the training and evaluation of event extraction systems. The GENIA Event (GE) corpus [124] is one of the largest and most widely used corpora of biomedical text annotated with event structures. It has been used to train a wide variety of biomedical event extraction systems, and a subset of this corpus formed the basis of the first BioNLP shared tasks (i.e., BioNLP'09 GE). The BioNLP'09 GE task [125] was largely based around a simplified subset of the original GE corpus, using only nine of the original 36 event types. Subsequent GE tasks have added complexity, by supplementing abstracts with full papers (BioNLP'11) [126], or by using an exclusively full-paper corpus, annotated with an extended range of event types (BioNLP'13) [126]. Further tasks of the BioNLP'11 and BioNLP'13 shared tasks have concentrated on different biomedical subdomains and/or target application areas, each defining a custom set of event types. These include Epigenetics and Post-translational Modifications (EPI), Infectious Diseases (ID) [127] (BioNLP'11), Cancer Genetics (CG) [128] and Pathway Curation (PC) [129] (BioNLP'13).

Events are centred on a *trigger* (often a verb or nominalisation that characterises the event). The trigger is then linked to an arbitrary number of participants or *arguments* [20, 130], which are assigned roles to denote their semantic contribution

towards the description of the complex relationship. For example, in the sentence "glnAP2 may be activated by NifA", the event trigger is the verb "activated".

The complexity of biomedical events means that arguments may be scattered throughout a sentence. Typically, arguments are structurally and/or semantically related to the event trigger, meaning that a thorough analysis of sentence structure can help to identify them. Accordingly, event extraction is usually aided by the use of semantic NLP techniques such as deep parsing. In particular, dependency parsing tools have been shown to be effective for event extraction, in terms of their ability to carry out a semantic analysis of the sentence, thus helping to identify event arguments and their likely semantic roles [131].

Although the event extraction task is far more complex than the detection of binary relations, the more detailed and expressive semantic information encoded in event structures means that their potential utility in text mining systems is far greater. This has helped to drive research efforts and has resulted in the growing popularity of such systems in the biomedical domain. Event extraction is currently being used to enable and increase the efficiency of several tasks in the biomedical domain, including the annotation of biomedical pathways, Gene Ontology annotation, and the enrichment of biological databases [20].

In the following section, we review some of the most notable reported approaches to the core event extraction subtasks.

Pipeline-based approaches

Most event extraction systems are developed as pipelines that divide the complete event extraction task into three phases: firstly, triggers are recognised and assigned a semantic label corresponding to an event type. The next step determines which entities or other events participate as arguments of the event. Lastly, links between triggers and arguments are created and assigned semantic roles, to form complete event structures [130, 132, 133]. Examples of event extraction systems that use the pipeline approach are the Turku Event Extraction System (TEES) [25] and EventMine [134]. In each system, sentences are represented as graphs with nodes representing named entities and triggers, and edges as event arguments. Both systems employ a multi-class SVM based classifier to assign an event class to each token if it is detected as a trigger, or a negative class otherwise.

TEES employs SVM-based pipelines using different sets of features. For trigger detection, these include: token features (e.g., word stem and character n-grams), syntactic features (e.g., dependency types obtained from the dependency parser) and frequency features generated from bag-of-word counts. After trigger detection, edge detection is used to predict the edges of the semantic graph, thus extracting event arguments [25]. Like the trigger detector, the edge detector is based on a multi-class SVM classifier. The classification employs a range of features including: token (e.g., character n-grams, named entity and type labels) and syntactic (e.g., shortest undirected paths of syntactic dependencies, according to results obtained through application of the Stanford Parser [135]). Each potential edge is classified by the model as any of the permitted event roles (e.g., theme, cause or none).

For trigger recognition, the EventMine system employs an SVM-based classifier with similar feature sets to TEES [25]. In addition to shortest paths between the tokens of interest, dictionary features are also employed. E.g., synonyms, hypernyms and other related forms of matching tokens provided by WordNet [136] and the UMLS SPECIALIST Lexicon [33]. For the task of edge detection, EventMine adopts features employed by TEES to determine the most appropriate semantic label for each candidate edge.

Joint approaches

Joint learning techniques (i.e., non-pipeline-based systems) have been proposed to address the potential problems caused by the percolation and accumulation of errors that is inherent in pipeline-based methods (i.e., errors introduced by one module in the pipeline have a knock-on effect on the output of subsequent modules). Examples of joint learning approaches that have been proposed for application to the event extraction task include joint inference models and dual composition models. Both of these techniques are described below.

Riedel and McCallum of the University of Massachusetts (UMass) were the first to explore joint inferencing of triggers, incoming/outgoing arguments and protein-protein bindings, through the use of Markov logic-based dual decomposition [137].

Building upon the dual decomposition model from UMass, the FAUST system [138] was developed. FAUST uses a stacking model to combine the results of the UMass system [139] with another biomedical event extraction system, i.e., the Stanford Biomedical Event Parser [140]. Specifically, the output of the Stanford

system is used as a feature in the UMass system. While the combined system performs well on the BioNLP 2011 shared task, it was found to produce many false positives which were not predicted by either of component systems when they were applied to the task individually.

Comparison

We provide in Table 2.4 a summary of the most notable approaches to event extraction, subdivided into two groups. The first group is pipeline-based methods, which divide the task into two sub-tasks, namely trigger recognition and argument recognition. Examples of systems that employ pipeline-based methods are TEES and EventMine. Both systems are based on machine learning, which facilitate ease of portability to new tasks, through training on different corpora. Such flexibility has allowed both systems to be adapted to several tasks within the different BioNLP shared tasks, as presented in Table 2.4. EventMine has also shown to be more widely applicable and has demonstrated its state-of-the-art performance when applied to texts of different types and belonging to different subject areas [130, 141]. The second group corresponds to joint model approaches which recognise both triggers and arguments simultaneously. Based on the F-scores obtained by the methods, ranging from 50-57%, which is lower than for relation extraction it can be clearly observed that event extraction is a very challenging and complex problem which is yet to be solved.

Table 2.4 Comparison of approaches to biomedical event extraction

Approach	Proponents or System	Key idea	Eval. Corpus	Precision	Recall	F1
Pipeline	TEES	multi-class SVM trigger & argument classifier	BioNLP ’09 Test	0.58	0.46	0.51
			BioNLP ’11 ID Test	0.48	0.37	0.42
			BioNLP ’11 EPI Test	0.53	0.52	0.53
			BioNLP ’13 CG Test	0.64	0.48	0.55
			BioNLP ’13 PC Test	0.55	0.47	0.51
	EventMine	multi-class SVM trigger & argument classifier, use UMLS and WordNet features	BioNLP ’09 test	0.58	0.48	0.53
			BioNLP ’11 ID Test	0.54	0.60	0.57
			BioNLP ’11 EPI Test	0.55	0.49	0.52
			BioNLP ’13 CG Test	0.55	0.48	0.52
			BioNLP ’13 PC Test	0.53	0.52	0.52
Joint	UMass	Markov logic, dual decomposition	BioNLP ’11 ID Test	0.62	0.46	0.53
			BioNLP ’11 EPI Test	0.41	0.28	0.33
	FAUST	stacking and stacked models	BioNLP ’11 ID Test	0.65	0.48	0.55
			BioNLP ’11 EPI Test	0.44	0.28	0.35

2.2 Biomedical resources

2.2.1 Corpora

The primary input for biomedical TM is text. An important first step towards developing more accurate TM systems is to collect and characterise text that satisfies various types of information needs [142]. One such method of characterisation is to annotate corpora of relevant text by adding layers of linguistic information. One common type of annotation is to add labels to terms, indicating the semantic class to which they belong [143]. Various other levels of annotation can be added to corpora, which include not only semantic annotations, but also discourse annotations (e.g., adding information about anaphoric links in a text) and syntactic annotations (e.g., POS tags).

For biomedical TM purposes, a popular approach to developing corpora for specific subdomains is to collect relevant MEDLINE abstracts, full-text articles, or more recently clinical narratives from EHRs. Biomedical corpora can be classified in different ways. For example, according to the domain that they cover (e.g., molecular biology, oncology), text type (e.g. literature articles, clinical narratives) or types of annotations added (e.g. semantic biomedical entities and relations, or syntactic POS and parse structure, etc.). A review of the most widely-used corpora in the biomedical domain suggested that such corpora need to have three main characteristics in order to be user-friendly and to encourage frequent use, i.e. high-quality documentation, balanced representation and information about Inter-Annotator Agreement (IAA) [142].

The MEDLINE database is the primary knowledge source for research in biology and medical science [18]. Various subsets of MEDLINE relevant to different biomedical subdomains have been collected both by different individual research groups and for community-wide evaluations. Some of these collections consist of abstracts that span a given period of time, e.g., the TREC Genomics track data [129] contains stand-off annotations covering ten years of MEDLINE citations (1994-2003) and has been used to support various tasks including IR, document classification and question answering.

There are several semantically annotated corpora for biomedicine which are publicly available. However until very recently, there were few corpora consisting of clinical text [144]. Of the large number of corpora in the biomedical domain, in the review below we only cover some of the most influential or recent ones.

2.2.1.1 Annotated corpora from the biomedical literature

GENIA

GENIA is the most influential and the most richly annotated biomedical corpus to date. It is one of the only biomedical corpora that contain high-quality linguistic and structural annotation, as well as different types of semantic annotations. The corpus consists of 1,999 abstracts in the area of molecular biology retrieved from MEDLINE using the MeSH terms '*human*', '*blood cells*' and '*transcription factors*'. Annotations include POS tags [145], syntactic parse information [146], biomedical entities [147], co-reference [148], relations [149, 150] and events [151]. Additionally, it is one of the three components of the BioScope corpus [152], which consists of GENIA abstracts, five full-text articles and a collection of radiology reports annotated for negation, uncertainty and their scopes.

PennBioIE

The PennBioIE corpus [153] is intended to cover the subject areas of cancer genomics and drug development. It consists of 1100 abstracts annotated for biomedical entities and for POS. A portion of the corpus is annotated for Penn treebank style syntactic structure [154].

While the above-mentioned corpora have been developed for the purpose of specific research projects, other semantically annotated corpora have been developed in the context of shared task evaluations for IE. Examples of these corpora are BioCreative [155, 156] and LLL [157].

NCBI disease corpus

The NCBI disease corpus consists of 793 PubMed abstracts annotated for disease mentions [158]. Each abstract is annotated by two annotators for disease mentions and corresponding concepts in MeSH or OMIM [34]. The annotation proceeded

through two phases: mention level and concept level. The disease mentions were annotated based on their relevance to biomedical information retrieval. Disease mentions are categorised into four classes: specific disease (e.g., clear-cell renal cell carcinoma), disease class (e.g.., cystic kidney diseases), composite mention (e.g., prostatic, skin, and lung cancer), and modifier categories (e.g., hereditary breast cancer families). The public release of the corpus consists of 6,892 disease mentions that are mapped to 790 unique disease concepts. The corpus constitutes a gold standard aimed at improving the state-of the-art in disease recognition and normalisation. It has already enabled the creation and evaluation of the first ML method for disease normalisation, i.e., DNorm [158].

MEDLINE abstracts cannot contain all the information presented in full-text articles, which contain important information (e.g., results and discussion) that is not reported in the abstracts. Therefore, there is growing interest in annotating corpora of full-text articles, especially according to the growing popularity of open access journal publishing such as BioMed Central [159]. The largest publicly available repository of original, full-text articles is the Open Access subset of PubMed Central[2]. Full-text, open access repositories have allowed the more recent creation of full-text annotated corpora. The ITI TXM is the first corpus of full-text articles. It consists of two parts: 238 full-text articles annotated for tissue expressions and 217 full-texts annotated for PPIs. The Colorado Richly Annotated Full Text Corpus (CRAFT) [160] consists of 97 full-text articles and more than 100 concept types annotated from 9 ontologies and terminologies (e.g., Chemical Entities of Biological Interest, Cell ontology, NCBI Taxonomy, etc.). The corpus adds to the growing body of semantically and syntactically annotated full-text collections (including the full-text portion of the BioScope collection mentioned above). BioCause is another example of full-text corpus; it includes a collection of 19 full text biomedical articles annotated for causality relation on the top of existing event annotations from the BioNLP shared task on infectious diseases. Table 2.5 presents a comparison of some important biomedical corpora.

[2] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist

Table 2.5 Some biomedical corpora and their characteristics

Corpus	Document Type/Size	Domain	Annotation type	Semantic types	Encoding/format
GENIA	2000 abstracts	molecular biology	entities, POS tags, relations, events	POS tags, syntactic parse information, biomedical entities, co-reference, relations and events	in-line XML
BioScope	GENIA abstracts, 9 full-text articles and 1954 radiology reports	radiology and molecular biology	negation, uncertainty and their scopes	modality cues and scope negation	in-line XML
PennBioIE	1100 abstracts	molecular genomics	POS, entities and relations	genes entities and POS tags	stand-off
GENETAG	20,000 MEDLINE sentences	molecular biology	entities	genes/proteins names	stand-off
NCBI disease	793 PubMed abstracts		entities	disease mentions	stand-off XML

Some biomedical corpora and their characteristics (Continued from previous page)

Corpus	Document Type/Size	Domain	Annotation type	Semantic types	Encoding/format
ITI TXM	238 full-text		entities relations	proteins, tissues, genes and tissue expression relation	stand-off XML
	217 full-text		entities and relations	protein entities and PPIs	stand-off XML
CRAFT	97 full-text articles	Genomics	entities	genes and gene products	stand-off
BioCause	19 full-text articles	molecular biology	discourse relations	proteins and chemical organism,	stand-off

2.2.1.2 Clinical annotated corpora

The corpora mentioned so far consist of documents drawn from the biomedical literature. However, as has already been mentioned, corpora of clinical text are much rarer. Gaining access to clinical records for research purposes is difficult due to confidentiality reasons; obtaining permission to release records for reuse by the wider research community is even more challenging. However, due to high demand, a number of such corpora have recently been developed and made available to the wider research community in support of research into clinical TM, thanks to pioneering efforts by medical research groups [12, 161]. These corpora consist of collections of de-identified clinical text, which have been made available according to compliance with user agreement requirements. These corpora include records in the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC II) database [162] and the Pittsburgh collection of clinical reports[3]. To our knowledge, the only annotated corpora of clinical records that have been released to the scientific research

[3] University of Pittsburgh NLP repository

community are those developed in the context of recent shared tasks, as described below.

I2b2

The i2b2 challenges have used collections of clinical records obtained from the Partners Healthcare Research Patient Data Repository, upon which a series of annotation efforts has been carried out to create layered linguistic annotation over the records. These efforts have provided training and testing data for a number of tasks, including de-identification of private health information in clinical records, classification of documents mentioning smoking status [30], recognising documents mentioning obesity and its co-morbidities [31], extracting information related to clinical concepts [78], performing clinical assertion classification [32] and carrying out co-reference resolution [163]. For the purpose of these tasks, a series of clinical corpora has been released by i2b2 containing different levels of annotations. For example, the records for the smoking status and obesity classification tasks are annotated at the document level, whereas the suicide notes used to carry out sentiment analysis are annotated at the sentence level. The i2b2 community was also interested in the detection of concept mentions and for this purpose, they released a corpus of clinical records annotated for clinical concepts (e.g., medical problems, tests, treatments, medications and dosages). Annotated entities were also annotated for assertion information (e.g., whether a medical problem is present/absent in a patient) and temporal relations (e.g., differences in dosages before and after surgery).

I. I2b2 recognising obesity corpus

The obesity challenge corpus consists of 1,237 discharge summaries for patients who had been hospitalised at some point since 1st December 2004 and were overweight or diabetic [31]. A total of 730 records were used for training and 507 records were held out for testing. Both datasets were annotated by two obesity experts from the Massachusetts General Hospital Weight Center. The corpus is annotated at the document level for mentions of obesity and 15 of its co-morbidities: Diabetes mellitus, Hypercholesterolemia, Hypertriglyceridemia, Hypertension, Atherosclerotic disease, Heart failure, Peripheral vascular disease, Venous insufficiency, Osteoarthritis, Obstructive sleep apnoea, Asthma, GERD,

Gallstones/Cholecystectomy, Depression, and Gout. Each document is labelled with one of four possible disease statuses for each disease, i.e.: Present, Absent, or Questionable or Unmentioned. The annotation process involved two tasks: textual and intuitive annotations. The textual annotation required there to be explicit reference to the disease in the text, while the intuitive annotation was based on inferring the disease status from the records, using the annotators' medical expertise. The agreement (Kappa) was 0.86 for the textual annotation task and 0.71 for the intuitive annotation task [31].

II. I2b2 extracting medication information corpus

This corpus consists of 1,243 de-identified discharge summaries annotated for medication (i.e. brand name and generic name) and medication-related information (dosage, frequency, duration and reasons). The training dataset consists of 547 discharge summaries, whilst the remaining 696 were held out for testing [78].

The unique aspect of this corpus, compared to other i2b2 corpora, is the involvement of the TM community in the generation of ground truth data. The community's input helps to demonstrate potential alternatives that can help to overcome the bottleneck of developing ground truth data.

The i2b2 team developed the annotation guidelines, which assume that none of the annotators who contributed to the annotation process has a medical, linguistic or computer science background. The development of the annotation guidelines went through an iterative process. The initial guidelines were released to a group of students from the University of Washington, who were asked to use these guidelines to annotate a small subset of the corpus. IAA was measured and feedback from the students helped the i2b2 team to revise the guidelines. After several iterations, the guidelines, together with 17 records annotated by the i2b2 team, were released to the teams participating in the i2b2 challenge [164]. For the community annotation experiments, 251 records were allocated to the i2b2 challenge teams, with 10 discharge summaries given to each person. A subset of the released records was also annotated by the i2b2 experts to compare the quality of the expert annotations with those produced by the community. IAA was measured by means of micro-averaged F-measures, which were above 0.90. The results showed both that community

annotators can produce annotations of a similar quality to those produced by experts and also that they can achieve IAA comparable to that of experts [164].

III. I2b2 Concepts/assertions and relations challenge

The corpus consists of de-identified discharge summaries from Partners Healthcare, Beth Israel Deaconess Medical Center as well as discharge summaries and progress notes from the University of Pittsburgh Medical Center. In total, there are 394 training reports and 477 test reports [32]. Annotations were made at the following three levels: concept, assertion and relations. The concepts annotation process involved annotating noun and adjective phrases that represent medical problems, treatments and tests.

The assertion annotation task involved annotating each identified medical problem with one of six categories of assertions: “present”, “absent”, “possible”, “conditional”, “hypothetical” and “not associated with the patient” [165].

The relation annotation task consisted of identifying different types of relations that hold between pairs of annotated concepts. Relations could hold between the following pairs of concept types: medical problems and treatments, medical problems and tests, and medical problems and other medical problems. Relation annotation was carried out at the sentence level (i.e., the two related concepts must occur within the boundaries of a single sentence). Previously-added annotations for both concepts and assertions were available for use during the relation annotation process. A total of 8 different relation types between three medical concepts were annotated:

Treatment improves medical problem (TrIP);
Treatment worsens medical problem (TrWP);
Treatment causes medical problem (TrCP);
Treatment is administered for medical problem (TrAP);
Treatment is not administered because of medical problem (TrNAP);
Test reveals medical problem (TeRP);
Test conducted to investigate medical problem (TeCP);
Medical problem indicates medical problem (PIP).

IV. I2b2 co-reference resolution Challenge

The corpus for this challenge consists of two separate corpora, i.e., the i2b2/VA and Ontology Development and Information Extraction (ODIE) corpus.

The i2b2/VA corpus contains 390 de-identified clinical records including discharge summaries and surgery progress notes. It is annotated for mentions and co-reference chains of five different classes (i.e. person, pronoun, problem, test, and treatment). The latter three types match those annotated in the i2b2 concepts/assertions and relations challenge (i.e., the challenge mentioned in the previous section) [163].

The ODIE corpus consists of de-identified clinical reports and pathology reports from the Mayo Clinic together with de-identified discharge records, radiology reports, surgical pathology reports and other reports from the University of Pittsburgh Medical Center. It is produced under the ODIE grant and made available to the i2b2 challenge under SHARP—Secondary Use of Clinical Data from the Office of the National Coordinator (ONC) for Health Information Technology [163]. It is annotated for ten concept categories: anatomical site, disease or syndrome, indicator/reagent/diagnostic aid, laboratory or test result, organ or tissue function, people, procedure, and sign or symptom.

Both the i2b2 and ODIE corpora were annotated by two independent annotators for co-reference pairs. Then the resulting pairs were post-processed to generate co-reference chains. The co-reference chains were reviewed by an adjudicator to resolve any disagreements between the annotators by adding/deleting annotations as necessary. The i2b2/VA corpus contains 5,227 co-reference chains, with an average chain length of 4,326 concept mentions and a maximum chain length of 122 concept mentions. In comparison, the ODIE corpus contains 419 chains, with an average chain length of 5,671 concept mentions and a maximum chain length of 90 mentions [163, 166].

V. I2b2 Identifying risk factors for Heart Disease (HD)

The corpus for this challenge consists of 1,304 de-identified clinical narratives representing 296 diabetic patients (2–5 records per patient) obtained from the Research Patient Data Repository of Partners Healthcare. The narrative records are annotated at document level for the explicit mentions of Coronary Artery Disease (CAD) or risk factors that are associated with its onset

(*diabetes*, *obesity*, *hyperlipidemia*, *hypertension*, *smoking status*, *family history* of CAD and related *medications*), the presence of clinical markers or indicator suggesting the presence of the risk factors (e.g., blood pressure measurement of over 140/90 mm/hg suggests that the patient has hypertension) and temporal attributes including present, before, during, or after the date on the record, giving the potential to create timelines of a patients' progress towards heart disease over the course of their longitudinal record.

This shared task differs from other i2b2 tasks in two ways: firstly, the records in the dataset are longitudinal in that they provide a snapshot of patient's health progress overtime. Secondly, the guiding concept when developing this dataset was to answer a clinical question about patients' health (e.g., ''how do diabetic patients progress towards heart disease, specifically coronary artery disease?") rather than annotating syntactic or semantic categories.

ShARe/CLEF

The ShARe/CLEF-eHEALTH [167] lab organised three shared tasks (i.e. disease names recognition and normalisation (task 1); mapping acronyms and abbreviations to UMLS Concept Unique Identifiers (CUIs) (task 2); and retrieving relevant documents to address patients' queries that arise when they are reading discharge summaries (task 3). The ShARe/CLEF corpus consists of annotations over subsets of the de-identified clinical records in version 2.5 of the MIMIC II database. The corpus contains a number of different document types including: discharge summaries, electrocardiography reports, echo reports and radiology reports. For tasks 1 and 2, a set of 200 training and 100 testing records was annotated for disorder mentions. These were subsequently mapped to appropriate UMLS CUIs. A disorder is any span of text that can be mapped to a concept within the disorder semantic group in the SNOMED CT terminology. The disorder semantic group consists of concepts belonging to one of the following UMLS semantic types: Congenital Abnormality; Acquired Abnormality; Injury or Poisoning; Pathologic Function; Disease or Syndrome; Mental or Behavioral Dysfunction; Cell or Molecular Dysfunction; Experimental Model of Disease; Anatomical Abnormality; Neoplastic Process; or Signs and Symptoms.

For task 3 a large collection of healthcare resources (i.e., one million documents collected from the Internet) was used, including the Drugbank, Diagnosia and Trip

Answers websites. These were made available to the CLEFeHealth 2013 participants through the Khresmoi project. The collection consists of web pages covering a broad range of topics (e.g. health and medicines) and those which target both the general public and healthcare professionals. The annotations include associating 55 queries corresponding to disorders identified in task 1, with corresponding text documents, and assessing the relevance of the identified documents to the queries.

SemEval

The SemEval-2014 task 7 [29] was designed as a follow up to the ShARe/CLEF eHealth 2013 tasks. The task reused the ShARe/CLEF corpus developed for tasks 1 and 2, but added further training data and a new test set (i.e. 133 discharge summaries). The new test data set was annotated by a senior annotator. This was followed by a correction step carried out by the same annotator using a checklist to identify the most frequent errors encountered in the annotation of the original corpus (i.e., the ShARe/CLEF corpus). However, as the gold-standard annotation of the test set is not released by the organiser, the detailed annotation information of the test set is not available.

Other efforts in the clinical domain

The above mentioned corpora are the only publicly available corpora in the clinical domain. However, various research groups have also published descriptions of annotated corpora of clinical records, used within their projects, but not made publicly available. For example, the CLEF corpus [168] is a collection of 20,000 records for patients diagnosed with cancer. The corpus is the most richly semantically annotated corpus for clinical IE, in terms of its annotation for a variety of clinical entities (e.g., drug or device, intervention, condition) and the relations among them (e.g., has indication, has finding). Ogren et al. [169] developed a corpus of clinical notes annotated for disorders, which was used to train and evaluate an NER system. Meystre and Haug [66] described a corpus of 160 clinical records of mixed types (diagnostic procedure reports, radiology reports, history and physicals, etc.) annotated for medical problems. This was used to evaluate an automatic system to extract lists of problems.

Most of the corpora in the biomedical domain have been released to the research community to facilitate reusability. Data reuse saves important amounts of human effort, time and money. For example, the GENIA corpus has been enriched several times with further annotations to include event and meta-knowledge annotations [170]. We ourselves exploited the corpus created in the context of the i2b2 recognising obesity challenge as the basis for constructing a semantically annotated corpus for phenotypic information related to CHF. The corpus is annotated at concept level and includes four major classes: causes, risk factors, signs or symptoms and non-traditional risk factors that highlight the role of kidney failure in the progress of heart failure [171]. Fu et al. utilised 1,000 clinical records from the MIMIC II data base and 30 full-text papers from the biomedical literature to construct a semantically annotated corpus for phenotypic information related to COPD [172, 173]. The corpus is annotated for fine-grained COPD phenotype entities (e.g., problem, treatment, test, etc.) linked to concepts in ontologies (e.g., Human Disease Ontology (DO) [174], Uber Anatomy Ontology UBERON [175], etc.). Later the literature articles subset of the COPD corpus was exploited by Batista-Navarro et al. [176] as a part of the User Interactive Task of BioCreative V to demonstrate Argo, an interoperable web-based text mining platform, suitable for semi-automatic phenotype annotation. The corpus was split into 15 papers for training the TM component constituting the automatic COPD phenotype curation workflows, and 15 papers for curation. Due to time constraints a document has been defined as a smaller chunk of text (e.g., section paragraphs according to each paper's metadata). 124 documents were randomly selected for the curation task. The first 62 documents were used for the pure manual annotations by 5 domain experts and the remaining documents were used for the semi-automatic annotation mode of the task.

The produced COPD phenotype corpus was annotated with four phenotype related entities (i.e., medical condition, signs or symptoms, drugs, proteins) and the relations between COPD and other mentions. Argo proved to accelerate the annotation process and show promising results, achieving an F-score of 0.66. This is close to the IAA of 0.68, indicating that automatic concept annotation workflows perform comparably with human annotators. Table 2.6 presents a comparison of some of the well-known clinical corpora.

Table 2.6 Some clinical corpora and their characteristics

Corpus	Document Type/Size	Annotation type	Semantic types	Encoding/ format	Availability
Recognising Obesity corpus	1,237 discharge summaries	present, absent, questionable for obesity + 15 comorbidities	obesity and its co-morbidities	in-line XML	upon request
Medication information corpus	1,243 discharge summaries	medications, dosages, frequencies, modes, reasons, durations, list/narrative	medications, dosages, modes, frequencies, durations and reasons	XML	upon request
Concepts/assertions and relations	394 discharge summaries and 477 progress note	concepts, assertions, relations	problem, treatment and test	stand-off	upon request
Co-reference resolution	814 different document types including: discharge summaries and progress notes	entities and co-reference chain	person, pronoun, problem, test, and treatment		upon request
Identifying risk factors for heart disease over time	1,304 clinical narratives longitudinal	CAD risk factors and temporal attributes	CAD risk factors and temporal attributes	XML	upon request
ODIE	164 pathology reports, discharge summaries, radiology reports, and	entities and co-reference chain	anatomical site, disease or syndrome, indicator/reagent/diagnostic aid, laboratory or test result, organ or tissue function, people, procedure, and sign or symptom.	stand-off	upon request

Some clinical corpora and their characteristics (Continued from previous page)

Corpus	Document Type/Size	Annotation type	Semantic types	Encoding/ format	Availability
ShARe/CLEF	300 different clinical record types	entities and UMLS CUIs	entities belong to SNOMED CT disorder	stand-off	upon request
SemEval	ShARe/CLEF 2013 data set in addition to 133 discharge summaries.	entities and UMLS CUIs	not available	not available	not available
COPD	1,000 clinical records and 30 full-papers	entities linked to pertinent ontologies	problem, treatment, test	stand-off	
COPD Biocreative V	30 full-papers	entities linked to pertinent ontologies and relations	medical condition, signs or symptoms, proteins and drugs	stand-off	upon request

2.2.1.3 Annotation Process

High-quality annotation, which encodes information about the interpretation of text, is best carried out by humans, since they are most suited to interpreting natural language text. Since annotations are usually used for ML purposes, it is important that they are consistent in order that systems can learn accurately from them. High quality and consistent annotations can be ensured through the preparation of annotation guidelines, which help to ensure that all human annotators have a clear and shared understanding of the task to be undertaken, thus helping to reduce potential inconsistency. Annotation guidelines need to include clear examples illustrating the contexts in which instances should (or should not) be annotated.

There are three main methods that have been applied in the annotation of biomedical text: 1) completely manual annotation in which annotation is started from scratch, based only on annotator's knowledge and expertise; 2) assisted (semi-automatic) annotation, in which the output of an automatic annotation tool is firstly applied to the documents to be annotated; the annotations that are output by the tool are then manually reviewed, corrected and/or augmented by expert annotators; 3) ontology-based annotation, in which only terms and relations present in existing knowledge sources are annotated. Each one of these approaches has its own advantages and disadvantages [20]. For example, the semi-automatic annotation is usually more consistent but it is sometimes biased. Similarly, the ontology-based annotation is likely to be biased because it can only be based on information that is encoded in the knowledge resource used as a basis for the annotation. Using multiple annotators to annotate each document by more than one annotator can help to compensate for potential biases.

2.2.1.4 Approaches to manual annotation

There are three main approaches that have been taken to generating manually annotated corpora:

1) Traditional annotation method: this method has been followed in the creation of almost all the large-scale annotated corpora produced in the context of

NLP research, and it has demonstrated to be highly effective. Annotation is reliant on bringing together a team of people with different types of expertise, to carry out different tasks. NLP experts usually design the guidelines, possibly in collaboration with a domain expert. A separate set of people (also usually domain experts) carry out the annotation, which is usually adjudicated by a more senior domain expert. Technical support staff may also be required to develop, configure or provide assistance in using annotation software. Detailed guidelines for the annotation process are generated prior to the start of the annotation process, and may be subsequently revised to take into account problems and issues that arise as annotation progresses. Although this approach is usually very successful, it can also be very costly, since the processes of developing and refining the guidelines, and training the annotators, can all be very time consuming [177].

2) Crowd sourcing: Instead of being carried out by domain experts, crowd-sourced annotation is carried out by non-experts, via online labour markets, such as Amazon's Mechanical Turk. This approach has been found to work well for simple tasks that are not reliant on high levels of domain expertise. The advantage of crowd-sourcing is that the cost of generating annotations is lower than if they were produced in the traditional way, i.e. by domain experts. Since crowd-sourcing usually involves large numbers of annotators, the annotations can also often be produced more quickly than by a small number of domain experts. The low cost of the annotation, combined with the large number of people often willing to participate, can help to ensure that the annotations produced are of a sufficiently high quality to be useful in the training of NLP systems. By obtaining multiple sets of annotations for the same corpus, a voting scheme can be applied, such that only annotations that are agreed on by a number of different annotators are taken to be correct [177]. This has been successful for the annotation of named entities for clinical trial documents [178].

3) Community annotation: this type of annotation is initiated through the release of a set of initial annotation examples and guidelines to the research community. Researchers from the community then contribute towards

producing more annotations. An example of a successful application of this method is the i2b2 challenge in 2009 where the corpus was annotated by both the challenge organisers and research teams who participated in the challenge. This kind of annotation is very fast and reliable if it is coordinated well [177].

2.2.1.5 Annotation tools

In addition to the IE tools that can assist in carrying out semi-automatic annotation, several tools have been developed specifically to aid in the manual annotation process. Examples of commonly-used tools to annotate biomedical text include Knowtator [179], GATE Teamware [180], Callisto [181], Ellogon [182], WordFreak [183], Extensible Human Oracle Suite of Tools (eHOST) [184] and Brat Rapid Annotation Tool (BRAT) [185]. In order for the annotation tools to be useful they must be easy to use, support various annotation types and allow collaborative annotation.

2.2.2 Knowledge resources

Biomedical resources constitute important sources of domain-specific knowledge used to drive data annotation, data integration and NLP tools.

2.2.2.1 Lexical resources

A rich set of knowledge sources has been developed for the biomedical domain, including both terminologies and ontologies. These resources are employed within a variety of text mining systems, and are often a prerequisite for the successful operation of the systems. Domain-specific lexical resources play a fundamental role in supporting various types of NLP tasks, including NER (e.g. MetaMap and BioLexicon [186]), relation extraction and event extraction. They can also help to facilitate interoperability among systems [187].

UMLS SPECIALIST lexicon

In the biomedical domain, the major lexical resource employed in text processing is the UMLS SPECIALIST Lexicon [188]. In addition, specialised resources can help

in analysing text from biomedical subdomains, such as gene and protein names or chemical and drug names [188].

The UMLS SPECIALIST Lexicon collects lexical items (e.g., biomedical terms that consist of single word or multi word expressions) that are frequently observed in biomedical text. For each term, the lexicon records information about their characteristics, including syntactic information (POS), morphological information (base form and inflectional variants) and orthographic information (spelling variants) [188]. This information can be very important to improve the performance of domain-specific NLP tools, such as POS taggers and parsers.

WordNet

WordNet [136] is an electronic lexical database for the English language, aimed at supporting NLP applications. It has been under development at Princeton University since 1985. In WordNet, terms are organised into sets of synonymous terms (called synsets), each of which represents a single underlying concept. The current version of WordNet (3.1) integrates over 117,000 synsets, which are organised into separate hierarchies for nouns, verbs, adjectives and adverbs, with links between them. For example, the synonyms "renal" and "kidney" occur within different synsets in separate hierarchies (since "renal" is an adjective and "kidney" is a noun). However, a specific relationship (pertainymy) relates the two synsets together. Several types of relations are encoded between synsets in the noun hierarchy, which include hyponymy (specific-generic) and meronymy (part-whole). WordNet is not specialised in any subdomain. However, because of its modest coverage of biomedical terms, its use in biomedical TM research is limited.

2.2.2.2 Terminological resources

Biomedical terminologies consist of lists or hierarchies of terms used in a particular domain or subdomain, usually together with their synonyms. Such terminologies can be extremely important to support NER tasks. The hierarchical organisation used in most terminologies allows for the encoding of parent-child or more-general-to-more-specific relationships, which can be exploited by relation extraction tasks. The UMLS Metathesaurus is an example of a terminological resource that integrates a large number of terminologies.

UMLS Metathesaurus

The UMLS Metathesaurus is a large repository of biomedical concepts and is the major component of the UMLS. The UMLS Metathesaurus is linked to the other two knowledge sources (i.e., semantic network and SPECIALIST Lexicon): firstly, all concepts in the Metathesaurus are assigned at least one semantic type from the semantic network; secondly, many of the terms that appear as concept names in the Metathesaurus also appear in the SPECIALIST Lexicon [188]. The UMLS Metathesaurus is organised by grouping all synonymous terms together into concepts, each of which is assigned a unique identifier. The Metathesaurus is the most comprehensive biomedical resource, its wide scope being facilitated by the range of source vocabularies integrated within it. These include: NCBI for identifying organisms, Gene Ontology for gene products, MeSH for biomedical literature and SNOMED CT for clinical terms [18, 33]. This comprehensive coverage by the Metathesaurus represents link between the vocabularies and the subdomain they represent. These features of the UMLS knowledge resources are exploited by MetaMap, a tool specifically designed to recognise UMLS Metathesaurus concepts in text.

2.2.2.3 Ontological resources

The aim of biomedical ontologies is to encode information about classes of entities (e.g., substances, qualities and processes) of biomedical significance. Examples of these classes include anatomical entities such as mitral valve and processes such as blood circulation. Unlike biomedical terminologies which are concerned with the names of entities, biomedical ontologies are concerned with the definition of biomedical classes and the semantic relations that hold between them. In practice the distinction between terminologies and ontologies is somewhat blurred, since some ontologies collect names of entities, whilst most terminologies also have some degree of hierarchical organisation that reflect the relations among entities.

UMLS semantic network

The UMLS semantic network consists of a set of broad categories or semantic types, together with the semantic relations that hold between them. The semantic types

provide a consistent means to categorise all concepts represented within the UMLS Metathesaurus, whilst the relationships provide a useful means to link together these concepts [188]. The scope of the semantic network is very broad, including 133 semantic types and 54 relation types. Such wide coverage is advantageous in allowing the semantic categorisation of a broad range of terms belonging to a variety of biomedical subdomains [188]. In the Metathesaurus, the structure of each source terminology is preserved. Hence, the relations among concepts are either inherited from the underlying source or specifically generated. Because of the multiple sources and their differing structures, the Metathesaurus cannot provide the consistent kind of structure expected from an ontology.

The semantic network is developed independently of the terminologies integrated in the Metathesaurus. It constitutes a high-level ontology for the biomedical domain. The semantic network is organised around the opposition of the two single-inheritance hierarchies: one for entities and the other for events. The immediate children for *Entity* are *Physical Object* and *Conceptual Entity* while *Event* has *Activity* and *Phenomenon or Process* as direct descendants. In addition to the taxonomy, associative relationships belonging to five different subcategories are defined between semantic types: physical (e.g., part-of, branch-of, ingredient-of), spatial (e.g., location-of), functional (e.g., complicates, causes), temporal (e.g., co-occurs-with), and conceptual (e.g., evaluation-of, diagnoses).

2.3 Summary

This chapter introduced the state-of-the-art in biomedical IE. In particular it provided a detailed review of the methods used to process documents from heterogeneous biomedical sources (i.e., biomedical literature and clinical records). Additionally, we provided a review of notable biomedical resources including: corpora and other knowledge sources.

Chapter 3 Heterogeneity in Biomedical Text

Biomedical text, including both literature articles and EHRs, constitutes a rich source of disease-phenotypic information. However, each text type has a different focus and perspective. EHRs include information about individual patient diagnoses, medication and family history, whilst scientific articles report on the latest research findings and results, together with advances in knowledge relevant to different diseases [49, 189]. Thus, although EHRs and scientific articles provide information that is complementary to each other, this information can be difficult to combine using automated methods. Problems of integrating information may occur according to the different styles of writing, language structure and vocabulary used within each type of text. Accordingly, potentially important unknown associations, which can only be discovered by considering information from both types of text, may be overlooked.

In this chapter, we identify and investigate the phenomenon of linguistic sublanguage variations that can occur in different genres of text within the broad domain of biomedicine. Although a potentially wide variety of text genres falls under this domain (e.g., periodicals, letters, book reviews, case reports, etc.), our investigation is specifically focussed on the different ways in which phenotypic information may be expressed in two types of textual sources that constitute particularly important sources for such information (i.e., EHRs and articles from the literature).

3.1 Background

Differences between the features of text belonging to completely different types and subjects (e.g., biomedical text and newswire) is a well-studied topic. Varying characteristics between such text types, such as sentence length/structure and semantic features, are known to affect the portability of NLP tools [130, 190, 191]. A number of studies have explored the linguistic differences between non-technical and scientific language. For example, in [192], two characteristics of academic writing which differentiate it from general English are described. Firstly, academic writing tends to be structurally ‘compressed’ and it is very common for detailed

information to be integrated into the discourse by means of the modification of phrases, rather than through the addition of extra clauses. For example, consider the following sentence, which is typical of the sentence structure used when writing for a general audience: " *the perspective that considers the participant's point of view and facilities that have been developed to treat waste"*. In contrast, the same information is likely to be expressed in a much more compact way for an academic audience, e.g., "*the participant perspective and facilities for waste treatment".* Secondly, in academic writing it is common to use less elaborate and less explicit language by omitting non-essential information, through the frequent use of passivisation, nominalisation and noun compounding.

Harris [193] proposed the notion of *sublanguage,* which is defined as a subset of general language. He hypothesised that the informational structure and form of specialised (i.e., domain-specific) language can be represented in the form of sublanguage grammar. The sublanguage grammar can then be used by a language processor to extract and encode entities and the relations between them in the text. This theory of sublanguage provides the basis for sublanguage processing within specialised domains.

Based upon Harris' theory, several works have studied variations between sublanguages. For example, Friedman et al. [194] analysed the properties of the different sublanguages that occur within clinical reports and molecular biology articles. They defined restricted ontologies for each domain, highlighted frequent patterns of co-occurrence of words/phrases that occur within each domain and discussed the similarities and differences between them. They concluded that the establishment of a sublanguage grammar is difficult, at least when this is accomplished by carrying out a manual analysis of sample corpora from the two domains. Accordingly, the use of ML techniques can be beneficial in helping to automate or semi-automate the process of discovering how relationships between entities are expressed within a given sublanguage.

3.2 Comparison of biomedical scientific and clinical sublanguages

Previous studies have shown that different subdomains within the biomedical literature (e.g. molecular biology, pharmacology, etc.) exhibit different

linguistic characteristics [17]. In this research, we demonstrate that the types of linguistic variations that occur between two different biomedical text types (EHRs and the biomedical literature) in terms of the expression of the same types of phenotypic information can be substantial.

Clinical sublanguage primarily expresses descriptions of entities and events associated with a patient's state, such as clinical findings, treatments and procedures. These are most commonly expressed using noun phrases [194]. However, the sublanguage used within biomedical scientific text tends to describe complex events associated with biomolecular substances and their interactions. In contrast to clinical text, such events in the biomedical literature are often expressed using verbs or their nominalisations.

Since the sublanguage of a particular scientific field reflects the underlying information, it is not surprising that the variant sublanguages used in different types of biomedical text, e.g., clinical (where the text consists of EHRs) and scientific literature articles, are substantially different. However, they also exhibit a number of interesting similarities.

First of all, in these two text types there is an overlap between both the entities mentioned and the subject matter covered. For example both EHRs and the biomedical literature are concerned with diseases, cells, tissues and molecular components, such as genes and proteins. Therefore, the grammars of clinical and biomedical scientific sublanguages share these informational categories. However, in clinical text, the focus is on describing causes, symptoms and treatments of a given disease (e.g., *She had mild increase in her hypertension during the hospitalization, likely secondary to fluid overload. Her hypertension managed with ACE inhibitor*). In contrast, in literature articles it would be more common to see discussions about molecular level interactions that may lead to a disease occurring (e.g., *They identified 4 chromosomal regions on chromosomes 6, 15, 5, and 2, which showed significant linkage to genes that influence individual blood pressure variation*) [194, 195].

There are also overlaps in terms of the modifiers used in each sublanguage. For example, in both clinical and biomedical scientific text, entities may have modifiers concerned with evidence, change, quantification, degree and body location. In clinical text, such modifiers are frequently used in conjunction with diseases, symptoms and treatments, etc., to provide accurate and detailed descriptions about patient health status. For example, in clinical text the sentence "*slight improvement*

in asthma" includes both a modification of *asthma*, referring to a change in the condition (i.e., *improvement*) and a modification of the information about the change, referring to the degree of change (i.e., *slight*). Similarly, the sentences "*increased lower extremity edema"* and "*increased swelling of lower extremities*" include modifiers referring to change (i.e., *increased*) and location (i.e., *lower extremities*). A further example from clinical text is "*asthma not ruled out",* in which the condition is modified by evidential information (i.e., *not ruled out*).

In biomedical scientific text, modifiers are also used but usually in a different way to clinical text. In biomedical text, mentions of diseases, symptoms etc. often occur unmodified (e.g., *leg edema*). However, according to the nature of the information contained within literature articles, it is more common for modifiers to be used in the context of providing summaries about the advances in knowledge relevant to different diseases. An example is the use of the evidential modifier "*these results suggest that"* in the sentence "*these results suggest that menaquinone-7 improves disease activity in patients with rheumatoid arthritis"*.

Although there are some similarities in the ways in which information can be expressed in each text type, there are also some significant differences. Figure 3.1 shows an example text snippet taken from an EHR that describes a patient's health condition upon arrival at hospital. The figure also shows a text snippet from a biomedical literature article, which also reports information about patient status. From the figure, it can be appreciated that literature articles constitute formal text that conforms to conventions of structure, readability and grammaticality. For example, information is presented clearly in the form of full sentences. In contrast, EHRs, which are intended to be used only in a hospital context by doctors, are usually less structured, with short or incomplete/ungrammatical sentences and contain many domain-specific abbreviations or acronyms (whose expansions may be ambiguous are/or not fully explained). Furthermore, there is often a large degree of orthographic and lexical variability, while spelling and/or grammar mistakes are frequent [189], with up to 10% of words being misspelt [43]. Compared to literature articles, therefore, the general textual characteristics of EHRs pose many challenges for TM analysis.

A further difference between the two text types is that there is considerable divergence in the nature of the relationships specified between entities. For example, although there is some overlap in the types of entities mentioned, the semantic

relationships in which these entities are typically involved tend to vary considerably between clinical and biomedical scientific text. For example, in the clinical domain, diseases are usually associated with procedures (*chest x-ray showed pneumonia*) or treatments (*on Bactrim for urinary tract infection*). Conversely, in biomedical text diseases are primarily associated with genomic variations (*BRCA1 and BRCA2 are the two genes in which mutation is associated with hereditary breast and ovarian cancers*).

A major difference between the sublanguages used in each of the text types is the complexity of the entities and relations. In the clinical domain, information is commonly described using noun phrases, which are generally modified by adjectives or nouns (e.g., *elevated right heart filling pressures, significant left atrial dilatation*). The relationships involving entities can vary in complexity. In the simplest case, a relation may associate a single finding with a modifier (e.g., *heart was enlarged, blood pressure is low*). However, the information expressed by relations is often more complex. For example, it is common for several modifiers to be associated with a finding (e.g., *Abdomen is soft, nontender, nondistended*), for one or more findings to be associated with a procedure and/or treatment (e.g., *Echo showed normal ejection fraction, normal systolic function, and normal valve motion*), for one or more findings to be associated with a negation modifier (e.g., *The patient denies shortness of breath, chest pain, orthopnea*), for causality to be expressed between entities (e.g., *anemia due to iron deficiency and chronic renal insufficiency*) and for temporal ordering to be specified (e.g., *coronary artery disease status post non-ST segment myocardial infarction*). In biomedical scientific text, the emphasis is rather on descriptions of biomolecular pathways, which may be expressed by complex interactions and other relations. Pathway relations that describe the interactions between substances are often expressed using verbs (e.g., *p53 binds to il2*). Frequently, the nominalised form of the verb (e.g., *activation*) is used to allow for nesting, for example, where the cause of an interaction is specified to be one or more other interactions (e.g., *activation of protein kinase C and elevation of cAMP interact synergistically to raise c-Fos and AP-1 activity in Jurkat cells*).

Tables 3.1 and 3.2 present a comparison in terms of our two compared sublanguages in the biomedical domain (i.e., clinical and biomedical scientific text), in terms of both similarities and differences.

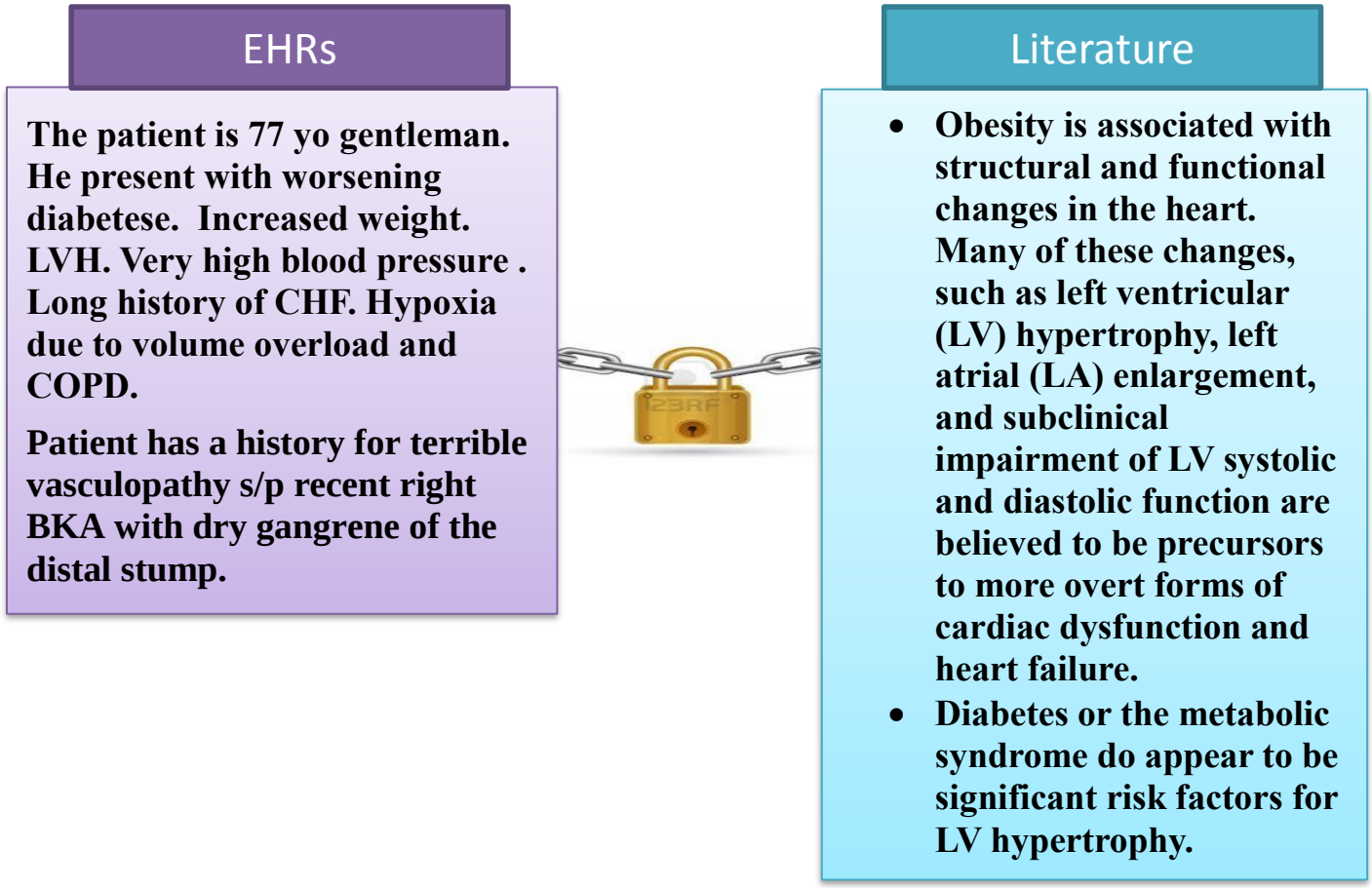


Figure 3.1 Phenotypic information encoded in two different types of text

Table 3.1 Comparison of textual features in clinical and biomedical scientific text

Textual feature	**Similarities between clinical and scientific text**	**Differences between clinical and scientific text**
Subject matter	both text types cover similar subjects and contain mentions of similar types of entities, such as diseases, cells, tissues and molecular components, such as genes and proteins.	• **EHRs**: disease descriptions specify causes, symptom, treatments etc. Molecular components occur in pathology reports and they denote findings of tests associated with molecular markers. • **Biomedical literature**: disease mentioned in context of biomolecular interactions.
Use of modifiers	both text types use modifiers relating to evidence, change, quantification and degree.	• **EHRs**: frequent use of modifiers in conjunction with diseases, symptoms and diagnostic procedures to provide detailed descriptions. • **Biomedical literature**: usually used in context of providing summaries about advances in knowledge relevant to different diseases

Table 3.2 Comparison of semantic-level variability in clinical and biomedical scientific text

Type of semantic variability	**Biomedical scientific text**	**Clinical text**
Complexity of entities	entities are less descriptive with few modifiers.	entities are descriptive and dominated by nouns and pre-modifiers (generally adjectives or nouns).
Complexity of relations	• disease information primarily associated with genomic variations and molecular interactions. • dominated by complex and highly nested relations between biological substances, described using verbs (e.g., *activate*) or their nominalisations (e.g., *activation*).	• diseases usually associated with procedures or treatments. • the relation can be simple to connect single finding with associated modifiers (e.g.. body location). • relations can be complex, e.g. connecting several findings with procedures, or denoting causality between entities. Relations may have additional information associated with them, such as modality.

3.3 Variability in expressing phenotypic information in EHRs and the biomedical literature

Since both EHRs and biomedical literature articles contain different, but complementary types of valuable phenotypic information, combining details from each source can be useful in uncovering new disease-phenotypic associations. The discovery of such associations can be instrumental in accelerating scientific progress towards an enhanced understanding of the etiology of human diseases and in facilitating better disease prevention and treatments.

The recognition of phenotypic concepts presents a particular challenge, since each concept can often be expressed in text in a number of different ways. Indeed, phenotypic concepts can even correspond to complete sentences, e.g., "*two brothers died of heart disease"*. Different types of variations may occur both within a particular text type and across different text types. Examples of variations that can occur between clinical and biomedical sublanguages in expressing same phenotypic information include lexical (*light-headedness* vs. *lightheadedness*), syntactic (*jugular venous pressure is elevated* vs. *elevated jugular venous pressure*) and semantic variations (*hypertension* vs. *high blood pressure*). Further examples of these and other types of variations are provided in Table 3.3. Variability in the expression of phenotypic concepts may occur according to the different levels of experience and backgrounds of the clinicians who author the text, together with the different styles of writing used in articles and EHRs. However, it is important to take such linguistic differences into account, since they can constitute a potential barrier to the successful application of TM methods. NLP tools that perform well on textual data from one source may fail to perform at the same level on other sources, unless the tools are tailored to these alternative sources in some way [17]. Mapping or normalising mentions of various types of phenotypic information that appear in both EHRs and literature articles to concepts in domain-specific knowledge resources such as UMLS [33] can help to draw generalisations about information that may be expressed in text in many different ways. The normalised phenotypic concepts can be used as a first step towards the automatic integration of knowledge that is dispersed within these two text types.

The above analysis highlights the importance of developing TM tools that are specifically tailored to extracting phenotypic information from different textual sources, in order to

facilitate a more in-depth understanding by clinicians of the factors surrounding deadly diseases such as CHF, and thus to allow for advances in their treatment. However, the many challenges to be faced in developing robust TM tools for application in this domain go towards explaining the current paucity of research in this area. To the best of our knowledge, the work described in this thesis is the first effort to extract and integrate phenotypic information from two different biomedical sources, using TM techniques.

Table 3.3 Differences in expressing the same phenotypic concepts in EHRs and the literature

Type of variability	**EHR mentions**	**Article mentions**
Synonymy	Drop in blood pressure	Hypotension
Lexical	light-headedness	Lightheadedness
Syntactic structure	left ventricle is dilated	left ventricular dilatation
Word ordering	cardiac output decrease	decreased cardiac output
Spelling variation	hyperkalemic	Hyperkalemia
Modification	moderate left ventricular enlargement	left ventricular enlargement

3.4 Summary

In this chapter we have investigated the differences between two sublanguages (i.e., clinical and scientific) in the biomedical domain. The fact that both sublanguages cover broadly the same subdomain means that they naturally exhibit certain similarities. Since they cover similar subjects, the same types of entities are mentioned within them (e.g., diseases, findings, drugs, etc.). However, there are also some significant differences in the features of the sublanguages. Clinical sublanguage, found within clinical reports, is highly descriptive (e.g. modifiers are frequently associated with concepts) in order to provide accurate details about patient health status. However, since this text is not intended to be published, it is common to encounter features such as ungrammatical sentences and spelling errors. In contrast, biomedical scientific language found within literature articles is often structurally compressed

and less elaborate, with modifiers being used in a different way. However, compared to clinical text, information in the literature is presented clearly and follows grammatical conventions. A further difference between the two text types is that there is considerable divergence in the nature of the relationships specified between entities.

The differences between the language used in EHRs and the literature have motivated our interest in comparing and contrasting the performance of TM tools on different text types/subdomains, in exploring domain adaptation methods, and in developing a manually annotated domain-specific corpus that is representative of the two different text types, to act as a gold standard to train and evaluate different TM techniques.

Chapter 4 Corpus Development

Annotated corpora in the biomedical domain mostly consist of texts drawn from the biomedical literature. Typically abstracts are sourced from MEDLINE (e.g., GENIA [124], PennBioIE [153], BioInfer [116] and NCBI disease corpus [158]). Further studies have shown that applying information extraction techniques to full-text articles is also a feasible task [127, 196]. This has resulted in increased interest towards the development of annotated corpora containing full-text articles (e.g., BioScope [152], CRAFT [197] and the BioNLP 2011 infectious disease dataset [127]). In the biomedical domain, a variety of corpora with different levels of annotation have been developed. These levels include syntactic (e.g., sentences [147], tokens [145], dependencies [116]), semantic (e.g., named entities [147], relations [127], events [124, 198]) and discourse (e.g., discourse relations [170, 199]), whilst the underlying texts may belong to different sub-domains (i.e., molecular biology, anatomy, chemistry etc.). However as we explained in Section 2.2.1, despite the richness of the available annotation very few of the previously annotated semantic information is relevant to phenotypes.

Although clinical corpora containing EHRs are rare (due to confidentiality and privacy concerns [32]), a small number have recently become available. These have been created for shared tasks to support the development of clinical NLP methods.

In addition to these shared tasks, several research groups have published descriptions of annotated clinical corpora developed and used within their own research, although they have mostly not been made publicly available. These corpora vary in terms of the text type and annotation granularity. For example, a corpus that is annotated for medical problems in narrative text clinical documents is described in [66]. However, the corpus is annotated at the document level, which makes it more suitable for developing and evaluating IR methods, rather than for supporting the extraction of fine-grained information about phenotypic concepts.

Several other clinical corpora have been enriched with text-bound annotations that encode the exact locations of a phenotypic concept within the text [168]. CLEF was one of the first corpora to include detailed semantic and fine-grained text-bound annotations. It is annotated with a variety of clinical entities (e.g., drugs, investigation, conditions and results), relations

between these entities such as: has_finding which holds between condition and result entities. For example, in the following sentence "*This patient has had a lymph node biopsy which shows melanoma in his right groin"*, the result *melanoma* is a finding of the investigation *biopsy*. However the corpus has not been made publicly available. Subsequently, South et al. produced a corpus [200] which is annotated for all mentions of signs or symptoms, medications and procedures relevant to inflammatory bowel disease. The ShARe/CLEF corpus of clinical notes is annotated for specific disorder mentions including diseases and signs or symptoms. The annotated concepts are linked to terms in the SNOMED CT vocabulary [28]. Another scheme [201] has similar specifications to the ShARe/CLEF corpus, in that its purpose was to annotate entities pertaining to the disorder semantic group in SNOMED CT vocabulary. However, the approach to performing the annotation was different from ShARe/CLEF as an automatic tool [63] was used to pre-annotate the corpus. The corpus developed in the context of the i2b2 concepts and relation challenge is a further example of a clinical resource that can support a variety of IE tasks [32]. The corpus is annotated for named entities concerning medical problems, tests and treatments and the relations between them.

In order to build upon existing semantically annotated corpora in the clinical domain, we have developed the PhenoCHF corpus. PhenoCHF is comparable to some of the other corpora annotated above, in that many of the annotated entities fall under the general definition of the UMLS disorders semantic group. However, we have used a finer-grained classification of such entities than previous efforts, in order to capture more detail about information that is particularly important in the study of CHF according to guidance from a domain expert. In contrast to most previous clinical text annotation efforts, the corpus also identifies relationships between the annotated entities.

PhenoCHF [171] is unique amongst annotation efforts within the clinical NLP community, in its integration of information from both EHRs and literature articles to allow the training of systems that are sufficiently robust to recognise relevant information in heterogeneous, and potentially complementary, sources as we discussed in Chapter 3. The annotation scheme, whose design was guided by a domain expert (i.e., a cardiologist) includes both entities and relations pertinent to CHF.

In the remainder of this chapter, we describe our work on developing PhenoCHF corpus as a resource for training and evaluating different TM tools. We include details about the composition of the corpus, the design of the annotation schema and guidelines, and statistics

regarding the reliability/consistency of the annotated information.

4.1 Description of the corpus

4.1.1 Corpus composition

The portion of our corpus containing information from EHRs consists of a set of discharge summaries, which constitute a subset of the data released for the second i2b2 shared task, known as "recognising obesity" [31]. The challenge dealt with the extraction of information about obesity and 15 of its comorbidities, such as CHF, hypertension and diabetes mellitus as described in Section 2.2.1. The original annotation for the discharge summaries consists only of document level annotations, which encode whether or not obesity or any of its comorbidities are mentioned in the summary.

The discharge summaries in the PhenoCHF corpus were chosen by filtering the original i2b2 corpus, such that only those summaries for patients with CHF and kidney failure were retained. This was achieved by searching for summaries containing the disease names *CHF* and *renal failure*, acronyms (*CRF, CRI*) or synonyms (e.g., *renal insufficiency, kidney failure)*. A total of 300 discharge summaries matched these search criteria.

The second part of PhenoCHF consists of the 10 most recent full-text articles (at the time of query submission) retrieved from the PubMed Central Open Access database, using the following query which was determined by a domain expert: "*Heart failure Clinical presentation" OR "Heart failure clinical features" OR "Heart failure symptoms" OR "Heart failure clinical manifestation" OR "Heart failure clinical picture" AND ("Chronic renal failure "OR "Renal failure" OR "Chronic renal insufficiency" OR "Renal insufficiency" OR "Kidney failure" OR "CRF"OR"CRI").*

All documents in PhenoCHF were manually annotated by medical experts for phenotypic information related to CHF. The annotation includes entity mentions relating to four semantic categories of phenotype-related information, as shown in Table 4.1.

Table 4.1 Types and statistics of entity mentions annotated in the PhenoCHF corpus

Semantic categories	**Description**	**# of annotated mentions in narrative EHRs**	**# of annotated mentions in literature articles**
Cause	Any medical problem that contributes to the occurrence of CHF	1320	1107
Risk factor	A condition that increases the chance of a patient having CHF	1335	408
Sign or symptom	Any observable manifestation of a disease which is experienced by a patient and reported to a physician	2449	304
Non-traditional risk factor	Conditions associated with abnormalities in kidney functions that put a patient at higher risk of developing *signs or symptoms* and *causes* of CHF	308	329

4.1.2 Schema

The design of the annotation schema was guided through an analysis of the relevant discharge summaries, in conjunction with a review of comparable domain specific schemata and guidelines, i.e., those from the CLEF and i2b2 shared tasks. The schema is based on a set of requirements developed by a cardiologist. Taking into account our chosen focus of annotating phenotypic information relating to the CHF disease, the cardiologist was asked firstly to determine a set of relevant entity types that relate to CHF phenotypic information and the role of the decline in kidney function in the cycle of CHF (exemplified in Table 4.2), secondly to locate words that modify the entities (such as polarity clues, i.e., words that negate entities) and thirdly to identify the types of relationships that exist between these entity types in the description of phenotype information (see Table 4.3). The types of annotation are described in a schema, shown in Figure 4.1. Following the manual annotation of entities and relationships, annotated entities in the corpus are mapped semi-automatically onto semantic types in UMLS with the aid of MetaMap.

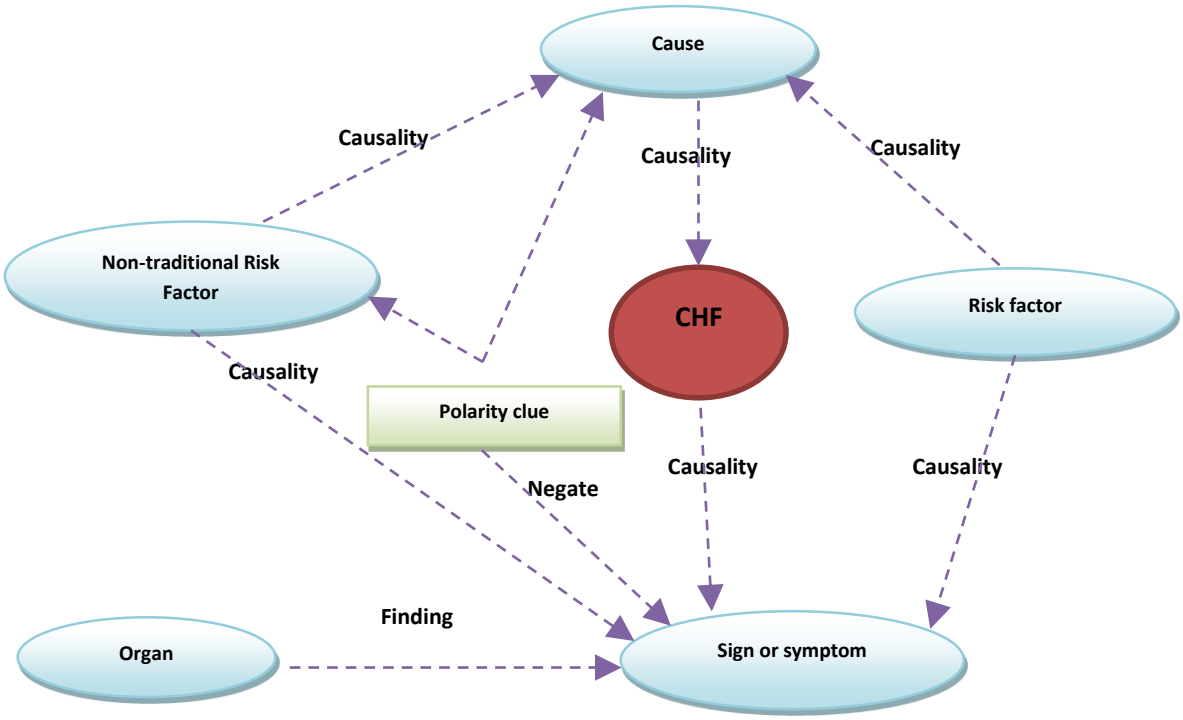


Figure 4.1 Annotation Schema: ovals represent entities, rectangle represents negation modifier and lines represent relationships

Relation annotation

Three types of intra-sentential relationships have been defined (see Table 4.3). Each type of relationship links two specific types of entities. The relationships are based on UMLS semantic network relations, e.g. the *causality* relation in the PhenoCHF corpus is based on *causes* in the UMLS semantic network that holds between two diseases or a disease and a pathologic function. The *finding* in PhenoCHF relation is mapped to the UMLS *manifestation* relation that holds between 'sign or symptom' and 'body part or organ'.

Table 4.2 Annotated phenotype entity classes

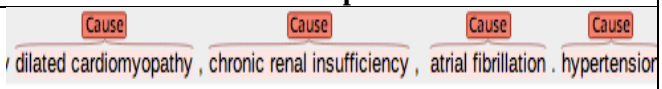
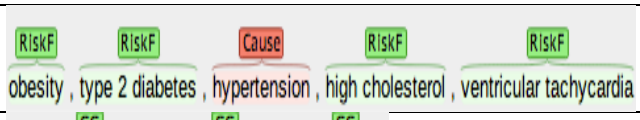
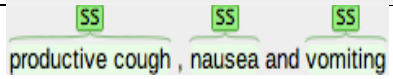
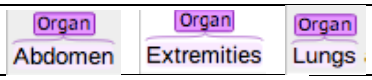
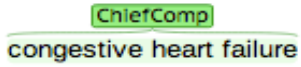
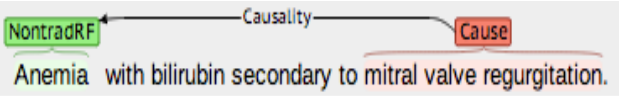
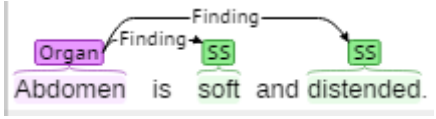
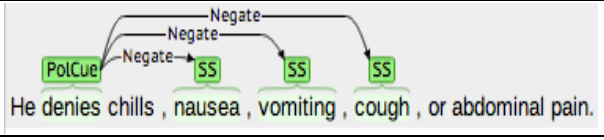
Entity Type	Description	Example
Cause	any medical problem that contributes to the occurrence of CHF	
Risk factors	a condition that increases the chance of a patient having CHF disease	
Sign or symptom	any observable manifestation of a disease which is experienced by a patient and reported to the physician	
Non-traditional risk factor	conditions associated with abnormalities in kidney functions that put the patient at higher risk of developing "signs or symptoms" and causes of CHF	
Organ	any body part	
Chief complaint	mentions of CHF	

Table 4.3 Description of annotated relations

Relation Type	First argument type	Second argument types	Description	Example
Causality	Chief complaint Cause Risk factors Non-traditional risk factor	Non-traditional Cause	This relationship links two concepts in which one concept causes the other to occur.	
Finding	Organ	Sign or symptom	This relationship links the organ to the manifestation or abnormal variation that is observed during the diagnosis process.	
Negate	Polarity cue	Finding Cause Non-traditional risk factor	This is a one-way relation that relates a negation attribute (polarity clue) to the condition it negates.	

4.1.3 Development of annotation guidelines

Consistent annotation across the entire corpus is a prerequisite of a high quality and reliable gold standard.

The same annotation schema and guidelines were used for both the discharge summaries and the scientific full articles. In the latter, annotations were omitted for: organ entities, polarity clues and relations. This decision was taken due to the differing ways in which phenotypic information is expressed in discharge summaries and scientific articles. In discharge summaries, phenotypic information is explicitly described in the patient's medical history, diagnoses and test results, whereas scientific articles summarise results and research findings. This means that certain types of information that occur frequently in discharge summaries are extremely rare in scientific articles, such that their occurrences are too sparse to be useful in training TM systems. Hence, these were not annotated.

We developed the guidelines through an iterative process. They were firstly tested by providing the two annotators who are doctors with a small common set of records to be annotated independently. An analysis of errors and disagreements in this annotation set was used to refine the guidelines. Specific issues were concerned with difficulty in differentiating between causes and risk factors (e.g., annotators were confused whether "*diabetes"* and *"hypertension"* are causes or risk factors of CHF) and the choice of a correct annotation span (e.g., whether to include the modifier "*left*" in the annotation of "*left atrial enlargement"*). Another source of error was that only the first mention of a phenotype was annotated, rather than all of its mentioned instances. The guidelines were updated to make the span decision easier for the annotators by refining the definitions for cause and risk factors and providing more examples for the correct spans. After updating the guidelines, agreement between the annotators improved, and the whole corpus was subsequently annotated by both annotators. For the complete guidelines the reader is referred to Appendix A.

In addition to the guidelines, the annotators were supported through regular meetings allowing the discussion of questions and other issues that arose during the annotation process. The cardiologist who helped to design the scheme acted as the adjudicator in these meetings and was responsible for making the final decision to resolve all problems and discrepancies.

Figure 4.2 shows the most prevalent phenotypes in the corpus and their distribution in the discharge summaries and articles. In discharge summaries, there is a large

emphasis on describing the signs or symptoms of the disease. These play a much less significant role in scientific articles, where the dominant topics are non-traditional risk factors and the etiology of CHF.

4.1.4 Annotation tool

The annotation was carried out using the BRAT [185], a highly-configurable and flexible web-based tool for textual annotation. BRAT is simple to configure for our requirements and easy for non-technical annotators to use. These factors contributed to our choice.

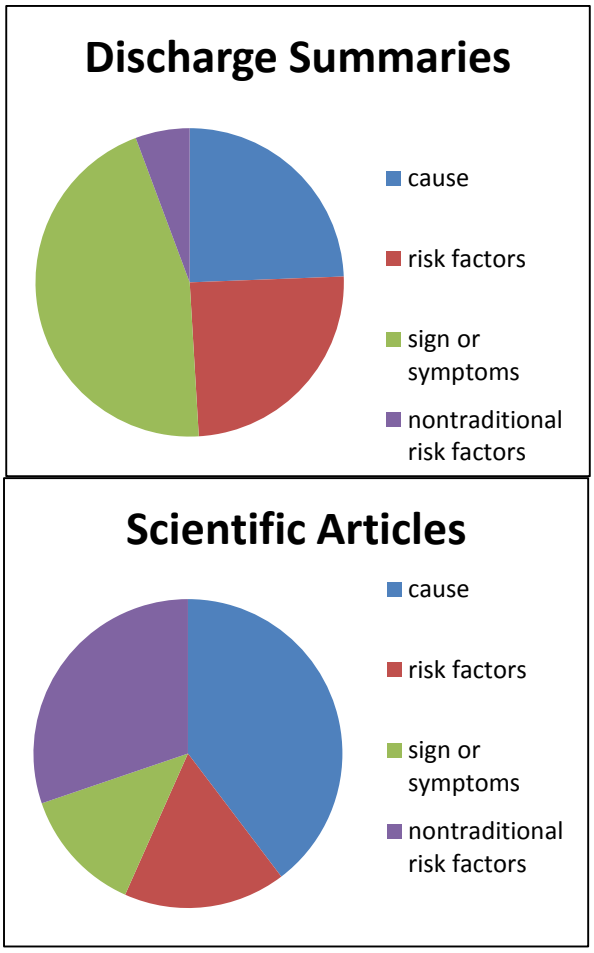


Figure 4.2 Distribution of phenotype information in the corpus

4.1.5 Evaluation

Annotations in the corpus should reflect the instructions provided in the guidelines as closely as possible, in order to ensure their high quality. A standard means of providing evidence regarding the reliability of annotations in a corpus is to calculate a

statistic known as the IAA. A high IAA score provides assurance that the two annotators can produce consistent annotations when working independently and separately.

There are several different methods of calculating IAA, which can be influenced by the exact nature of the annotation task. The simplest is to calculate the percentage of absolute agreement by dividing the number of agreed annotations by the total number of annotations [202]. However, absolute agreement is not considered very accurate, in that it does not take into account that some proportion of agreement between the two annotators can be expected by chance [203].

Accordingly, a more widely used coefficient of agreement is Cohen's Kappa [204]. However, the calculation of Cohen's Kappa requires that the total number of annotated items is known in advance, meaning that it is unsuitable in our case. Instead, we use the measures of precision, recall and F-measure to indicate the level of inter-annotator reliability [205]. In order to carry out such calculations, the set of annotations produced by one of the annotators is considered as the 'gold standard', i.e., the set of correct annotations.

Precision is the percentage of correct positive predictions annotated by the second annotator, compared to the first annotator's assumed gold standard. Precision is calculated as the ratio between the true positive (TP) entities and the total number of entities annotated by the second annotator (the sum of TP and False Positives (FP)).

$$P = TP / TP + FP$$

Recall is the percentage of positive cases recognised by the second annotator. It is calculated as the ratio between the True Positive (TP) entities and the number of named entities that the second annotator was expected to recognise, based on the gold standard (the sum of TP and False Negatives (FN)).

$$R = TP / TP + FN$$

F1-measure is the harmonic mean between precision and recall.

$$F1\text{-}measure = 2* (Precision * Recall) / Precision + Recall$$

We have calculated separate IAA scores for the discharge summaries and the scientific articles. Table 4.4 summarises agreement rates for term annotation in the discharge summaries, showing results for both individual entity types and macro-averaged scores over all entity types.

Relaxed matching criteria were employed, such that annotations added by the two annotators were considered as a match if their spans overlapped. For example, if one annotator annotated the span *increased shortness of breath,* and the other annotated only *shortness of breath,* this would count as a match.

In comparison to related annotation efforts, the IAA rates are quite high. However, it should be noted that the number of targeted classes and relations in our corpus is quite small and focussed compared to other related corpora.

Agreement statistics for scientific articles are shown in Table 4.5. Agreement is somewhat lower than for discharge summaries. This could be due to the fact that the annotators (clinicians) are more used to dealing with discharge summaries in their day-to-day work, and so are more accustomed to locating information in this type of text. Scientific articles are much longer and generally include more complex language, ideas and analyses, which may require more than one reading to fully comprehend the information within them.

Table 4.6 shows the agreement rates for relation annotation in the discharge summaries. The agreement rates for relationships are relatively high. This can partly be explained by the deep domain knowledge possessed by the annotators and also by the fact that the relationships identified were relatively simple, linking only two pre-annotated entities.

Table 4.4 Term annotation agreement statistics for discharge summaries

	Causality	**Risk factor**	**Sign or Symptom**	**Non-traditional risk factor**	**Polarity clue**	**Organ**	**Macro-average**
Total number of annotations	1320	1335	2449	308	492	432	-
TP	1265	1263	2388	266	276	407	-
FN	55	72	61	42	18	25	-
FP	40	50	28	60	10	40	-
Precision	0.97	0.96	0.98	0.81	0.96	0.91	0.93
Recall	0.95	0.94	0.97	0.86	0.93	0.94	0.93
F-score	0.95	0.94	0.97	0.83	0.94	0.92	0.92

Table 4.5 Term annotation agreement statistics for scientific articles

	Cause	Risk factor	Sign or Symptoms	Non-traditional risk factor	Macro-average
Total number of annotations	357	272	153	118	-
TP	284	225	120	85	-
FN	73	47	33	33	-
FP	27	34	19	17	-
Precision	0.91	0.86	0.86	0.83	0.86
Recall	0.76	0.82	0.78	0.72	0.77
F-score	0.82	0.84	0.82	0.77	0.81

Table 4.6 Relations annotation and agreement statistics for discharge summaries

	Causality	Finding	Negate	Macro-average
Total number of annotations	125	364	692	-
TP	102	343	657	-
FN	23	21	35	-
FP	12	19	29	-
Precision	0.94	0.95	0.96	0.95
Recall	0.80	0.94	0.95	0.89
F-score	0.86	0.94	0.95	0.91

4.2 Summary

In this chapter, we have presented a detailed description of our procedure for the development of the PhenoCHF corpus. This consisted of the development of a novel annotation scheme, specifically tailored to annotating phenotypes within the context of CHF and kidney failure. Subsequently, we manually applied the scheme to create an annotated corpus of documents from two different biomedical sources (i.e., literature articles and EHRs). The corpus has been developed to act as a gold standard resource to carry out domain adaption and development of TM tools that can extract and integrate phenotypic information from both text types. We will use this corpus to drive our research in the following chapters of this thesis.

Chapter 5 Phenotypic extraction

Due to the large volume of clinical information and test results in EHRs, it is very time consuming for clinicians to read EHRs to have a better idea about the patient's phenotypic information. TM techniques have been applied successfully to extract different medical information from clinical records. Such techniques can be extended to allow the extraction of phenotypic information and the relations between them from free text.

Following the creation of the PhenoCHF corpus [171], we proceeded to carry out experiments to automatically extract phenotypic information form the PhenoCHF corpus [206]. As we discussed in Chapter 4, PhenoCHF corpus is annotated for phenotypic entities and the relations between them. In this chapter, we will discuss the methods we have used to extract phenotypic mentions. Presented in Figure 5.1 is the workflow we have applied on the PhenoCHF corpus.

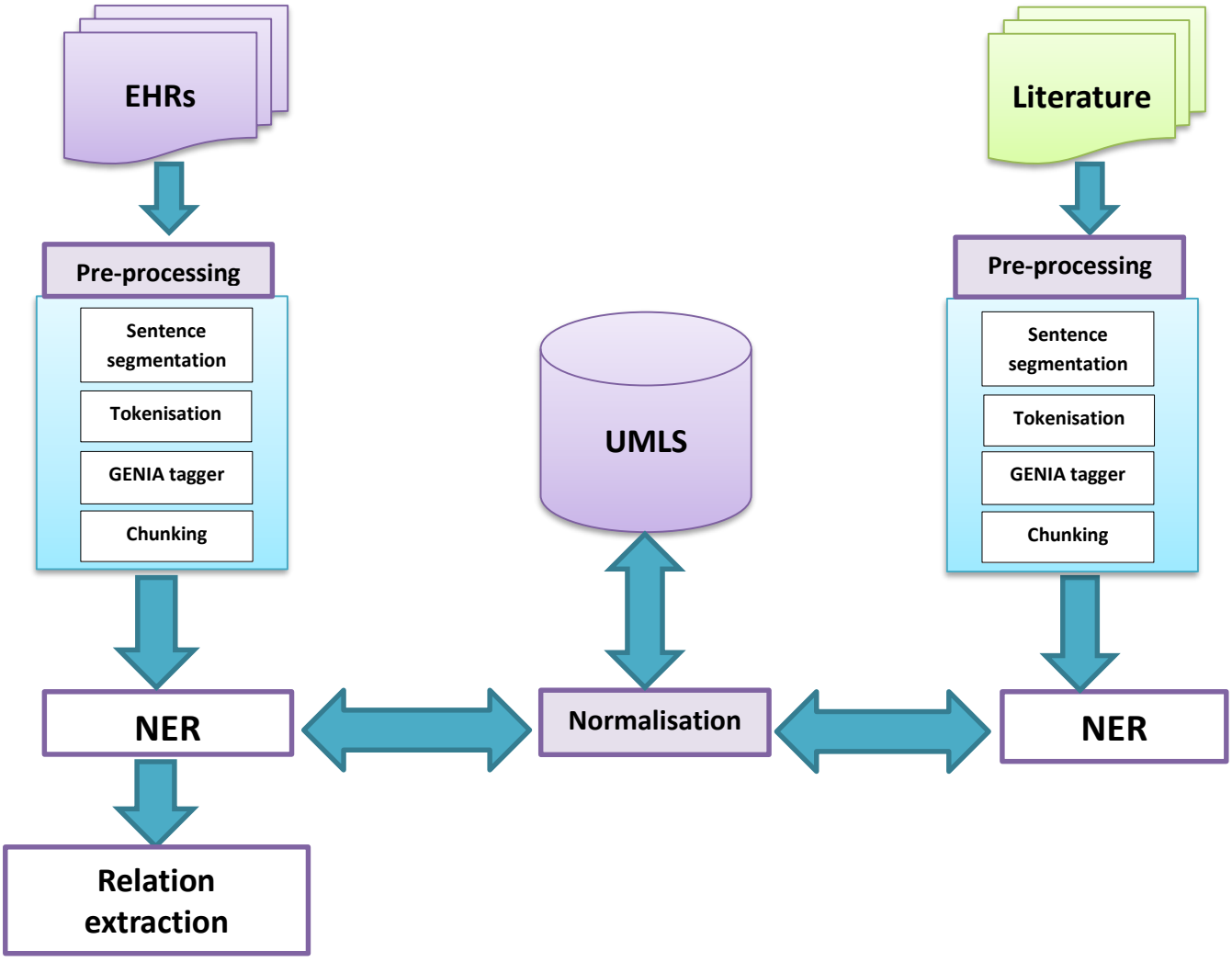


Figure 5.1 Workflow to process the PhenoCHF corpus to extract CHF phenotypic information

5.1 Phenotypic entity recognition

The recognition of phenotypic entities is a prerequisite for the extraction of more complex information (e.g., relations and events that involve these entities) and the integration of dispersed information.

The task of extracting phenotypic mentions can be seen as a typical NER task. It involves determining the boundaries of the mentions and assigning semantic types. In our case, this corresponds to the types cause, risk factor, non-traditional risk factor and sign or symptom.

As described in Chapter 2, there are several existing TM systems that aim to extract various types of semantic annotation from clinical texts, including MetaMap [63], cTAKES [56], i2b2 HITEX [207] and MedLEE [67]. These systems mainly use dictionary-based methods, which aim to map mentions of clinical concepts found in text to entries in the UMLS Metathesaurus.

The main issue with the above-mentioned systems is that they can only match mentions of concepts whose lexical form matches, or is closely related to, known variants of the concepts in UMLS. Generally, semantic variants of concepts whose lexical form is not related to existing entries cannot be handled by these systems. These problems represent potentially significant drawbacks for dictionary-based systems. Given that such dictionaries are usually manually curated, it is impossible for them to be kept up to date to cover all relevant concepts and their variants that may occur within text. This can reduce their ability to recognise all instances of categories of interest. Furthermore, it is not always the case that a particular sequence of words will denote a medical concept in all cases. The exact interpretation will sometimes depend upon the textual context. However, dictionary-based methods do not usually take context into account.

More intelligent methods of extracting entities are required, in order to capture the nuances of the expression of clinical entities in text. Terms in clinical text may vary from forms that are listed in dictionaries and their interpretation may be dependent on context. Additionally in the specific context of our work, a significant drawback of UMLS is that it does not include semantic categories that correspond directly to specific types of phenotypic information.

In the following sections, we demonstrate how the PhenoCHF corpus can be used to train ML models to recognise different types of information relating to phenotypes

automatically. We performed a range of different ML experiments, using different splits of the data in the corpus for training and testing, different algorithms and different sets of features. We compare the results obtained by the ML models to those achieved using two alternative approaches to phenotypic NER, i.e. dictionary and rule based methods.

In the first set of experiments, we considered each part of the corpus (i.e., literature articles and discharge summaries for EHRs) separately. Each of these portions of the corpus was divided into a training set (80%) and test set (20%). Thus, for the clinical discharge summaries, 240 records were used for training and 60 records were used for testing, while for the full text literature articles, 8 articles were used for training and 2 for testing. Different machine learning algorithms (HMM, Maximum Entropy Markov Model (MEMM) and CRFs) were applied to each of the training sets, and the resulting learned models were evaluated against the test set. We considered a variety of features and evaluated their contribution towards the performance of the machine learning models. In our second set of experiments, we trained models on the complete set of documents from one portion of the corpus (i.e. either EHRs or literature articles) and evaluated the performance on the other portion of the corpus (i.e. either EHRs or literature articles). The aim of these experiments was to determine the potential portability of the models to find out whether a model trained on one type of text could successfully recognise entities in another type of text. Subsequently, we carried out experiments to determine whether training on a mixture of text types (EHRs and literature articles) can result in a more robust and better performing model. Finally, we investigated the potential portability of our best performing model by applying it to different annotated corpora (i.e. ShARe/CLEF 2013, HD risk factors and COPD phenotype corpora).

5.1.1 Methodology

In this section, we firstly describe the methods applied in carrying out our baseline experiments for NER, i.e. the dictionary-based and rule-based approaches, after which we provide a detailed account of the ML methods that we applied.

5.1.1.1 Dictionary-based method

For the dictionary-based method, we applied MetaMap to PhenoCHF. MetaMap maps phrases found in free text to concepts in the UMLS Metathesaurus. Although the UMLS Metathesaurus does not have any semantic categories corresponding directly to fine-grained phenotype information, previous studies [208] have found that information relating to phenotypes usually falls under the semantic types belonging to the *disorder* semantic group in UMLS. Our own annotators have also confirmed this.

In UMLS, semantic groups consist of a number of individual semantic categories. The disorder semantic group in UMLS contains 12 semantic categories, i.e.: *Acquired Abnormality, Anatomical Abnormality, Cell or Molecular Dysfunction, Congenital Abnormality, Disease or Syndrome, Experimental Model of Disease, Finding, Injury or Poisoning, Mental or Behavioral Dysfunction, Neoplastic Process, Pathologic Function* and *Sign or Symptom*.

Given that by default, MetaMap recognises instances of all categories of the wide variety of semantic concepts that are included within the UMLS Metathesaurus, we configured the tool such that it would only recognise concepts belonging to the *disorder* semantic group. The aim was to try to reduce the number of falsely identified phenotypic concepts.

5.1.1.2 Rule-based method

For the rule-based approach, we exploited an existing system that applies pattern-matching rules to free text to facilitate the recognition of semantic information. The system, called Conceptual Annotations for Facts, Events, Terms, Individual Entities and Relations (CAFETIERE) [61], allows the formulation of rules that can match phrases (e.g., entities) using various textual features of both the phrase itself and its textual context. Through the manual examination of phenotype annotations and their contexts in the training portion of PhenoCHF, we developed a set of rules (i.e., 45 and 37 rules for records and articles respectively). These rules exploit a range of textual features including syntactic (i.e. POS), semantic (UMLS semantic type) and lexical (word shape, prefix and suffix) to capture common patterns that denote the presence of phenotype information.

Each rule can specify up to three *contexts*. The left and right contexts specify patterns that must occur either before or after the entity to be annotated (but which do

not form part of the entity), while the centre context specifies features of the tokens that form part of the entity to be annotated. The syntax for a rule is left_context \ constituents / right_context, where there may be 0 or more left or right context items and at least one constituent item. Items on the right hand side are comma-separated except before a slash. Each item, including the phrase item, is described by a set of comma-separated expressions enclosed in square brackets, where an expression has the form: feature operator value.

An example of the right hand side of a rule is shown below where it specifies a syntactic pattern to capture certain types of phrases that can describe signs or symptoms. The rule aims to annotate phrases (left hand side detail not shown) such as *heart is enlarged*, *abdomen was distended*, *leg is swollen*, i.e., phrases that describe a characteristic of a body part.

```
\
[syn=NN|NNP]{1,3}[20]?,[sem=beverb]?,[syn=VBN]
/
```

The above rule right hand side only specifies a centre context, i.e., any matching token sequence will be annotated, regardless of its surrounding textual context. The centre context consists of 3 token specifications. The first specification matches a token that is syntactically tagged as a noun or proper noun. The *{1,3}* after this token specification means that the rule will match up to three tokens with this specification. This is to allow matching with multi-word body parts, including those that include proper names as one of the tokens, e.g., *Achilles tendon.* The second token specification will match different forms of the verb *to be* (e.g., *is*, *was, were*) while the final token specification matches past participles of verbs.

An example of a slightly more complex rule right hand side is shown below. This rule aims to recognise certain types of cause phenotypic concepts, such as "*chronic obstructive pulmonary disease"*, *"mild mitral insufficiency"* and "*atrial fibrillation"*.

\

[syn=JJ|NN|NNS|VBG|NNP|VB|VBN]{0,3},

[syn=NN|NNS,token="*ension"|"pressure"|"disease|"failure"|"dysfunction"|"insufficiency"|"flutter"|"regurgitation"|"syndrome"|"fibrillation"|"infarction"]

/

Once again, the rule only specifies a centre context. The *token* feature is used to specify all or part of a token to match. The | character is used to specify alternative matching tokens, while the * character will match any sequence of characters within the token, to allow for more general token specifications. For example, **ension* will match *tension, hypertension,* etc. Thus, the rule will match phrases whose final word is one that typically denotes a cause. The first two token specifications allow the matching of longer phrases that fully describe this cause, including sequences of preceding nouns and/or adjectives.

5.1.1.3 Machine learning-based methods

We cast the problem of NER as a sequence labelling task, i.e., the automatic assignment of labels to a sequence of tokens.

The set of possible labels is defined by a chosen encoding scheme. We use the most commonly encoding representation begin-inside-outside (BIO) encoding as shown in the example in Table 5.1 on Page 103.

We carried out experiments with a number of existing ML algorithms for sequence labelling such as HMM [209], MEMM [210] and CRFs [211]. Each of which can predict the most likely sequence of labels given a sequence of words. Specifically, we have employed the CRFSuite implementation of CRF [212] and the Mallet implementation of MEMM [213]. Additionally, since the traditional HMM does not assume independent features [214], i.e., each observation is independent from its context, we adapted the approach reported in [215] to allow the integration of multiple features.

Each of the ML algorithms can use different sets of features representing different types of linguistic information relating to the input sequence of tokens. To generate

these features, a text pre-processing pipeline, which consists of a set of existing tools, as described below, is applied to the raw text.

Pre-processing pipeline

All documents in PhenoCHF corpus (EHRs and literature articles) were pre-processed with the pipeline described below.

Sentence splitting

The plain text in the PhenoCHF corpus is firstly processed using the LingPipe MEDLINE sentence splitter model [216]. This model was designed specifically to process biomedical text. It uses a set of rules to determine the sentence boundaries. The rules take into account possible endings for sentences (e.g., full stop, question mark), impossible penultimate tokens (e.g., abbreviations or acronyms such as personal titles like *Dr*) and impossible starts of sentences (e.g., percentage sign).

Tokenisation

We use the GENIA tagger [47] to segment each sentence into tokens. The maximum entropy model has been trained on both general- and biomedical-domain documents and has demonstrated robustness in tokenising biomedical text. We use the model adapted for biomedical text.

Part-of-speech and chunk tagger

The GENIA tagger has been shown to achieve the state-of-the-art performance in providing syntactic information for biomedical text (i.e., it achieves a precision of 97-98% for POS tagging on biomedical text) [47]. It takes as input tokenised sentences and outputs syntactic information for each token (i.e., POS and chunk tags). Tagging is based on a maximum entropy model trained on both general-and biomedical-domain documents [47].

Feature set

The input to the ML algorithms is formulated by combining the information in the BIO format of the gold standard annotation on PhenoCHF corpus with various types of features obtained through the application of the pre-processing pipeline described above. These features include word level information (i.e., bags-of-words), syntactic

information (e.g., part-of-speech and chunk tags) and morphological features (i.e., prefixes and suffixes of the words). Table 5.1 shows an example of the features available as input to the ML algorithms, for the following sentence: *He had coronary artery disease and myocardial infarction.*

Table 5.1 Example of features available for machine learning input, following the application of the pre-processing pipeline

BIO tag	Token	POS	Chunk
O	He	PRP	B-NP
O	had	VBD	B-VP
B-CAUSE	coronary	JJ	B-NP
I-CAUSE	artery	NN	I-NP
I-CAUSE	disease	NN	I-NP
O	and	CC	O
B-CAUSE	myocardial	JJ	B-NP
I-CAUSE	infarction	NN	I-NP

We specifically extracted the following features for potential use in our ML experiments:

Part-of-speech (POS) tags

Unigram and bigram features of POS tags within a window of two tokens before and after the active token were extracted. Given the example in Table 5.1 and using '*artery'* as the active token unigrams {*VBD*}, {*JJ*}, {*NN*}, {*NN*}, {*CC*}, and bigrams {*VBD, JJ*}, {*JJ, NN*}, {*NN, NN*}, {*NN, CC*} would be extracted.

Chunks

Unigram and bigram features of chunks tags within a window of two tokens before and after the active token were extracted. Again using '*artery'* as the active token in the example above, the unigrams {*B-VP*}, {*B-NP*}, {*I-NP*},{*I-NP*}, {*O*} and bigrams {*B-VP, B-NP*}, {*B-NP, I-NP*},{*I-NP, I-NP*},{*I-NP, O*} would be extracted.

Morphological features

We generated affix-capturing features, based on the observation that certain prefixes and suffixes are quite common amongst phenotypic expressions, e.g., the prefix *hyper-* in hyperthyroidism and the suffix *-emia* in *lipidemia* and *anemia*. Using the gold standard annotations in the training set, a registry of prefixes/suffixes of lengths two to five was automatically compiled (following [217]); this list served as a look-up list during feature extraction.

We used the training dataset to generate the most frequent prefixes and suffixes of length two to five. Then each of prefix/suffix *x* is evaluated according to the following equation:

$$Weight(x) = \frac{(\#IN_X - \#OUT_X)}{\#X}$$

Where $\#IN_x$ is the number of times that the prefix/suffix *x* occurs within a phenotypic concept; $\#OUT_x$ is the number of times that the prefix/suffix *x* occurs outside a phenotypic concept; #X is the total number of times that prefix/suffix *x* occurs within the corpus. The value *weight(x)* is highest for prefixes and suffixes that are most likely to occur inside phenotypic expressions, and which are least likely to occur outside of such expressions. Prefixes and suffixes with a weight above a certain value may thus be considered to be potentially predictors of words that are likely to form part of a phenotypic concept. Accordingly, we select candidate prefixes/suffixes with a weight above 0.70.

Table 5.2 shows the result of matching the tokens from the example sentence against our affix lists, which are provided in Appendix A.1. Tags resulting from this matching process are incorporated into the feature set.

Table 5.2 Example of tokens sequence tagged with matches against our affix lists

	Prefix				Suffix			
Token	**Size 2**	**Size 3**	**Size 4**	**Size 5**	**Size 2**	**Size 3**	**Size 4**	**Size 5**
He	O	O	O	O	O	O	O	O
Had	O	O	O	O	O	O	O	O
History	O	O	O	O	O	O	O	O
of	O	O	O	O	O	O	O	O
Heart	O	O	O	O	O	O	O	O
Disease	O	O	O	O	O	O	O	O
Anemia	O	O	O	O	O	O	emia	O
Hypertension	hy	hyp	hype	hyper	O	O	sion	nsion
Hypercholesterolemia	hy	hyp	hype	hyper	O	O	emia	O

Experiments

As described above, we carried out a number of different ML experiments. In the first set of experiments, we considered each text type of the corpus (i.e., literature articles and discharge summaries for EHRs) separately. Each portion of the corpus was divided into a training set (80%) and test set (20%). Thus, for the clinical discharge summaries, 240 records were used for training and 60 records were used for testing, while for the full-text literature articles, 8 were used for training and 2 for testing. Different ML algorithms were applied to each of the training sets, and the resulting learned models were evaluated against the test set. We considered a variety of features, and we systematically evaluated the contribution of using different features towards the performance of the ML models.

In our second set of experiments, we trained models on the complete set of documents from one portion of the corpus (i.e. either EHRs or literature articles) and evaluated the performance on the other portion of the corpus (i.e. either EHRs or literature articles). The aim of these experiments was to determine the potential portability of the models, i.e., to find out whether a model trained on one type of text could successfully recognise entities in another type of text.

Subsequently, we carried out experiments to determine whether training on a mixture of documents (both EHRs and literature articles) can result in a more robust and better performing model. For these experiments, we used 5-fold cross validation to split the data. The purpose was to reinforce our results from the first set of experiments and ensure that the trained model did not overfit the datasets.

Finally, we investigated the potential portability of our best performing model by applying it to different annotated corpora (i.e. ShARe/CLEF 2013, HD risk factors and COPD phenotype corpora).

5.2 Results and Discussion

We have evaluated the performance of both the baseline and ML methods introduced above, by calculating precision, recall and macro-averaged F-score against the gold standard annotations in the test portions of the PhenoCHF corpus.

The evaluation also compares results achieved using different matching criteria (i.e., both exact and relaxed matching). For exact matching, the start and end offsest of the predicted phenotypic entities must be the same as those in the gold standard data, wheras for relaxed matching, it is sufficient for the start and end offsets of the recognised phenotype to overlap with the gold standard.

Tables 5.3 and 5.4 show the results obtained for the experiments that were carried out separately on the two portions of the corpus (i.e., EHRs and literature articles), using 80% training and 20% tests sets introduced above. These results are also visualised in Figures 5.2 and 5.3. Results are shown for different ML algorithms, as well as for the baseline rule-based and dictionary-based methods.

The lowest performance was achieved by the dictionary-based MetaMap method. This method produced many FPs, even though we restricted the semantic types recognised to those belonging to the *disorder* group. This is partly because MetaMap recognised all disorders, regardless of whether they were concerned with CHF, and also because not everything that is a disorder can be classed as phenotypic information. However, in the gold standard, such disorders were only annotated if they were mentioned in the context of CHF. For example, *gout* is recognised as a disorder by MetaMap, but is not annotated as a phenotype term in PhenoCHF. Using MetaMap produces low recall due to spelling mistakes in the corpus (e.g., *aneamia* instead of *anaemia*), which MetaMap is not designed to handle. Additionally, a large number of phenotypic terms consist of multiple words, which MetaMap often recognises as multiple different terms. For exact matching, recall is low because MetaMap cannot recognise phenotypes which consist of adjective or modifier + medical term, i.e., "*moderate to mild cough"*.

In terms of the other methods, it can be observed that, even though the rule-based method achieved the highest F-score among all experiments, for both text types of the corpus, comparable performance is achieved by one of the ML algorithms (i.e., CRF) on the EHRs portion of the corpus. Although the ML methods achieve considerably lower performance than rule-based methods on the literature part of the corpus, it is much smaller than the EHRs part of the corpus, and therefore the training data available is much sparser. This means that machine learning models are unlikely to have sufficient evidence to predict phenotype information accurately in the unseen test set.

The high performance of the rules suggests that it is possible to formulate general patterns that denote the existence of phenotypic information. However, the rule-based approach does have several drawbacks. Firstly, it is a very labour–intensive and time consuming process. The human rule-writer must carefully read many documents and generalise the textual patterns denoting the probable occurrence of concept mentions. Additionally, given that these patterns can vary between text types, it will usually be necessary to create a new set of rules for each different text type to be considered. Indeed as has been explained above, we needed to create roughly equal sized (but different) sets of rules for the EHRs and literature data sets. Thus, portability of this method is a large issue. In contrast, given an appropriate annotated training corpus, ML algorithms can generalise patterns many times more quickly.

An analysis of the errors produced by the rules revealed that most of the FNs occur because the test set contains forms of phenotype annotations that were not present in the training set, and thus the rules do not cover them. E.g., "*mitral valves mildly thickened"* and *"increased swellings in both hands"*. The source of several FPs is that some of the rules are not sufficiently restrictive as some non-phenotype terms share similar syntactic patterns to phenotypic terms. For example, consider the following rule:

```
\
[syn=NN|NNP]{0,3},[sem=beverb]?,[syn=VBN|JJ,token!=
"normal"|"regular"|"stable"]
/
```

The phrase "*abdomen is benign"* is incorrectly recognised by this rule as a *sign or symptom*, because many signs or symptoms are expressed using a similar syntactic pattern i.e., "*abdomen is distended"*. From this rule, it can be appreciated that we tried to address such examples to a certain extent, since the specification of the last token in the central context means that it will not match the words *normal, regular* or *stable.* These restrictions aim to filter out terms that refer to normal conditions, e.g. "*chest is normal"* and "*heart is regular"*. However, once again due to examples in the test data that do not occur in the training data, we were unable to filter out all such FPs.

The CRF algorithm achieves the highest F-score amongst the evaluated machine learning methods, followed by HMM and MEMM. CRF also achieves a greater balance between precision and recall than the other algorithms. Overall, our experiments demonstrate that ML methods exhibit good levels of precision but lower recall. This is partly due to the fact that machine learning algorithms are sensitive to textual heterogeneity, such as the use of different vocabulary and different writing styles [218].

Table 5.3 Comparative evaluation of different machine learning methods on the discharge summary (EHRs) set, for MEMMs, HMMs and CRFs. Only the results from the model with the best performing combination of features are presented

	Exact Match			**Relaxed Match**		
Methods	**P**	**R**	**F**	**P**	**R**	**F**
MetaMap	0.22	0.29	0.25	0.39	0.51	0.44
Rules	0.88	0.86	**0.87**	0.92	0.93	**0.92**
MEMMs	0.67	0.33	0.52	0.87	0.60	0.54
HMMs	0.90	0.63	0.74	0.90	0.65	0.76
CRFs	0.88	0.77	0.82	0.90	0.86	0.88

Table 5.4 Comparative evaluation of different machine learning methods on literature articles set, for MEMMs, HMMs and CRFs. Only the results from the model with the best performing combination of features are presented

	Exact Match			**Relaxed Match**		
Methods	**P**	**R**	**F**	**P**	**R**	**F**
MetaMap	0.42	0.25	0.30	0.67	0.33	0.44
Rules	0.83	0.88	**0.85**	0.88	0.90	**0.89**
MEMMs	0.18	0.55	0.24	0.20	0.56	0.28
HMMs	0.30	0.55	0.39	0.32	0.58	0.41
CRFs	0.48	0.62	0.54	0.53	0.69	0.60

For each algorithm, we have evaluated the effects of using different feature sets in conjunction with the different ML algorithms. This is shown in Table 5.5 for EHRs and Table 5.6 for the literature articles. Compared to the bags-of-words (BOW) baseline, it can be observed that all additional features contribute towards improving performance. POS features appear to contribute most towards improving the precision, with the addition of chunk features usually contributing little to improving the overall performance. However, adding prefix and suffix features results in an additional small boost in F-score in all cases, such that the highest performance is achieved when all 3 sets of features are used in addition to BOW.

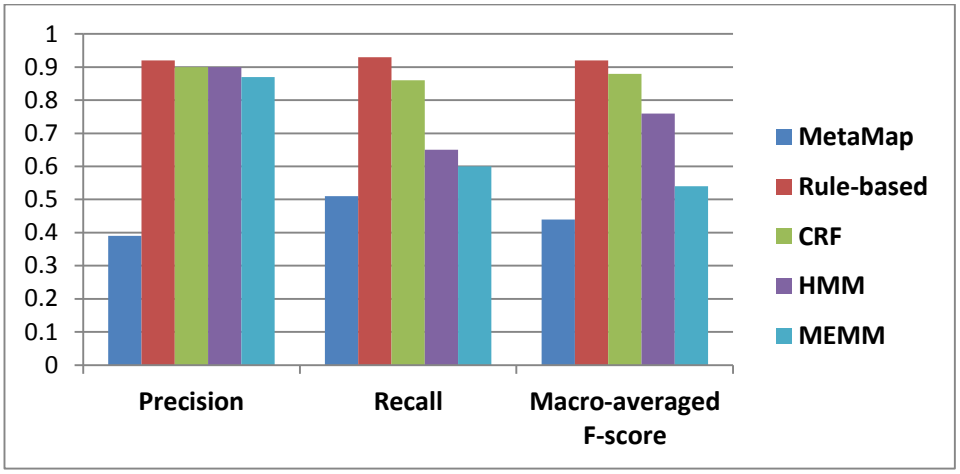


Figure 5.2 Visualisation of comparative evaluation of NER methods on discharge summaries

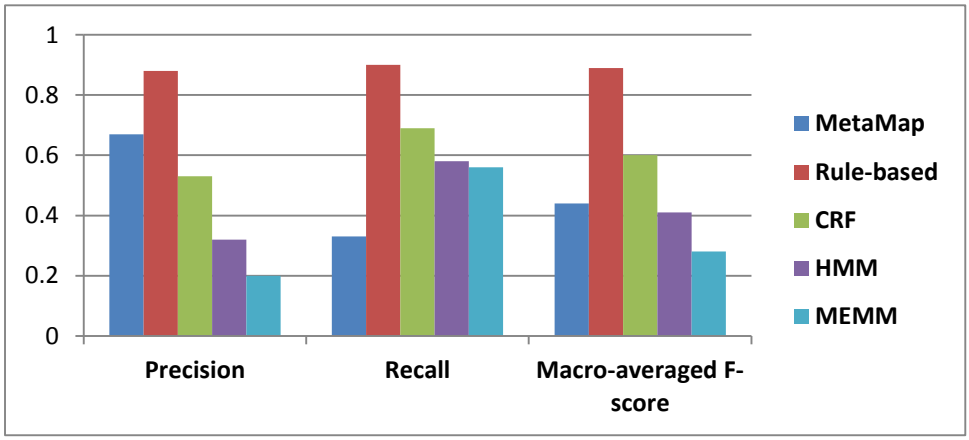


Figure 5.3 Visualisation of comparative evaluation of NER methods on articles

Table 5.5 the contribution of features in each machine-learning based method on discharge summaries

Method	P	R	F
HMM(BOW) baseline	0.86	0.59	0.69
HMM (BOW+POS)	0.89	0.62	0.73
HMM (BOW+POS+CHUNK)	0.88	0.62	0.73
HMM (BOW+POS+CHUNK+Prefix & Suffix	0.90	0.63	0.74
MEMM(BOW) baseline	0.57	0.43	0.49
MEMM (BOW+POS)	0.61	0.47	0.53
MEMM (BOW+POS+CHUNK)	0.56	0.51	0.53
MEMM (BOW+POS+CHUNK+Prefix & Suffix	0.67	0.33	0.52
CRF(BOW) baseline	0.84	0.75	0.78
CRF (BOW+POS)	0.87	0.72	0.80
CRF(BOW+POS+CHUNK)	0.89	0.75	0.81
CRF (BOW+POS+CHUNK+Prefix & Suffix	0.88	0.77	0.82

Table 5.6 The contribution of features in each machine-learning based method on literature articles

Method	P	R	F
HMM(BOW) baseline	0.28	0.41	0.32
HMM (BOW+POS)	0.27	0.54	0.36
HMM (BOW+POS+CHUNK)	0.29	0.41	0.37
HMM (BOW+POS+CHUNK+Prefix & Suffix	0.30	0.55	0.39
MEMM(BOW) baseline	0.49	0.11	0.18
MEMM (BOW+POS)	0.24	0.21	0.22
MEMM (BOW+POS+CHUNK)	0.22	0.23	0.22
MEMM (BOW+POS+CHUNK+Prefix & Suffix	0.18	0.55	0.24
CRF(BOW) baseline	0.54	0.38	0.45
CRF (BOW+POS)	0.57	0.47	0.51
CRF(BOW+POS+CHUNK)	0.57	0.47	0.51
CRF (BOW+POS+CHUNK+Prefix & Suffix	0.48	0.62	0.54

The top part of Table 5.7 shows the results of our experiments in which we trained models on one part of the corpus (i.e. either literature articles or EHRs) and tested the model on the other part of the corpus. According to the results of the previous experiments, training was carried out using the CRF algorithm and all features.

These results show that the CRF model trained on the discharge summaries performs with reasonable accuracy on the literature articles. The level of precision is

particularly encouraging, given the good size of the discharge summaries part of the corpus and the fact that the annotations provide evidence of the wide range of different ways in which phenotypic information can be expressed. A model trained on this class of documents has the potential to perform well when applied to other text types. In contrast, the model trained on the literature articles demonstrates much less potential for portability, as evidenced by the results obtained when it is applied to the discharge summaries. However, this result is to be expected according to the results obtained in the previous experiments. If there is insufficient training data for a literature-trained model to perform with high accuracy when applied to other literature articles, then it is understandable that the performance is even lower when the model is applied to documents with different characteristics.

However, the above results do not mean that the literature part of the corpus is not useful at all for training. As shown in the bottom row of Table 5.7, we carried out a 5-fold cross validation experiment over the combined corpus of clinical records and articles, in which each fold consists of data from both articles and records. The results are higher than any of the experimental results shown in Tables 5.3 and 5.4, in which training and testing was carried out only on a single part of the corpus. Although the experimental setup is different (i.e., cross validation rather than a training and test set), the results in Table 5.7 strongly suggest that training on information from multiple text types can result in a classifier that is not only robust to heterogeneous text types, but which can perform with higher accuracy than if only a single text type is used for training. This result is in contrast to other studies (e.g., [219]), which have reported that pooling corpora of different text types decreases the performance of the trained model. However, the difference in our case is that the two portions of the corpus were annotated according to a common set of guidelines.

Table 5.7 Results of CRF model training and evaluation on different document types

Evaluation data	**Training Data**	**P**	**R**	**F**
PhenoCHF discharge summaries	**articles**	0.79	0.47	0.58
PhenoCHF articles	**discharge summaries**	0.56	0.29	0.38
PhenoCHF (full) 5-fold cross validation		0.89	0.83	0.86

5.3 Evaluation

In order to further demonstrate the portability and utility of the best performing CRF model trained on our corpus, we evaluated the PhenoCHF model that is trained on the full PhenoCHF corpus (i.e., EHRs and the literature articles) on three of the gold standard corpora reviewed in Section 2.2.1, namely the ShARe/CLEF 2013 task1, i2b2 heart disease, and the COPD phenotypic corpus as presented in the Table 5.8.

In the following sections we provide an analysis of the results we obtained by applying the PhenoCHF model on the three above-mentioned corpora.

5.3.1 The ShARe/CLEF 2013 task1

As reviewed in Section 2.2.1, the ShARe/CLEF 2013 task 1 corpus consists of 300 clinical records split into a training set of 200 records and a test set of 100 records. The records were annotated for mentions of disorder terms, which were mapped to corresponding concepts in the SNOMED CT terminology.

As in UMLS, SNOMED CT concept types are organised into semantic groups, one of which is *disorder*. The SNOMED CT *disorder* semantic group corresponds almost exactly to the UMLS *disorder* semantic group, in that it consists of the following semantic types: *Congenital Abnormality, Acquired Abnormality, Injury or Poisoning, Pathologic Function, Disease or Syndrome, Mental or Behavioural Dysfunction, Cell or Molecular Dysfunction, Experimental Model of Disease, Anatomical Abnormality, Neoplastic Process* and *Sign or Symptom*. The only difference is that while the UMLS *disorder* group includes the *Finding* semantic type, the corresponding SNOMED CT *disorder* group does not. As has been previously discussed in the context of UMLS, therefore, the fact that the *disorder* group covers many types of phenotypic information means that there is some level of overlap in the annotated entities within ShARe/CLEF 2013 and PhenoCHF. However an important difference between the annotations is that, whilst in PhenoCHF annotated information is restricted to that concerning CHF, the annotations in the ShARe/CLEF corpus correspond to all instances of terms relating to disorders. Accordingly, the disease-specific models trained on PhenoCHF cannot be expected to recognise all the disorders annotated in the ShARe/CLEF corpus.

In order to provide a fair comparison of the ability of the PhenoCHF model to recognise disorder information in the ShARe/CLEF corpus, we only applied our model to a subset of the records in the corpus, i.e., those which are specifically concerned with heart disease. We created this subset by finding records in the ShARe/CLEF corpus which contained annotated terms relating to heart disease disorders. We exploited the hierarchical structure of concept classes in SNOMED CT, retaining only those records (135 from the training set and 76 from the test set) containing annotations that could be mapped to concepts within the heart disease subtree of SNOMED CT.

As can be seen in Table 5.8 on Page 120, the results are somewhat lower than those achieved when applying our model to the PhenoCHF corpus. Whilst this could be partly due to the fact that the ShARe/CLEF corpus includes reports of different types than those included in the PhenoCHF corpus, we know that it is also due to the differing annotation scopes of the ShARe/CLEF and PhenoCHF corpora. Although we tried to restrict our evaluation to only those records in the ShARe/CLEF corpus that cover a similar subject to documents in PhenoCHF, it is still the case that *all* instances of disorders within these documents are annotated, rather than only those that specifically relate to CHF.

Thus, in order to provide a more realistic estimate of the ability of our model to recognise information relating to CHF in the ShARe/CLEF subset, we asked our expert annotators (doctors) to review both the FPs and FNs that were output by the PhenoCHF model. They were asked to identify how many of the FPs output by the PhenoCHF model (in comparison to the ShARe/CLEF annotations) actually represent valid phenotypic information, and how many of the FNs represent information that is beyond the scope of CHF (and hence could not be expected to be recognised by our model).

The results of this expert annotation revealed that the majority of FPs recognised by our model represent valid phenotype information in the context of CHF, and are recognised by the PhenoCHF model according to the wider range of semantic types that are annotated in PhenoCHF, compared to the ShARe/CLEF corpus. In particular our *sign or symptom* category partly corresponds to the UMLS *Finding* semantic type (e.g., "*chest pain"*), which is not within the scope of the ShARe/CLEF corpus. We found that many *sign or symptom* annotations occur within the echocardiogram and radiology reports of the ShARe/CLEF corpus. A smaller number of FPs were found to

be genuine errors made by the PhenoCHF model, but these were found to correspond largely to cases where non-phenotype terms share the same morphological form as correct phenotype terms, and so they were incorrectly recognised by our model. As an example, the suffix *-uria* is common amongst phenotype information related to CHF, especially non-traditional risk factors (e.g., "*dysuria"*). However, the same suffix is sometimes used in non-phenotypic terms, e.g., "*cystinuria"*.

The FNs were mainly due to the broader scope of ShARe/CLEF annotation, compared to the very focused scope of PhenoCHF. Accordingly, for example, *endocarditis* is annotated as a disorder in the ShARe/CLEF corpus, but was not recognised by the PhenoCHF model because it is beyond the scope of phenotypic information related to CHF. A further source of error is acronyms and abbreviations. Although there are many such examples in PhenoCHF, such as "*CHF*" for '*Congestive Heart Failur*e', "*CAD"* for '*Coronary Artery Disease'* and "*MR"* for '*Mitral Regurgitation'*, there is a proliferation of different acronyms in the ShARe/CLEF corpus, e.g., "*PAFIB", "LBBB", "CHB", "PDA"*. This is partly due to the fact that the corpus was specially designed to allow the evaluation of acronym recognition and resolution, in addition to the more general task of disorder recognition. Given that many of the acronyms in the ShARe/CLEF corpus correspond to disorders that our model is not trained to recognise, many of them were missed.

When we remove the FPs that correspond to real phenotypic information, as well as the FNs that are beyond the scope of our task, the precision and recall and F-score are 0.68, 0.73 and 0.68 respectively, for ShARe/CLEF, for the model trained on the complete PhenoCHF corpus (i.e., PhenoCHF).

Phenotypic information related to CHF in the ShARe/CLEF corpus is shown in figure 5.4 and as it appears the most prevalent phenotype is sign or symptom followed by cause. The least prevalent was non-traditional risk factor.

As CHF phenotypic information falls within the scope of the *disorder* semantic group, it was also in our interest to test the performance of the CRF model trained on the ShARe/CLEF corpus, which includes a broad coverage of information belonging to the *disorder* semantic group, on the PhenoCHF corpus. For this purpose we trained the CRF model (i.e., ShARe model) using an identical set of features that were used to train the PhenoCHF model as presented in Section 5.1.1.3, and we evaluated the ShARe model to extract phenotypic information from the PhenoCHF corpus. The result of the ShARe model on the PhenoCHF corpus is presented in Table 5.8.

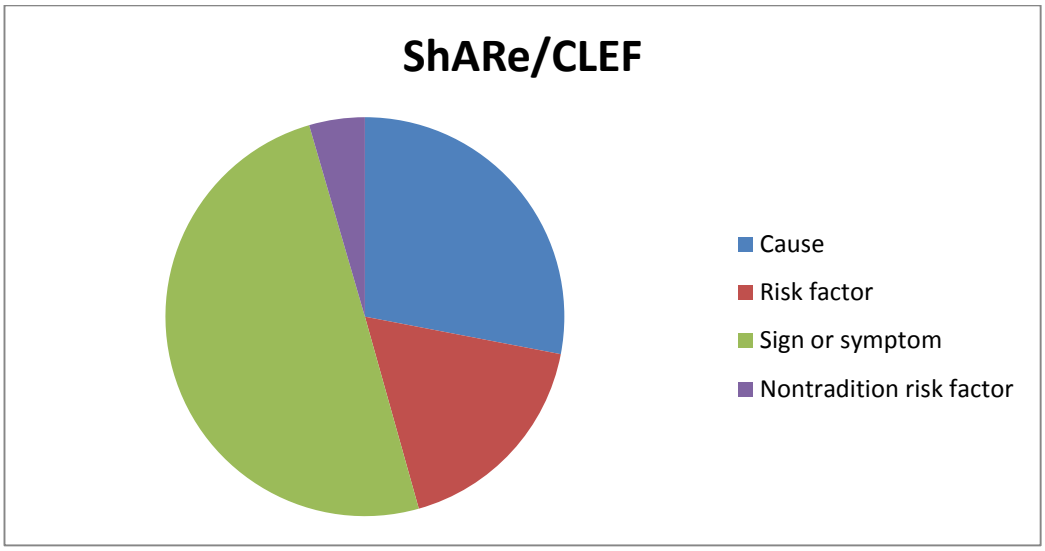


Figure 5.4 The distribution of the types of phenotypic concepts relating to CHF in the ShARe/CLEF corpus

As can be observed, the ShARe model achieved higher recall than precision; this is largely due to the broad scope of the annotations within ShARe/CLEF making a model trained on this corpus robust enough to recognise most of the CHF phenotypic entities. The manual analysis of FNs has shown that most of the entities missed by the ShARe model belong to the UMLS *Finding* semantic group, which is beyond the scope of the ShARe/CLEF annotations; hence the ShARe model exhibits difficulties in recognising the wider variability in the syntactic structure in which finding information can be expressed when compared to *disorder* instances in ShARe/CLEF (e.g., creatinine is elevated, weakness in the extremities, and grandmother and aunt died of sudden cardiac death).

Meanwhile, the main source of FPs is the broader scope of annotations within ShARe/CLEF in comparison to the very focused scope of PhenoCHF. Although we restricted our experiments to those records in the ShARe/CLEF corpus that cover heart disease information, the ShARe model still recognises all disorder entities in the PhenoCHF corpus, and the majority of FPs produced by the ShARe model are correct instances of disorders. However, the FP entities are beyond the scope of CHF and hence counted as FPs (e.g., "*platelet dysfunction*", "*lung cancer*", "*hyperplasia*" and "*pancreatitis*").

Overall, the performance of the PhenoCHF model on the ShARe/CLEF corpus achieved a better result in comparison to the result of the ShARe/CLEF model when

applied on the PhenoCHF corpus. However our results suggested that phenotypic information is covered by the UMLS *disorder* semantic group, which includes all the semantic types under the SNOMED CT *disorder* semantic group in addition to the *Finding* semantic type. We believe that adding *Finding* annotations on top of ShARe/CLEF annotations will bring wider coverage of phenotypic information in the ShARe/CLEF corpus and make it a very useful resource for phenotypic information extraction.

5.3.2 i2b2 heart disease corpus

As reviewed in Section 2.2.1, the i2b2 heart disease corpus (HD) is a collection of 1,304 records split into a training set of 790 records and a test set of 514 records annotated for heart disease risk factors including obesity, CAD, hypertension, hyperlipidaemia, diabetes, smoking, family history of premature CAD and any medication used to treat the risk factors or indicators.

Manual analysis of annotation in the HD risk factors corpus revealed that the PhenoCHF and the HD risk factors corpora are partially overlapping in their annotation scopes (i.e., heart disease risk factors). More specifically, the PhenoCHF corpus shares the HD risk factors corpus with all risk factor entities excluding medications. However, the two corpora differ in semantic categories. For example, the HD risk factors corpus includes seven classes (obesity, CAD, hypertension, hyperlipidaemia, diabetes, smoking, family history), while the PhenoCHF corpus annotates these risk factors with different semantic categories. For example, obesity, hyperlipidaemia, diabetes and family history are annotated as risk factors in PhenoCHF, while CAD and hypertension are annotated as a cause.

To provide a fair evaluation of the PhenoCHF model against the HD risk factors corpus, we removed the annotation for medication mentions from the HD risk factors corpus, and then applied the PhenoCHF model to extract heart disease risk factors from the i2b2 HD risk factors test dataset as shown in Table 5.8.

Error analysis of the results produced by our PhenoCHF model revealed that the majority of FPs represent valid phenotype information in the context of CHF. In particular, the *sign or symptom* category, which is not within the annotation scope of the i2b2 HD risk factors corpus, represents the largest portion of FP predictions. The reason for the large amount of *sign or symptom* information being recognised within the HD risk factors corpus is because two thirds of the records in the corpus are either

for patients who have been diagnosed with CAD or who developed CAD over the course of their records. Therefore, *signs or symptoms* related to heart diseases are very common in the records. A smaller number of FPs originated from risk factor mentions that are either related to a member of the patient's family (e.g., "*father developed coronary artery disease"*) or negated (e.g. "no history of hypertension").

The FNs were mainly due to the differences between the annotation guidelines for the PhenoCHF and HD risk factors corpora. While the annotations in PhenoCHF corpus include only the explicit mentions of the disease names that represent CHF phenotypes, the five disease classes (CAD, diabetes, obesity, hyperlipidaemia, hypertension) that represent risk factors in the HD risk factors corpus are annotated through either explicit mention of the disease or the indicator of the disease. Different diseases have a different number of indicators. For example, obesity has two indicators — Body Mass Index (BMI) and Waist Circumference (WC). Diabetes has two indicators — haemoglobin levels above 6.5 and glucose levels over 126. Some indicators are challenging and are expressed in long text spans, such as CAD test result indicators (e.g., "*MIBI was read as positive for moderate to severe inferior ischaemia"*, "*stress test with MIBI imaging that perhaps showed an abnormality"*). Since the PhenoCHF model is not trained to extract such information (i.e., disease indicators) the PhenoCHF model failed to recognise some risk factors. Machine learning algorithms are sensitive to textual heterogeneity, which also affected performance. Examples of missed risk factors are: "177/90" and "blood sugar was noted to be greater than 600".

The distribution of the types of phenotypic concepts relating to CHF in the HD risk factors corpus, recognised by PhenoCHF model, is shown in Figure 5.5. It is worth noting that the most prevalent phenotypic type is sign or symptom followed by cause and risk factor, whilst the least prevalent type is non-traditional risk factor.

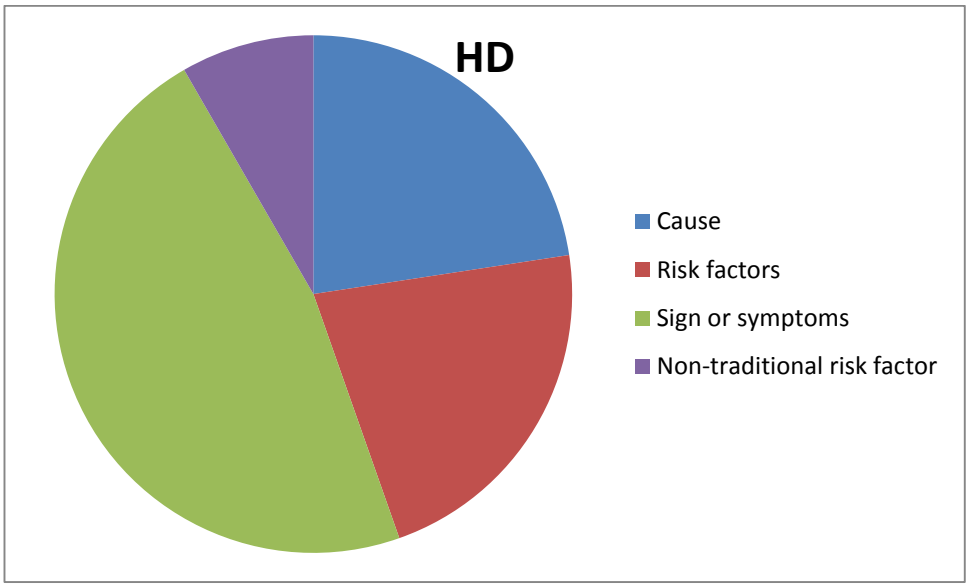


Figure 5.5 The distribution of the types of phenotypic concepts relating to CHF in the HD corpus

Manual analysis of the results obtained by applying our PhenoCHF model on the HD risk factors corpus revealed that the scope of risk factors information falls within the cause and risk factors phenotypic classes in the PhenoCHF corpus. We further tested the performance of the CRF model trained on the HD risk factors corpus, namely the HD model, using the combination of features presented in Section 5.1.1.3 to extract cause and risk factors entities from the PhenoCHF corpus. The results are shown in Table 5.8. The HD model achieved high precision in extracting cause and risk factor entities related to the seven classes in the HD risk factors corpus (obesity, CAD, hypertension, hyperlipidaemia, diabetes, smoking, family history). However, the HD model achieved low recall as it failed to recognise causes and risk factors that are not related to the five classes (e.g., the causes "*chronic kidney disease*", "*uremic cardiomyopathy*" and the risk factors "*peripheral vascular disease*", "*asthma*", "*stroke*"). The main source of FPs is due to the differences in the annotation guidelines between the two corpora as explained above. For example, the HD trained model recognised "blood glucose of 151" as a risk factor because it is an indicator of diabetes. Whereas this annotation is not included in PhenoCHF, therefore it is counted as a FP in the result.

5.3.3 COPD phenotypic corpus

We also evaluated the PhenoCHF model on the COPD-phenotype corpus by Batista-Navarro et al. [176]. As described in Section 2.2.1. The COPD corpus is annotated for

four phenotypic entities related to COPD (medical conditions, signs or symptoms, proteins, and drugs). COPD and CHF are commonly prevalent co-morbidities, hence COPD mention is annotated as a cause for CHF in the PhenoCHF corpus. Therefore, the two corpora (COPD and PhenoCHF) overlap in many phenotype entities. However, the PhenoCHF corpus is not annotated for drug and protein entities. To provide a fair evaluation of our PhenoCHF model on the COPD corpus we removed the annotations pertaining to drugs and proteins. We tested the performance of the PhenoCHF model to extract information related to medical conditions and signs or symptoms.

As shown in Table 5.8, the results revealed that the PhenoCHF model was able to recognise most medical conditions and signs or symptoms information. This is not surprising, considering what we mentioned earlier: CHF and COPD are comorbidities and hence there are many phenotypes shared between the two diseases. More specifically, the COPD phenotypes (medical condition, signs or symptoms) fall within the scope of the following phenotypic classes: cause, risk factor and sign or symptom in the PhenoCHF corpus. As can be observed from Table 5.8, the PhenoCHF model achieved high precision and low recall, largely due to the phenotypic entities that are specifically related to COPD (e.g., muscle atrophy, rhinovirus infection, cystic fibrosis and skeletal muscle dysfunction) and are not related to CHF. By looking at the FP results we noticed that most of the FPs are correct examples of phenotypes in the context of CHF as annotated in the PhenoCHF corpus. We also believe that some of the FP results which comprise correct CHF phenotypes are also related to COPD (e.g., reduced oxygen delivery, fluid retention, reduced exercise capacity, fluid retention and reduced mixed venous oxygen saturation). But the annotations were missing in the COPD corpus due to differences in the definitions of semantic types between the two corpora. For example, *medical condition* in the COPD corpus typically contains mentions of diseases related to COPD or its comorbidities, while *signs or symptoms* are often composed of any observable irregularity manifested by a COPD patient. In the PhenoCHF corpus, however, the semantic category *cause* refers to any disease that directly contributes to cause CHF. *Risk factor* contains any mention of a condition that put the patient at high risk to develop CHF. Sign or symptom category refers to any observable manifestation of CHF.

The distribution of the types of phenotypic concepts relating to CHF in the COPD corpus, recognised by the PhenoCHF model is shown in Figure 5.6. It is worth noting

that the most prevalent phenotypic type is cause followed by sign or symptom, whilst the least prevalent types are risk factor and non-traditional risk factor.

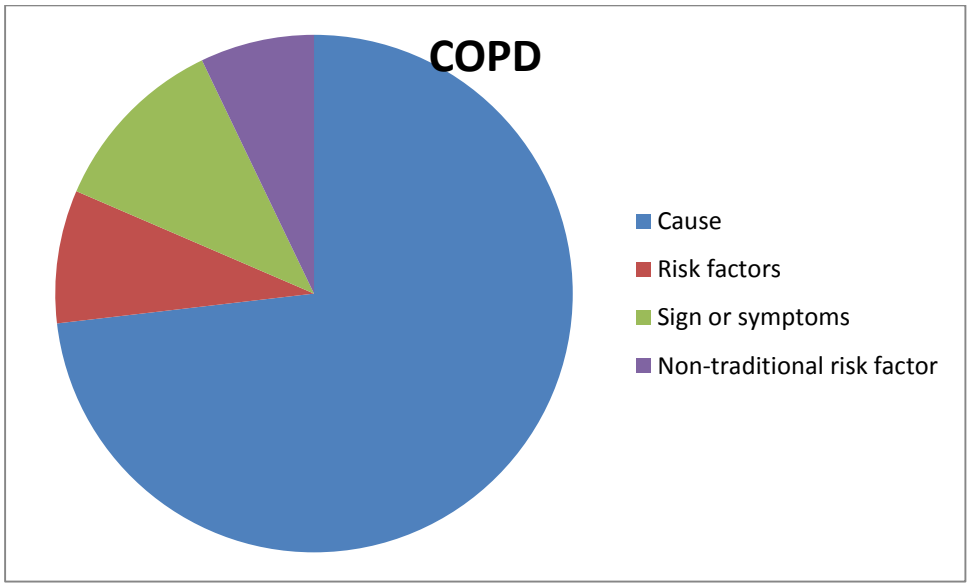


Figure 5.6 The distribution of the types of phenotypic concepts relating to CHF in the COPD corpus

It is also of interest to see the performance of the CRF model trained on the COPD corpus to extract cause, risk factor and sign or symptom entities from the PhenoCHF corpus. For this purpose, we trained a CRF model using the combination of features presented in Section 5.2.1.3 on the COPD corpus. We applied the model to extract cause, risk factor and sign or symptom information from the PhneoCHF corpus. As can be observed from Table 5.8, the COPD model achieved lower precision and recall in comparison with the PhenoCHF model. This can be explained by the richer and wider coverage of annotations in the PhenoCHF corpus when compared with the annotations in the COPD corpus. Therefore, the COPD model exhibits lower recall when applied to PhenoCHF. The low recall achieved by the COPD model is also due to the heterogeneous documents that comprise the PhenoCHF corpus. The PhenoCHF corpus is a combination of EHRs and literature articles, where EHRs represent the largest portion of the PhenoCHF corpus. However, the documents in the COPD corpus consist of only literature articles; therefore, the COPD model exhibits difficulties in recognising phenotypic entities with complex characteristics and a variety of syntactic structure in the EHRs portion of the corpus (e.g., mild concentric left ventricular hypertrophy, mitral valve was thickened). It is also observed that while most of the

FPs are correct *signs or symptoms* in the context of COPD (e.g., "*cancer"*, "*cystic fibrosis"*, "*mortality and death"*), the COPD model produced many spurious negative results which are not related to COPD or CHF (e.g., "*glaucoma"*, "*osteoarthritis"* and "*pancreatitis"*).

Table 5.8 Comparative evaluation of PhenoCHF model on overlapping corpora: ShARe/CLEF, HD risk factors risk factors and COPD phenotype corpora. Corpus refers to the data that was used for testing and model refers to the data that was used for training

Corpus	**Model**	P	R	F
HD	**PhenoCHF**	0.57	0.62	0.59
PhenoCHF (cause and risk factor annotations)	**HD**	0.85	0.36	0.51
COPD	**PhenoCHF**	0.85	0.68	0.75
PhenoCHF(cause and sign or symptom)	**COPD**	0.74	0.51	0.60
ShARe/CLEF	**PhenoCHF**	0.32	0.71	0.44
PhenoCHF	**ShARe**	0.30	0.54	0.40

5.3.4 Results

We evaluated our best performing model, PhenoCHF, on three overlapping corpora. We also leveraged these overlapping corpora to create different models concerned with phenotypic extraction and evaluated them in various ways. The results of these tests are shown in Table 5.9.

The results we obtained are encouraging, considering that each corpus has unique characteristics. For example, ShARe/CLEF contains clinical records of various types (discharge summaries and electrocardiogram, echocardiogram, and radiology reports), COPD consists of literature articles, and the HD risk factors corpus is characterised by its annotations for explicit disease mentions (e.g., "*diabetes"*) or its indicators (e.g., "*high level of glucose"*) as well as long annotation span (e.g., catheterisation showed multi vessel non-obstructive CAD). The result also showed the value of PhenoCHF as a resource to train NER models that take full advantage of the evidence of different means of expressing phenotypic information and different writing styles in order to deal with the different characteristics present in different corpora.

Finally, we carried out experiments to determine whether training on overlapping corpora with shared semantic types can result in a more robust and better performing model. For this purpose, we trained several CRF models, combining the semantic type

annotations that are shared across different corpora. Since the *heart disease risk factors* in the HD corpus overlap with the annotations under the cause and risk factors categories in the PhenoCHF corpus, we merged the cause and risk factors annotations from PhenoCHF with risk factors annotations from the HD risk factors corpus. For reasons of fairness, the semantic categories in the HD risk factors corpus have been replaced with cause or risk factor categories based on PhenoCHF guidelines as follows: the categories CAD and hypertension are replaced with cause, while the categories hyperlipidaemia, diabetes, smoking, and family history are replaced with risk factor. Similarly, medical condition and signs or symptoms annotations are merged with the overlapping subset of annotations in PhenoCHF (cause, risk factor and sign or symptom). While it was straightforward to replace categories in the HD risk factors and COPD corpora to the PhenoCHF categories, the entities in the ShARe/CLEF corpus are annotated under one general class (i.e., disorder) and the disorder information is scattered within the PhenoCHF categories, which makes it difficult to replace the disorder category into PhenoCHF categories. For reasons of fairness we merged the full PhenoCHF corpus with the subset of ShARe/CLEF which is concerned with heart diseases and unified the semantic categories within the merged corpora of ShARe/CLEF and PhenoCHF into the *phenotype* category.

Using the following merged corpora: ShARe/CLEF+PhenoCHF, HD risk factors+PhenoCHF, COPD+PhenoCHF, we trained and evaluated single CRF models using 5-fold cross validation, where each fold consists of data from both merged corpora (e.g., ShARe/CLEF and PhenoCHF). The macro-averaged evaluation results for the three merged corpora are presented in Table 5.9.

Merging corpora allows the training of a wide-coverage, state-of-the-art phenotypic extraction model from multiple corpora with partial semantic annotation overlap. The degradation in the results of training a single model from partially overlapping corpora is due to the creation of spurious negative instances from one corpus for cases that correspond to positive instances in terms of the scope of another corpus.

Having observed that the recall of the CRF models on the three corpora is significantly lower than precision, we tabulated their recall scores for each phenotypic subtype (Table 5.10) in COPD+PhenoCHF and HD risk factors+PhenoCHF merged corpora. This allowed us to identify the subtypes which are most difficult for NER models and, hence, are pulling down the overall recall. It can be observed that the most challenging subtypes for the NER model trained on merged corpora are *cause*

for the COPD+PhenoCHF corpus and *risk factor* for the HD risk factors+PhenoCHF corpus. This can be attributed to the sparsity of annotations under these two categories. The sparsity prevented the CRF models from learning the relevant features, leading to low recall for these types.

Table 5.9 Results for 5-fold cross validation over the merged corpora

	P	**R**	**F1**
ShARe/CLEF+PhenoCHF	0.81	0.76	0.78
HD+PhenoCHF	0.77	0.57	0.66
COPD+PhenoCHF	0.88	0.80	0.83

Table 5.10 Recall scores for each phenotypic subtype in the HD risk factors+PhenoCHF and COPD+PhenoCHF corpora

Merged corpus	**Cause**	**Risk factor**	**Sign or symptoms**
HD+PhenoCHF	0.60	0.53	-
COPD+PhenoCHF	0.78	0.81	0.81

5.4 Summary

In this chapter, we have demonstrated how the PhenoCHF corpus can be successfully used in the development of systems targeted at the extraction of information relating to CHF and CKD, through the application of different techniques to extract phenotypic entities.

To demonstrate the value of PhenoCHF in developing ML NER systems, we compared the results obtained by training different ML models against other baseline methods, i.e., those based on dictionary lookup and hand constructed pattern matching rules. Our results showed that ML methods can significantly outperform dictionary-based methods, and that the best performing ML algorithm, i.e. CRF, compares favourably to rule-based methods when sufficient training data is available, especially taking into account the amount of manual work saved when ML techniques are applied.

In terms of our ML results, we also systematically demonstrated that a ML tagger trained to recognise phenotypic information in one type of text is fairly robust to changes in document types. We also found that training a model on a pooled corpus consisting of two different documents types, but which are annotated using a common set of guidelines, can result in a model with both improved performance and greater robustness/portability when applied to different document types. This was demonstrated by applying such a model to a corpus with overlapping scope to PhenoCHF, i.e., the ShARe/CLEF, HD risk factors and COPD phenotype corpora.

Chapter 6 Phenotypic relation extraction

Relation extraction systems that have the ability to detect more than two arguments have the potential to stimulate a qualitative advance in medical information extraction and enrich existing medical knowledge resources and databases with new and more complex relations and associations.

In this chapter we will discuss the automatic extraction of *n*-ary relations between phenotypic entities from EHRs subset of the PhenoCHF corpus.

6.1 Relation extraction

In chapter 5 we described an automatic method to extract phenotypic entities from both EHRs and literature articles. The extracted phenotypic entities represent a rich source of information that can contribute to individual patient care [12]. For example, they can be used to provide clinicians with a summary of the medical history of a patient, and they identify patient-specific characteristics which can be used to determine a suitable personalised treatment plan to a patient with CHF.

Furthermore, the automatic extraction and classification of phenotypic entities can facilitate entity-based searching of documents, which can be far more effective than simple keyword-based search. However, clinicians are interested not only in retrieving all instances of documents that mention a particular entity, but also in locating specific pieces of knowledge involving the entity that are reported in the records, and which can help them to answer questions that arise during the diagnosis process.

An example of such a question is *What are the causes of hypoxia in patients with CHF?*. In seeking answers to this question, clinicians are only interested in medical records for patients known to have CHF and, more specifically, those in which *hypoxia* is mentioned to occur as a result of other medical conditions. Some examples of sentences that would fulfil the clinicians' information need are as follows: '*Congestive heart failure along with obesity underlying restrictive lung disease could be the cause of hypoxia*' and '*the patient had worsening hypoxia related to restrictive lung disease*'.

In order to allow the results of search systems to match more closely the requirements of clinicians, research into relation extraction aims to carry out a deeper

analysis of the text, with the aim of identifying and/or characterising relations and the entities that participate in them. The output of such analyses can be used as the basis for developing advanced semantic search systems that allow queries to be performed over this structured knowledge, rather than simply over keywords or entities [71]. This in turn helps in the retrieval of a more focussed set of results in response to a query, which could assist a researcher in formulating hypotheses that could, for example, be subsequently explored in clinical trials.

The PhenoCHF corpus is annotated for both entities and a number of different types of relations in which they are involved, i.e., those denoting *causality*, *finding* and *negation*. We consider the relations in our corpus to be complex for two reasons. Firstly, two of the relation types (i.e. *causality* and *finding*) are *n*-ary, meaning that they can link together an arbitrary number of arguments (i.e., possibly more than two) in the same sentence. Consider the following sentence:

The patient course was significant for chronic renal insufficiency in the setting of a low ejection fraction, congestive heart failure, and volume overload.

In this sentence, the phenotype *chronic renal insufficiency* is stated to be caused by three other phenotypes, i.e.: *low ejection fraction*, *congestive heart failure* and *volume overload*. Hence, a relation is identified in which *chronic renal insufficiency* has 3 different causes. A second reason why the relations are complex is that there may be an overlap between *negation* and *finding* relations, e.g., when the finding relation is negated. This is illustrated in the following sentence:

cardiac: irregular rate and rhythm, no murmurs or rubs

There are three *finding* relations involving *cardiac* and two of them (i.e., those involving "*murmurs"* and "*rubs"*) are negated. Therefore, it is very important to have a joint view over both *finding* and *negation* types, in order to account for the fact that the findings involving "*murmurs"* and "*rubs"* are negated.

Until now, there has been little research into extracting complex relations or events from clinical records, mainly due to the lack of available corpora in which such detailed information has been annotated. One example of *n*-ary relation extraction from clinical records is the i2b2 medication challenge, which requires the extraction of medications and medication-related information (i.e., medication, dosage, mode, frequency, duration and reason), followed by the identification of links between medications and

medication-related details. Although the relations in this corpus are *n*-ary, the relations are not semantically typed. The relation extraction task is tackled as a classification problem to determine whether or not a medication and its attributes are related. The best performing system participating in this challenge was developed by Patrick et al. [220], who used a cascade approach based on two machine learners: CRF to extract medication information, and SVM to determine whether or not pairs of entities were related. As a final step, a rule-based method is applied to connect all the related entities together to build complex *n*-ary relations.

The easiest way to deal with *n*-ary relations is to factorise all the relations into sets of binary relations; this is done by pairing all entities within a sentence, which may or may not be related. Consider the following sentence: *He had renal failure due to heart failure and poor forward flow*. The entities within this sentence can be factorised into the following pairs: *("renal failure", "heart failure"), ("renal failure", "poor forward flow")* and *("heart failure", "poor forward flow").* If an entity pair corresponds to the arguments of a relation in the gold standard, then it is assigned a class of that relation type. Otherwise, the class *none* is assigned. For example, the corresponding relation types for the pairs in the above example are as follows: *causality*, *causality* and *none.*

The precision of *n*-ary relations is equal to p^{n-1}, where *n*-1 refers to the number of binary relations into which the complex relations were factorised. For example, for a 4-ary relation, there are 3 different binary relations. Assume the precision for the binary relation extraction is 0.8; the precision of 4-ary relation extraction will be $0.8^3 = 0.512$. Therefore, applying binary relation extraction methods for *n*-ary relation extraction will result in low performance [221].

McDonald et al. [222] proposed a framework to extract *n*-ary relations with three arguments; the *variation* relation that is targeted corresponds to alternations in nucleic acid levels and is formalised as follows: (*location, initial-state*, and *altered-state*). The described framework divides the task of 3-ary relations into two stages: the first stage involves extracting all possible binary relations, after which a graph is constructed in which the edges represent binary relations between pairs of entities. In the second stage, the maximal cliques are scored to find potential 3-ary relation instances. This system obtained an F-score of 0.64. The lack of using rich syntactic and semantic features adversely affected the performance.

The EventMine system can deal with *n*-ary relations, nested relations and negation, and it has been shown to achieve a state-of-art performance for several event extraction tasks [26, 130]. Accordingly, when suitably adapted to the clinical domain, EventMine possesses the functionality necessary to allow the extraction of relations of the type annotated in the PhenoCHF corpus. The pipeline-based, modular nature of the system makes it straightforward to adapt to new tasks. Additionally, its adaptability to new domains has been demonstrated, and it can achieve state-of-the-art performance by addressing only the major differences between the two tasks [130].

The above-mentioned advantages of EventMine motivated us to investigate how it could be adapted from its original purpose of extracting events from biomedical scientific papers, to the task of extracting of *n*-ary relations between phenotypic entities related to CHF and CKD, with the aim of capturing the effect of the CKD in worsening heart conditions. To demonstrate the suitability of EventMine for our purposes, we compare its performance in extracting relations from PhenoCHF with other state-of-the-art supervised ML methods, which are able to extract only binary relations.

6.1.1 Methodology

In this section, we describe our application of two types of methods to the task of extracting phenotypic relations from clinical records. Firstly, we explain our method of applying binary relation extraction techniques to the problem. Subsequently, we explain how we adapted the EventMine system to allow the extraction of such relations.

6.1.1.1 Binary relations extraction

Since there are 3 types of relations within the scope of our task (i.e., *causality, finding* and *negation*), the relation extraction problem was addressed as a multi-class classification task, in which the classifier determines the type of relation that exists between each pair of entities. We used a four-class classification model, in which the classes correspond to the three relations of interest, plus the additional class *none*, which is assigned when there is no relation between a pair of entities.

To prepare the data for training, all annotated relations were factorised into sets of binary relations; this was done by pairing all entities within a sentence that may or may not be related, and assigning one of the four possible labels, as explained above.

Following the factorisation of the entities, we used the Weka package tools [223] to compare the performance of different classifiers in identifying and classifying relation instances in the test set. We carried out experiments using both Naïve Bayes and Random Forest classifiers, using a range of different features, as described below.

Feature sets

Due to the prevalence of complex sentence structures in biomedical text, effective relation extraction systems must carry out a deep analysis of sentence structure. This is able to provide syntactic information which, in turn, supports semantic knowledge acquisition [20]. Accordingly, we supplemented lexical and semantic features with syntactic features such as grammatical relations between words, as well as the results of dependency parsing, i.e., PAS.

Our features can be categorised into two groups: entity-related features and context features. The feature extraction process used a different external knowledge source (i.e., UMLS) [33], and incorporates the output of different parsers, in order to provide a rich set of syntactic features. Specifically, we applied the GENIA tagger [47] to obtain part of speech tags and chunk sequences, and the Enju parser [224] to obtain the shortest dependency paths between pairs of entities.

Entity-related features

The entity-related features provide information about the two entities or arguments of the relation, including lexical features (i.e., sequence of words that constitute the entity), syntactic features (i.e., POS and chunk sequence obtained from the GENIA tagger) and semantic features (i.e., phenotypic entity classes and UMLS semantic types).

Context related features

Context features include information about the text surrounding pairs of potentially related entities. These include words that occur before the first argument and after the second argument, BOW between the two arguments, the chunk sequence of the words

occurring between the two arguments and the shortest dependency path between the two arguments, which corresponds to the PAS obtained from the Enju parser.

PAS paths are constructed by finding the shortest path that connects two tokens of interest within a parse tree. If two entities are arguments of the same predicate, the shortest path between them naturally traverses through the shared predicate, as shown in Figure 6.1. In cases where two entities belong to different PASes that share a common argument, then the shortest path traverses through this shared argument. An example of this latter case is shown in Figure 6.2, where the shortest path between "*shortness of breath"* and "*volume overload"* traverses through the node for *she* [225].

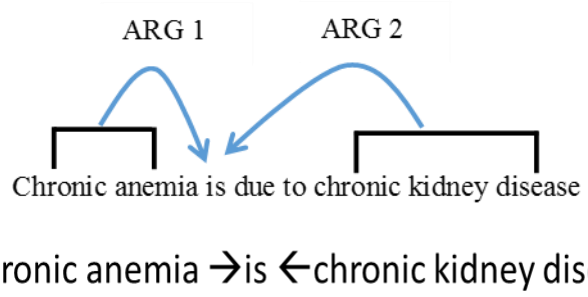


Figure 6.1 Dependency and shortest path between the two related entities

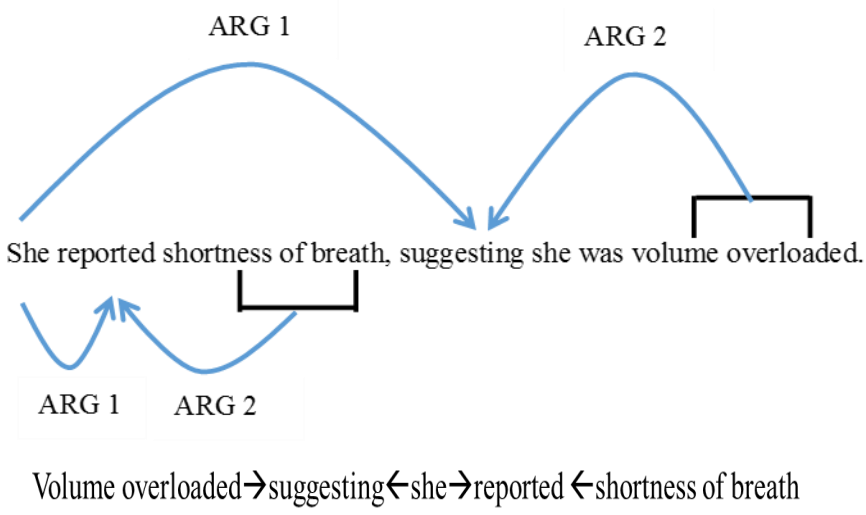


Figure 6.2 Dependency and shortest path between the two related entities

PASes are useful in this task as they represent relations in an abstract manner. For example, the PAS obtained for the sentence *Anaemia causes renal failure* is the same

as that obtained for other possible syntactic variations, including passivisation: *renal failure is caused by anaemia* and relativisation: *renal failure that anaemia causes"*. PAS paths thus normalise the syntactic variability that can be used to express the same information, and allow the construction of general representations of relations between pairs of entities.

Table 6.1 shows examples of all of the features extracted for the sentence *Chronic anemia is due to chronic kidney disease*.

Table 6.1 Examples of the used features for the sentence chronic anemia is due to chronic kidney disease

Entity-related feature		
	Entity 1	**Entity 2**
Sequence of words	Chronic anemia	Chronic kidney disease
POS	JJ NN	JJ NN NN
Chunk	B-NP I-NP	B-NP I-NP I-NP
PhenoCHF class	Non-tradional risk factor	Cause
UMLS class	Disease or syndrome	Disease or syndrome
Context-related feature		
Sequence of words between the two entities	Is due to	
Sequence of POS	VBZ JJ TO	
Chunk sequence	B-VP B-ADJP B-PP	
Shortest path	Chronic anemia→ is ← chronic kidney disease	
Word before the first entity 1	Demonstrated	
Word after entity 2	None	

6.1.1.2 Adaptation of EventMine to the extraction of relations from clinical records

Each module in EventMine applies a one-versus-rest SVM to solve multi-class classification problems using a combination of features, including lexical and syntactic features obtained from multiple parsers i.e., Enju [224] and GENIA Dependency (Gdep) [226]. A detailed description of EventMine can be found in Chapter 2.

Given that EventMine expects event representations as input, the relations in the PhenoCHF corpus were converted into events in order to allow EventMine to be trained to extract them. The conversion was carried out by treating all entities as event triggers and adding relations as arguments if the entity has outgoing relations. For

example, the entities (e.g., cause, risk factor and non-traditional risk factor) that are linked in *causality* relations were converted to PhenotypeE events and all sign or symptom entities linked to organ entity in *finding* relations were converted to FindingE events. Figure 6.3 shows an example of the original relation annotation in PhenoCHF, and Figure 6.4 depicts how these relations are converted into event structures.

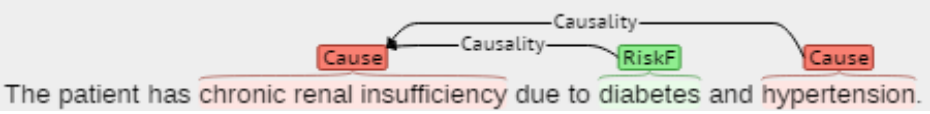


Figure 6.3 Causality relations

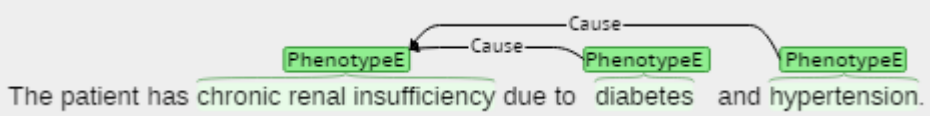


Figure 6.4 Converting the annotation of causality relations into events

Detailed analysis of the converted relations revealed that they have much in common with the events targeted by the GENIA Event Extraction task (GENIA) in the BioNLP ST 2013[126]. In particular, the converted relations in PhenoCHF can have multiple arguments and the relations are sometimes nested, i.e., one relation can have another relation as an argument. In other ways, however, the converted relations in the PhenoCHF corpus are simpler than GENIA events, in that there are only two event types (converted relations) — one pertaining to causality (PhenotypeE) to link a phenotype with its causes and another for finding (FindingE), to link sign or symptom with a corresponding organ. Furthermore, while in GENIA, events can overlap, i.e., a given text span can serve as the trigger for multiple events, this is not the case in the converted PhenoCHF corpus.

Given the above similarities between our converted relations and the types of events that EventMine is designed to extract, we were able to directly train EventMine on our converted corpus, using the same feature sets and configuration as in the original version of the system.

6.1.2 Evaluation

We firstly report on our results for binary relation extraction, and then compare these with the results obtained using EventMine.

The results of applying Naïve Bayes and Random Forest classifiers to the extraction of relations were evaluated using the gold standard PhenoCHF annotations, by splitting the dataset into 80% training and 20% testing. Since the phenotypic entity extraction was separately evaluated as discussed in Chapter 5 [206], the use of the gold standard entity annotations allows us to evaluate the relation extraction task independently. The relation extraction results were evaluated using the standard metrics of precision, recall and F1 measure. To evaluate the performance of extracting 3-ary and 4-ary relations, we compute the recall based on the assumption that, for a relation to be classified correctly, all the binary relations that constitute the 3-ary or 4-ary relations must be correctly identified. Incomplete 3-ary or 4-ary relations were counted as false positives. Tables 6.2 and 6.3 break down the evaluation results according to relation type and number of arguments.

Table 6.2 Random Forest classifier to extract relations

	2-ary			**3-ary**			**4-ary**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Causality	0.66	0.80	0.72	0.44	0.61	0.51	0.28	0.38	0.32
Finding	0.67	0.91	0.77	0.45	0.64	0.53	0.30	0.56	0.39
Negation	0.58	0.95	0.72	-			-		
Macro-averaged F-score	0.63	0.88	0.73						

Table 6.3 Naive Bayes classifier to extract relations

	2-ary			**3-ary**			**4-ary**		
	P	R	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Causality	0.57	0.72	0.63	0.32	0.52	0.39	0.19	0.33	0.24
Finding	0.62	0.85	0.70	0.38	0.58	0.45	0.23	0.51	0.31
Negation	0.57	0.80	0.67	-			-		
Macro-averaged F-score	0.59	0.79	0.66	-			-		

Error analysis showed that most of the FNs can be attributed to data sparsity, i.e., certain ways in which relations can be expressed are rare and/or may have unexpected structures, which do not occur in the training data. Therefore, the classifier struggles to classify such relations when they occur in the test set. A specific example of a *Causality* relation that is expressed in the test set in an unconventional way that does not appear in the training data is: *Renal failure likely secondary to poor forward flow*.

The syntactic structure of this sentence is incorrect, and it lacks a verb to connect the two arguments.

Meanwhile, the most common cause of FPs is incorrect relations that share characteristics with correct relations. For example, in the sentence *Edema is due to fluid overload*, the two concepts, "*edema"* and "*fluid overload"*, were classified as being involved in a *causality* relation because they share the syntactic structure of positive causality relations. Both *finding* and *negation* relations have low precision and high recall. This is mainly because *finding* and *negation* are straightforward to link a negation modifier with the negated entities for the negation relation, or to link an organ with sign or symptoms to form a finding relation. This pattern is also observed in other similar studies that extract relations from clinical records. For example, in a study by Roberts et al. [103] to extract relations, the recall for the *TeRP* relation, which links tests to medical problems, is 90%. This relation is very similar to the *finding* relation in our corpus. In Roberts et al., the description of the *negation-modifies* relation is very similar to the *negation* relation in our corpus and their system achieved a high recall of 98%.

The output of EventMine was evaluated using exact and relaxed evaluation methods, as shown in Table 6.4. The results are very encouraging, this can be partly explained by the narrow topic, i.e., the link between CHF and CKD and simpler nature of the task compared to other relation corpora from well-known extraction tasks in the clinical domain, such as the i2b2 and CLEF tasks. However, our relation extraction task can be considered more complex than the previously introduced i2b2 medication challenge, which only recognised un-typed relations. The best performing system for the medication challenge, by Patrick and Li [220], obtained an F-score of 0.85. Therefore, our best score of 0.77 can be considered satisfactory.

Table 6.4 Results of applying EventMine to extract relations

	Exact			**Relaxed**		
	P	**R**	**F**	**P**	**R**	**F**
Triggers	0.44	0.37	0.40	0.92	0.81	0.85
Argument	0.26	0.16	0.20	0.79	0.76	0.77
Negation	0.48	0.81	0.60	-	-	-

When evaluated using exact matching, the performance of EventMine is significantly lower than that obtained using relaxed matching. This can be partly explained by the fact that EventMine was originally designed to extract events from the biomedical domain where event triggers are usually expressed as short text spans, typically (e.g., *inhibit*) or nominalised verbs (e.g., *inhibition)*. In the converted phenotype relations, however, the triggers are the phenotypes themselves, which are usually expressed as longer sequences of words, i.e. noun phrases of varying complexity and verb phrases, as shown in Table 6.5. Accordingly, when applied to the current task, EventMine only partially recognised many of the event triggers. For example, it was only able to detect the headword of the potentially complex noun phrases, thus leading to low recall when exact matching criteria are applied.

Table 6.5 Different types of phrases corresponding to phenotypes

Phrase types	**Examples**
Simple noun phrases	• *reduced ejection fraction* • *chronic anaemia*
Compound noun phrases that contain coordinators	• *irregular rate and rhythm* • *mother and sister with heart disease*
Noun phrases with prepositions	• *family history of coronary artery disease* • *increased shortness of breath*
Verb phrases	• *jugular venous pressure is 6 cm* • *father died of a myocardial infarction*

Post-processing rules

In order to mitigate the low recall obtained for exact matching, a set of rule-based post-processing steps was developed, based on the outputs of the Enju [224] and GDep [226]. These include predicate-argument structures and word dependency relations, respectively. We studied the internal structure of noun phrases corresponding to triggers, in order to refine trigger detection. The rules used the Enju output to connect pre-modifiers in the same noun phrase to the head noun, by tracing the syntactic tree and finding all the predicates that have the head noun as an argument. This is illustrated in Figure 6.5, where *abnormalitie*s is the head word that was detected by EventMine, and *motion*, *wall* and *regional* are all predicates which have *abnormalities* as an argument. If the detected triggers are part of a noun phrase that contains coordination, then a rule is applied that uses the output of GDep to trace

head dependencies, as shown in Figure 6.6, where the head word is *non-compliance* and the words that are dependent on the head word are *medical* and *dietary*.

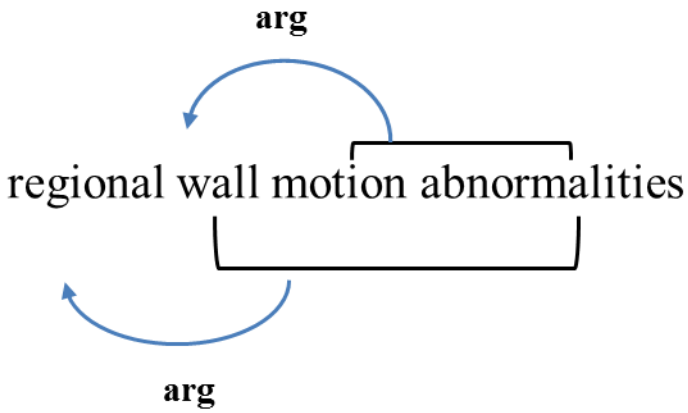


Figure 6.5 Post-processing rule to link the head word with the pre-modifiers

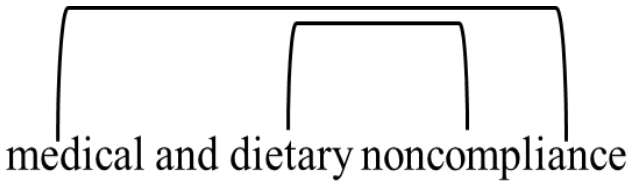


Figure 6.6 Post-processing rule to link the head word with the dependent words

These rules are only designed to connect pre-modifiers with headwords that constitute a noun phrase and compound noun phrases that contain coordinators (e.g., *and*). The current rules cannot help to fully recognise verbal phrases such as "*jugular venous pressure is high"* and noun phrases with following prepositional phrases, such as *"jugular venous pressure at angle of the jaw"*. However, the types of phrases targeted by the rules account for a large portion of the mis-identified triggers and through their application, we were able to improve the exact matching F-scores for trigger detection and argument detection by 26 and by 38 percentage points, respectively as shown in Table 6.6.

Table 6.6 Comparison of F-scores for the performance of EventMine using exact boundary matching before and after post-processing

	Exact matching by EvenMine	
	Without post-processing	**With post-processing**
Triggers	0.40	0.66
Argument	0.20	0.58

Normalising diverse phenotypic phrases to canonical expressions in relevant terminologies (e.g., UMLS) is a prerequisite for effective information extraction; this is discussed further in Chapter 7 of this thesis.

Comparing the results of EventMine with those obtained through the application of binary relation extraction methods showed that, although Naïve Bayes and Random Forest classifiers perform well on the 2-ary relations, the performance decreases when the 3-ary and 4-ary relations are considered. In contrast, the flexible and adaptable nature of EventMine means that it has a stable level of performance, regardless of the number of arguments involved in the relations. Unlike the binary relation extraction methods, EventMine combines all related entities, and outputs complex, *n*-ary relations as a single event structure.

To illustrate more clearly the differences in the outputs of the binary relation extraction methods and EventMine, consider the following sentence: *The patient became hyperkalemic, secondary to poor urine output and acute renal failure.*

The binary relation extraction methods (i.e., Naïve Bayes and Random Forest) output two different relations, each with two arguments, as follows:

- Relation ID: R1, Type: Causality, Argument1: hyperkalemic, Argument2: poor urine output
- Relation ID: R2, Type: Causality, Argument1: hyperkalemic, Argument2: acute renal failure

However, EventMine is able to combine the information into a single relation, which more explicitly encodes the fact that there are two separate causes for a single phenotype.

- Event ID: E1,Type: PhenotypeE, Trigger: hyperkalemic, Cause1: poor urine output, Cause2: acute renal failure

While, as explained above, one challenge of EventMine lies in extracting phenotypes expressed as long sequences of words, nevertheless we were able to address this to a large extent through the application of post-processing rules.

6.2 Summary

In this chapter, we have shown that the PhenoCHF corpus can successfully support the development of different methods to extract phenotypic relations from clinical text.

We have demonstrated that the corpus can be used to train both binary relation extraction methods, as well as more complex *n*-ary extraction methods, which more closely model the types of relations annotated in PhenoCHF. The latter was achieved by adapting EventMine, which was shown to outperform supervised ML-based methods for binary relation extraction. Our results illustrate that automatic detection of complex *n*-ary relations in medical records is a feasible task, and that EventMine can be successfully adapted to this task.

Chapter 7 Integrating phenotypic information from clinical records and literature articles

In Chapter 5, we discussed our experiments on the automatic extraction of phenotypic entities related to CHF from both EHRs and literature articles. In Chapter 6, we built upon this by extracting relations between these entities that are specified in EHRs (e.g., causality relations that highlight the interaction between CHF and kidney failure). Although EHRs and literature articles provide valuable information that can complement each other, it can be problematic to combine the knowledge contained within them, according to the varying ways in which information is expressed in each of these textual sources. However, such a combination is important since it can be vital to allow for the discovery of new disease-phenotypic associations which may not be apparent if only a single knowledge source is considered.

Mapping different and variant mentions of the same concept within the different types of textual sources to concepts in domain-specific resources such as UMLS is an important step towards bridging the gap between the information contained within the two sources. Normalising phenotypic entities in this way can help to draw generalisations about information that may be expressed in text in many different ways; it also constitutes an important first step towards the effective integration of complementary information dispersed within these sources, in order to facilitate new knowledge discovery and generation of new hypotheses.

7.1 Background

The problem of concept normalisation for genes and proteins has been extensively studied, according to its central role in a number of the BioCreative shared tasks [227-229], where participants are required to produce lists of genes and proteins mentioned in documents, and to link each one to a unique concept in a domain-specific database. A variety of methods including pattern matching, dictionary lookup, ML and heuristic rules were applied in the systems participating in these challenges. The shared tasks have resulted in the development of several new techniques that complement the more traditional dictionary-based methods, although they are largely customised for operation on biomedical literature.

The development of the NCBI disease corpus [158], which is comprised of the abstracts of biomedical articles, represents a rich source for researchers to explore normalisation methods for disease names, and several methods customised for disease name recognition have emerged [230-232]. To our knowledge, the ShARe/CLEF eHealth Evaluation Lab [28] is the only shared task that has focussed on normalising clinical concepts that occur in EHRs. The normalisation process aims to map textual mentions of disorder entities to concepts in UMLS [233]. However, the more specialised task of normalising phenotypic information in EHRs has not been previously attempted, probably according to the lack of a gold standard that can be used to evaluate existing techniques or support the development of new techniques.

This is not to say that automatic acquisition of phenotypic knowledge from text is not an active research area. Indeed, the *Phenotype day* events [234, 235] bring together a large number of researchers to promote advances in the state-of-the-art of phenotype knowledge acquisition and to support deeper understanding for phenotyping. Researchers have proposed solutions to a number of different topics including: 1) composition of ontologies to represent phenotypes [236-239] 2) tools and pipelines to support phenotypic data curation and integration with ontologies [172, 236, 240] 3) application of phenotype knowledge acquisition in real world applications (e.g., discovery of phenotype-genotype relation) [241-243]. However, as far as we are aware, using normalisation techniques to integrate EHRs and biomedical literature has not previously been attempted.

Although the PhenoCHF corpus [171] is annotated for mentions of phenotypic entities, the annotation process did not involve linking these mentions to unique concepts in an external knowledge resource. This meant that it was not possible for us to apply ML techniques to carry out the normalisation process, since training an ML system to carry out this task requires that links between entities and database identifiers are annotated. This is necessary so that the ML algorithm can learn how to model the similarity between annotated textual mentions and information about the corresponding concepts in the external knowledge resource [244].

The absence of such links in our annotated corpus limited the types of normalisation approaches that we could apply using the PhenoCHF corpus to those based on dictionary lookup and/or string similarity. Accordingly, our review of normalisation approaches in the following section concentrates on commonly used dictionary-based and string similarity methods.

7.1.1 Dictionary-based methods

A large number of approaches to concept normalisation in the biomedical domain rely at least partially on dictionary lookup techniques. Existing NLP systems (e.g., MetaMap [63], cTAKES [56] and SAPHIRE [245]) aim to recognise entities belonging to a variety of semantic categories, and to map them to concepts in the UMLS Metathesaurus. These systems employ mainly dictionary-based methods, i.e., they attempt to match phrases occurring in documents with synonyms of terms that are listed in UMLS. Heuristics are employed in some of these systems to allow recognition of textual concept mentions that do not match exactly with synonyms listed in the resource. These heuristics include normalizing case, removing inflections, generating derivations, using additional lexical resources or permuting the words contained within UMLS entries. The application of post-processing rules has also been shown to be beneficial [231].

Whilst such heuristics are successful in detecting term variants to a certain extent, e.g., those exhibiting different word order, they still largely assume that the actual words used to express a concept will be the same as those already present in the Metathesaurus. However, in reality, language use is highly creative, meaning that in practice, such an assumption is too restrictive.

Oellrich et al. [246] investigate the performance of four concept recognition tools (i.e., cTAKES; MetaMap; the National Center for Biomedical Ontology (NCBO) annotator [247] and the Biomedical Concept Annotation System (BeCAS) [248]) on ShARe/CLEF dataset to extract disorder mentions and map them to unique concepts in UMLS. Amongst the four systems, the best performance was achieved by MetaMap. However, a number of problematic issues were observed for all four of the NLP systems evaluated, which lead to decreased performance and incorrect recognition of the disorder concepts. Such issues include incorrect boundary detection, difficulties in correctly mapping abbreviations and problems in handling phrases containing coordination. These findings suggest that adding additional pre-processing steps that are aimed at resolving abbreviations and coordinations would be advantageous.

7.1.2 String similarity-based methods

Approximate string matching methods move beyond pure dictionary-based matching, in that they allow textual mentions to be mapped to dictionary entries that they closely

resemble, rather than requiring an exact match. By calculating a numerical measure of similarity between a recognised entity in text and entries in a dictionary, it is possible to determine the most appropriate entry to which the textual entity mention should be mapped.

String similarity methods may be classified into three main approaches: edit distance, token-based and hybrid approaches.

7.1.2.1 Edit Distance

The most widely-used means of determining the similarity between a pair of strings is to calculate the *edit distance* beween them. Edit distance metrics quantify the similarity between two strings by counting the minimum number of operations (e.g., insertion, deletion) required to transform one string into the other. A variety of methods of calculating edit distance has been proposed, examples of which are as follows:

The Levenshtein edit distance [249] corresponds to the minimum number of insertions, deletions or substitutions needed to transform a string *X* into another string *Y*.

Two other edit distance metrics that are broadly similar to each other are Jaro [250] and Jaro-Winkler [251] metrics. The Jaro metric is based on both the number and order of the common characters that are contained within two strings *X* and *Y*. A character is counted as common if it occurs in the other string within a given distance, which depends on the length of the string. Let X= x_1 , x_2, …….x_i and Y= y_1, y_2……y_j. The Jaro distance is calculated as follows:

$$\text{Jaro (X,Y)}=\begin{cases}\frac{1}{3}\left(\frac{m}{|X|}+\frac{m}{|Y|}+\frac{m-t}{m}\right)\end{cases}$$

Where:

- *m* is the number of the matching characters
- *t* is the number of *transpositions*

An identical pair of characters from *X* and *Y* are considered to match if they are not farther from each other than:

$$\left[\frac{max(|X|,|Y|)}{2} - 1 \right]$$

The Jaro metric was developed to detect spelling variations between a pair of strings being compared. Thus, the Jaro metric will range from 0 to 1, with a higher value signifying a greater similarity between the pair of strings being compared [252].

Winkler proposed an enhancement to the Jaro metric [251] based on his observation that spelling errors occur more commonly toward the end of a string. Winkler's enhancement assigns a higher comparator score to strings with common prefixes as they are more likely to be similar.

Jaro-Winkler uses the length *P* of the longest common prefix of *X* and *Y*. *P* is the constant scaling factor that gives more favourable similarity scores to strings that match towards the beginning, rather than towards the end [253]. Let *P'* = max (*P*,4)

$$\text{Jaro-Winkler (X, Y)} = \text{Jaro (X, Y)} + \frac{P' \cdot (1 - \text{Jaro(X,Y)})}{10}$$

Like the Jaro metric, the Jaro-Winkler metric will also range from 0 to 1, with a higher value signifying a greater similarity between the pair of strings being compared.

The edit distance metrics seem to be intended for single tokens [253]. To find the similarity between longer strings that include multiple tokens, token-based distance functions have been proposed.

7.1.2.2 Token-based edit distance metrics

Token-based similarity measures aim to find the similarity between strings that include multiple tokens. They convert two strings *X* and *Y* into (unordered) bags of tokens; a similarity function is then used to calculate the level of similarity between the two sets of tokens. Examples of token-based distance metrics are the Jaccard similarity and cosine similarity.

The Jaccard similarity function [254] measures the similarity between two strings *X* and *Y* as shown in the following equation:

$$\text{Jaccard (X,Y)} = (|X \cap Y|)/(|X \cup Y|)$$

Where (|X∩Y|) refers to the number of overlapped tokens between X and Y and (|X∪Y|) refers to the total number of token in the union of X and Y. The higher the score, the more similar X and Y are to each other. This makes the Jaccard measure biased toward short strings [253]. For example, let the string in query X = 'reduced ejection fraction', Y_1 = 'ejection fraction' and Y_2 = 'decreased ejection fraction', J(X, Y_1) =0.40 and J(X, Y_2) =0.33. Nevertheless, Y_2 is more similar to X since *reduced* and *decreased* are synonyms. Jaccard metric gives higher score for Y_1 as the number of total tokens in the union is smaller.

Cosine similarity or TF-IDF (Term Frequency- Inverse Document Frequency) is a further token-based similarity metric which is widely used in information retrieval tasks. The TF-IDF weighting scheme is a statistical measure used to evaluate the importance of a word within a document in a collection or corpus [255]. It works by comparing common tokens in the strings *X* and *Y*.

$$\text{TF-IDF}\ (X,Y) = \sum_{w \in X \cap Y} V(w,X) \cdot V(w,Y)$$

$$V'(w,X) = \log(TF_{w,X}+1) \cdot \log(IDF_w) \quad \text{and}$$

$$V(w,X) = V'(w,X) \Big/ \sqrt{\sum_{w'} V'(w,X)^2}$$

Where $TF_{w,X}$ is the frequency of word *w* in string X and IDF_w is the inverse of the fraction strings in the corpus that contain *w*.

However, the terms are weighted, such that rarely occurring words are assigned higher weights. As such, higher similarity scores are assigned to pairs of strings whose matching words have a distinguishing meaning, rather than function words like *the, of* etc. However, the measure is only useful when the strings share strictly identical tokens; it would not work, e.g., in the case that a word is misspelled in one of the strings being compared, e.g. "*high blood presser* "and "*pressure"* [257].

7.1.2.3 Hybrid distance function

To overcome the limitations of TF-IDF in comparing strings whose tokens may exhibit spelling variations, a "soft" version of TF-IDF was introduced by [253], which combines both character-based and token-based metrics. Using this method, the similarity score is determined by considering *similar* tokens, as well as identical tokens that appear in both X and Y. Accordingly, in this approach, a special set of words CLOSE $(\emptyset, X, Y)$ w $\in X$ is determined, such that there exists some word v $\in Y$ where $\text{sim}'(w, v) > \emptyset$ [256].

In order to calculate the similarity between two strings *X* and *Y*, the strings are firstly broken into tokens, which are weighted using the TF-IDF statistical weighting scheme. Let sim' be a secondary similarity function that performs well on short strings (e.g., Jaro-Winkler). Let CLOSE $(\emptyset, X, Y)$ represent a set of tokens $w \in$ X such that there is a similar token $v \in Y$, and for $w \in$ CLOSE $(\emptyset, X, Y)$, let D(w,Y)= $max_{v \in Y}$ dist (w, v). SoftTFIDF is then calculated as follows:

SoftTFIDF(X,Y) =

$$\sum_{w \in CLOSE\ (\emptyset, X, Y)} V(w,X) \cdot V(w,Y) \cdot D(w,Y)$$

SoftTFIDF is similar to cosine similarity, except that instead of requiring exact token matches, it computes approximate matches between tokens, using a secondary distance function (e.g., Jaro-Winkler) [258]. Tokens are considered sufficiently similar to each other if the secondary distance function returns a value that is above a pre-specified threshold (usually 0.9). However, if a similar token does not appear in both X and Y, then the SoftTFIDF value will be equal to the cosine similarity value [257]. Furthermore, SoftTFIDF does not consider the string length (i.e., the number of tokens) in computing the similarity, which can be very important in mapping a string to the most similar term. For example, in terms of meaning, "*worsening in exercise tolerance"* is more correctly mapped to '*reduced exercise tolerance'* rather than to '*exercise tolerance'*. However, the SoftTFIDF method would assign a higher similarity value to the latter term.

7.2 Methodology

We have developed a novel method called *PhenoNorm,* which integrates surface-level and semantic similarity measures to allow variant mentions of different types of phenotypic information to be mapped effectively to concepts in the UMLS Metathesaurus. We compare our method with two baseline methods, i.e., dictionary-based (using MetaMap) and string similarity based (using SoftTFIDF).

Pre-processing

To address the specific characteristics of our corpus, we applied a number of pre-processing steps. Different methods were applied both to the UMLS terminology itself, and to the phenotypic entities in the PhenoCHF corpus, to help to increase the accuracy of the subsequently applied mapping procedure.

Firstly, we filtered the UMLS vocabulary to include only those terms belonging to the disorder semantic group. Since all phenotypic concepts are expected to fall within this group, this filtering step ensures that irrelevant concept candidates are disregarded by our normalisation method. We built the inverted index to store the mapping between each word in the phenotypic entities and all UMLS terms in which this word appears.

As has been outlined previously, a detailed manual analysis of PhenoCHF revealed that phenotypic concepts can be expressed using diverse syntactic means. These include simple noun phrases (e.g. "*progressive renal failure"*), coordinated noun phrases (e.g., "*increased chest pain and fatigue"*), noun phrases followed by prepositional phrases (e.g., "*increasing dyspnea on exertion"*) and complete clauses or sentences (e.g., "*jugular venous pressure is elevated"*).

Our similarity method is not restricted to specific syntactic structures, and hence it can be applied in mapping annotations with all types of structures to appropriate UMLS concepts. However, special treatment is needed for coordinated noun phrase annotations, since they usually correspond to multiple individual concepts, each of which should be mapped to a separate UMLS concept. Accordingly, we carried out pre-processing of such annotations through the application of a rule-based module (Baumgartner et al. [259]), which uses POS and chunk information obtained from the GENIA tagger [47] to split coordinated phrases into appropriate separate entities, e.g., "*increased chest pain and fatigue"* is split into *increased chest pain* and *increased*

fatigue. A further pre-processing pipeline was applied to the entity annotations in PhenoCHF, in order to convert all letters to lower case, to tokenise multi-word annotations, to remove stop words, e.g., *the, is, was*, etc., and to expand abbreviations into their full forms. The latter task was achieved in two steps. Firstly, the MetaMap concept recognition tool was applied to map the abbreviations to appropriate UMLS concepts; the preferred full form of the concept was subsequently retrieved from UMLS. Given that acronyms can be ambiguous, and thus MetaMap will not always produce the correct mapping, the automatically derived expansions were manually checked and corrected by a doctor, with the aid of an online abbreviation and acronym resource (http://acronyms.thefreedictionary.com/).

7.2.1 PhenoNorm

Following the pre-processing steps, we applied our *PhenoNorm* method to carry out the mapping of phenotypic entity annotations to UMLS concepts. The method works as follows:

For a given phenotype annotation that consists of *n* tokens (*token(1),token(2)……token(n)*) , the following steps were undertaken (see Figure 7.1) :

- For each *token(i)*, the inverted index is consulted to determine all UMLS terms that include the token.

- Candidate terms with the most similar sets of tokens to the phenotype annotation are found by computing the intersection between the sets of hits retrieved for each *token (i)*. The set of candidates is then reduced by considering only those UMLS terms whose tokens match most closely to those in the phenotype annotation. If any of the candidate UMLS terms matches *exactly* with the phenotype annotation, then the algorithm terminates. Otherwise, the closest non-exact matches are sought. Firstly, it is determined whether any of the candidate terms share *all* words with the phenotype annotation (but possibly in a different order). If such candidates exist, then the algorithm moves on to step 3. If no such candidates exist, then the constraint is relaxed such that candidates with only (n-1) matching words will be considered, and so on.

- Each candidate term identified in step 2 is assigned a score based on its level of similarity to the phenotype annotation. Similarity is computed using Levenshtein edit distance metrics, i.e., the minimum number of character level operations (e.g., insertions, deletions) required to transform the phenotype annotation into the candidate UMLS term.

- The phenotype annotation is mapped to the UMLS concept associated with the candidate term with the minimum edit distance to the phenotype annotation.

- If the phenotype annotation does not contain any tokens that match with a UMLS term (e.g., *diabetesmellitus*, which occurs in UMLS as *diabetes mellitus*), then character n-grams are employed as the means of calculating similarity between the phenotype annotation and the UMLS terms (where n is 5 by default, and 3 if the length of the *token (i)* is less than 5). For each *token (i)* in the phenotype annotation, all UMLS terms containing the least frequent (rarest) n-gram in *token (i)* are retrieved, since rare n-grams tend to be the most informative. Steps 2-4 are then repeated.

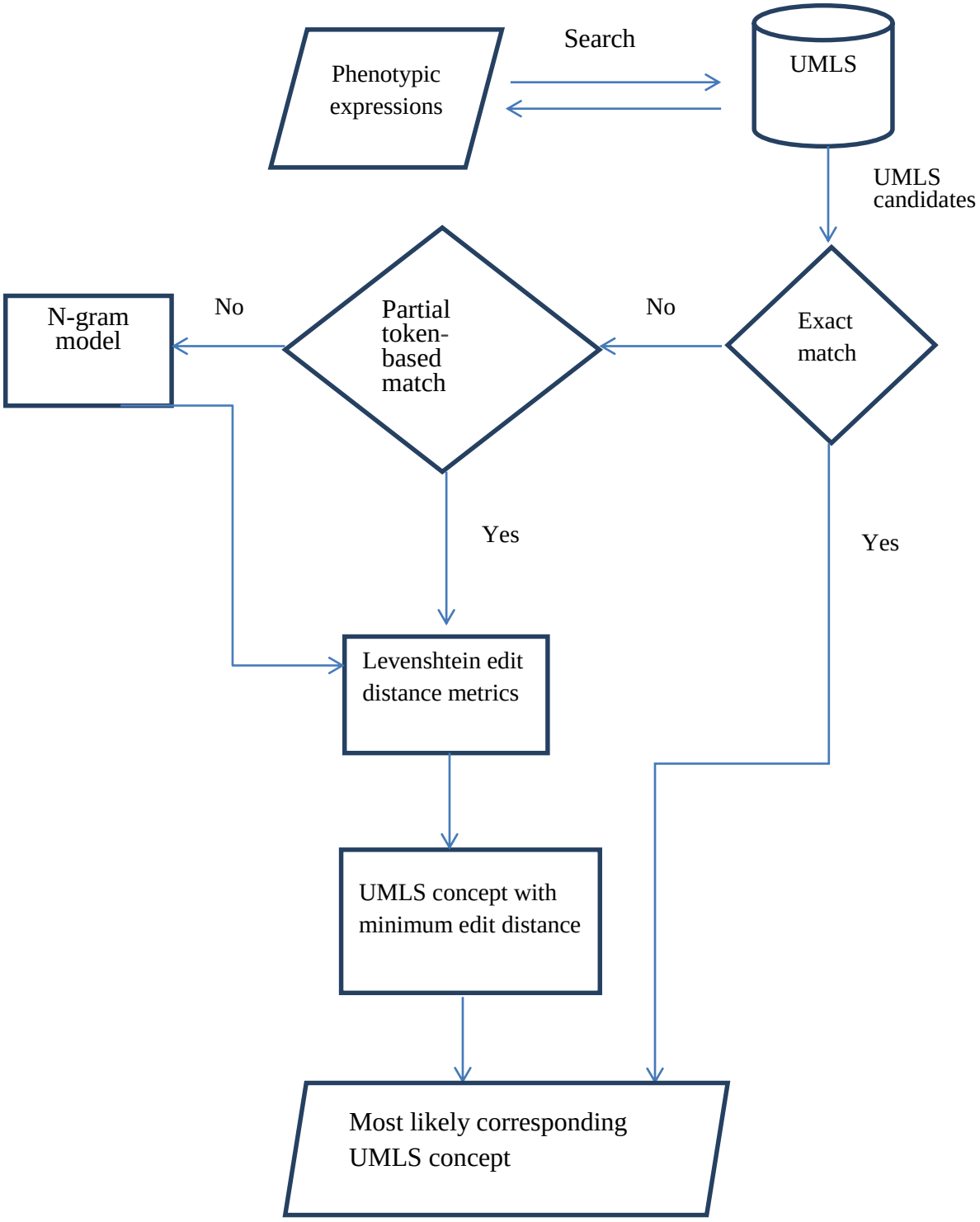


Figure 7.1 Workflow for the normalisation steps

Pre-processing rules

As mentioned above, one of the features of our method is that it gives a higher priority to terms sharing the greatest number of words, and it also favours terms that have a similar length to the phenotypic annotation, rather than considering overall edit distance. Although this strategy results in the correct mapping in the majority of the cases, the fact that the meaning of non-shared words can be different can cause incorrect mapping in some cases. For example, the entity annotation "*elevated pulmonary capillary wedge pressure"* is incorrectly mapped to the UMLS term '*decreased pulmonary capillary wedge pressure'* instead of '*increased pulmonary arterial wedge pressure'*. This is because the selected term shares four words in common with the annotated entity. This means that it is assigned a higher similarity score. However, the non-shared words in the phenotype annotation and the UMLS term that are linked by our method, i.e., *elevated* and *decreased* have completely opposite meanings, whereas the semantics of the words *elevated* in the phenotypic mention and *increased* in the correct UMLS term are similar. This demonstrates that it would be advantageous to take meaning as well as surface similarity into account when performing the mapping.

With this in mind, we applied pre-processing rules to the initial results of our mapping, in order to better account for semantic-level similarities between terms. We utilised WordNet [136], a large lexical database of English, in which words are organised into sets of synonyms (called *synsets*); synsets are linked together into a semantic network. For example, the synonyms *elevated* and *raised* occur within the same synset, and this synset is linked within the network to the synset that contains *increased*. The pre-processing rules used WordNet to generate variants of each phenotype annotation, by using WordNet to find the synonyms of each adjective or noun appearing in the original annotation. The generated phenotypic variants are used as input to be processed by the PhenoNorm method.

With the aid of these pre-processing rules, semantic-level variations such as "*increased pulmonary arterial wedge pressure"* can be generated from the original phenotypic annotation "*elevated pulmonary capillary wedge pressure"*, which allows PhenoNorm to successfully map the annotation to the correct UMLS term '*increased pulmonary arterial wedge pressur*e'.

7.2.2 Baseline techniques

Our baseline methods involved applying MetaMap [63] and the SofTFIDF string similarity method [253]. MetaMap firstly splits the input text into sentences and the noun phrases within each sentence are identified. For each noun phrase, MetaMap identifies possible mappings to UMLS concepts based on lexical lookup and on variant generation, and associates a score with each potential mapping. Similarly to the configuration of MetaMap described in chapter 5 to extract phenotypic entities, we configured the tool to recognise only those concepts belonging to the 12 categories under the disorder semantic group, since all phenotypic concepts fall within this semantic group.

To apply the SoftTFIDF method, we used the implementation provided in the secondstring package [260], which has achieved good results when applied to several different string-matching problems [253].

7.3 Results

The PhenoNorm and baseline methods were evaluated using the accuracy metric, which is defined as follows:

Accuracy= correct/total

where *correct* refers to the number of phenotypic entities mapped to the correct UMLS concept and *total* refers to the total number of entities in the gold standard. The correctness of the mappings to UMLS was determined by an annotator of the PhenoCHF corpus, who is a clinician.

The results of the evaluation of the PhenoNorm method are summarised in Table 7.1, showing the accuracy achieved for each phenotype category, both with and without the WordNet-based pre-processing step.

Table 7.1 Results of applying PhenoNorm method to the PhenoCHF corpus

Phenotypic categories	**Accuracy**	
	Without post-processing	**With post-processing**
EHRs		
Cause	0.89	0.90
Risk factor	0.74	0.75
Sign or symptom	0.78	0.83
Non-traditional risk factor	0.86	0.88
average	0.82	0.84
Articles		
Cause	0.91	0.91
Risk factor	0.87	0.88
Sign or Symptom	0.83	0.85
Non-traditional risk factor	0.86	0.88
average	0.87	0.88

The evaluation of the two baseline methods is reported in Table 7.2., in which their performance is compared to the best result achieved by the PhenoNorm method.

Table 7.2 Comparison of MetaMap, SoftTFIDF and the best result of PhenoNorm

Method	**Accuracy**	
	EHRs	**Articles**
MetaMap	0.46	0.56
SoftTFIDF	0.76	0.83
PhenoNorm	0.84	0.88

As can be observed in Table 7.1 PhenoNorm achieves good results and that in the majority of cases, the WordNet-based pre-processing helps to further increase the accuracy. The better performance on literature articles is likely to be due at least in part to the more formal nature of the writing, in which authors frequently use standardised forms to refer to clinical concepts. For EHRs, the performance for the *cause* and *non-traditional risk factor* categories is almost as high as for articles. The results are somewhat lower for *risk factor* and *sign or symptom* in EHRs as these two types of phenotypic information are expressed in long sequences and usually mentioned with qualifiers (e.g., *increased, reduced, elevated, moderate, severe,* etc.). For example, "*severe myocardial infarction*" is incorrectly mapped to '*recent*

myocardial infarction' instead of '*myocardial infarction*'. However, as can be observed, the pre-processing step is particularly helpful for boosting performance for *sign or symptom* entities in this text type (i.e., an increase in accuracy of 4.6% is achieved).

Literature article analysis

In the majority of cases, as can be verified by our encouraging results, the PhenoNorm method correctly maps most of the phenotypic mentions in literature articles to UMLS concepts. Examples of correct decisions determined by our method include mapping "*increasing chest pain"* to '*chest pain increasing in severity'*; "*significant jugular venous distention"* to '*increased jugular venous distention'*; and "*stenosis in left anterior descending"* to '*left anterior descending coronary artery stenosis'*.

An analysis of mapping errors produced by our method when applied to literature articles revealed that issues can occur when annotated entities do not correspond exactly to UMLS terms. For example, the phenotypic mention *increased oxygen requirement* has a similar meaning to the UMLS concept *hypoxia*. However, it is mapped incorrectly to the UMLS concept *increased insulin requirement*, since this concept shares two tokens with the phenotypic entity.

Another source of error results from the fact that this method favours strings with similar numbers of words, and those with the greatest number of shared words, rather than considering the overall edit distance between the strings. Although this feature is useful in most cases, this behaviour can, in a small number of cases, give incorrect mapping. For example, "*intermittent chest pain"* is mapped incorrectly to '*intermittent flank pain'* instead of the most similar UMLS term, i.e. '*chest pain'*. This wrong decision was made because there are two shared words between the two terms (i.e., *intermittent* and *pain*), and since the matched UMLS term is of the same length as the phenotype annotation.

EHR analysis

In the EHR subset of PhenoCHF, there is far more variation in the way that phenotypic concept mentions are expressed, and in certain cases, there is no corresponding UMLS term to which the mentions can be matched. This is the case, e.g., for "*troponin leak"* and "*low flow to the kidneys"*. Another type of mapping error can occur when a UMLS term contains all tokens in the phenotype annotation, even

though it actually represents a different concept, the annotation "*family history for coronary heart disease"* is mapped to the UMLS term '*family history: premature coronary heart disease'* instead of '*family history of coronary artery disease'*. However, this feature of our method was useful in other cases, e.g., to allow the successful mapping of the phenotype mention "*elevated right heart filling pressures"* to the UMLS concept '*elevated right atrial pressure'* and of the mention "*decreased breath sounds bilaterally"* to '*decreased breath sounds bilaterally bases'*.

PhenoNorm achieved superior results to both baseline methods, with considerably better performance for the EHRs. SoftTFIDF performed acceptably on average for both EHRs and literature articles datasets. MetaMap achieved the lowest results; this is to be expected, given the high variability in the means of expressing phenotypic concepts, and the fact that MetaMap relies largely on lexical lookup to determine how noun phrases should be mapped to UMLS concepts.

Error analysis revealed that incorrect mappings produced by SoftTFIDF occur mainly because it does not take into account the length of the phenotypic entity (i.e., number of tokens). For example, SoftTFIDF incorrectly maps the phenotypic entity "*moderately reduced left ventricular systolic function"* to '*moderate left ventricular systolic dysfunction'* instead of '*moderately or severely depressed left ventricular systolic function'*. This mapping is chosen by softTFIDF because it assigns a higher score to the shortest UMLS term with greatest number of common tokens without considering the overall number of tokens in the phenotypic concept.

Another source of error comes from the secondary similarity metrics; SoftTFIDF uses Jaro-Winkler as a secondary edit-distance metric to find similar strings. When applying Jaro-Winkler to compare terms, the similarity score is highly dependent on the similarity of the prefix (i.e., the first four letters) of the tokens being compared. For every phenotypic annotation, SoftTFIDF finds all UMLS terms with tokens in common with the phenotype annotation, and then the Jaro-Winkler metric is applied to further filter these UMLS terms, by finding the shortest UMLS term in which the prefixes of the tokens match those in the tokens of the phenotype annotation.

For example, the SoftTFIDF method incorrectly maps the phenotypic annotation "*reduced cardiac output"* to the more general UMLS term '*cardiac output'* instead of to the more specific UMLS term '*decreased cardiac output'*. This is because the prefix of the token *reduced* does not match any of the shortlisted UMLS terms that

have tokens in common with the phenotypic concept, and since the method does not account for semantic-level similarities.

The constraints of SoftTFIDF contribute to many errors and mean that the method is biased towards mapping to the shortest possible UMLS terms, meaning that important information expressed in qualifier concepts (e.g., *high, low, moderate, severe*), which determine the degree of the disease, is often ignored.

Similar patterns of errors were evident in the output of MetaMap. For example, FNs occurred frequently when phenotypic entities correspond to term variations that are not listed in UMLS (e.g., "*cardiac troponin leak"* which is a variant of '*cardiac troponin increased'*) or when annotations were mapped to more general concepts, e.g., "*high Jugular venous pressur*e" is incorrectly mapped to '*jugular venous pressure'* instead of to the correct, but more specific concept '*raised Jugular venous pressure'*. Another cause of FNs is spelling mistakes that occur in the clinical records. For example, "*eg edema"* is not mapped to '*leg edema'*, but rather to the more general concept '*edema'*. Furthermore, since MetaMap is designed to only recognise simple phrases, it fails to recognise phenotypic concepts that are expressed using different syntactic structures. For example, the phenotypic concept "*moderately reduced left ventricular ejection fraction"* is parsed by MetaMap into three different concepts, i.e., '*moderately'*, '*reduced'* and '*left ventricular ejection fraction'* and thus mapping it incorrectly into three different UMLS concepts instead of to the correct single concept '*left ventricular ejection fraction decreased'*. A further example is the phenotypic concept "*stenosis in left anterior descending",* which is parsed by MetaMap into three different concepts, i.e., '*stenosis'*, '*left anterior'* and '*descending'* and therefore it cannot be mapped to the correct, but much longer concept '*left anterior descending coronary artery stenosis'*.

In general, SoftTFIDF achieves a level of performance that is more comparable to the PhenoNorm method than the much lower results achieved by MetaMap. The main issues of SoftTFIDF are that it does not take into account the overall length of the term, and that the Jaro-Winkler secondary similarity metric does not always produce the correct results. However, the fact that it does not attempt to split long terms into shorter ones, as is the case with MetaMap, means that it achieves far superior performance to MetaMap.

To highlight the differences in the mappings to UMLS provided by the SoftTFIDF and the PhenoNorm methods, a comparison is provided in UMLS mappings Table 7.3.

Table 7.3 Comparison of the UMLS mappings produced by SoftTFIDF and PhenoNorm for the same phenotypic concepts

Phenotypic concepts	SoftTFIDF Mappings	PhenoNorm Mappings
moderately reduced left ventricular systolic function	moderate left ventricular systolic dysfunction	moderately or severely depressed left ventricular systolic function
High jugular venous pressure	jugular venous pressure	raised jugular venous pressure
Diminished exercise tolerance	exercise tolerance	impaired exercise tolerance
Mother died of myocardial infarction	subsequent myocardial infarction of other sites	family history of myocardial infarction
Reduced cardiac output	cardiac output	decreased Cardiac Output
Coagulation imbalance	Imbalance	coagulation abnormal
Uremia-associated dyslipoproteinemia	Uremia	dyslipoproteinemia
Increased pulmonary artery systolic pressure	pulmonary artery mean systolic pressure	doppler echocardiography: increased derived systolic pressure of pulmonary artery.
significantly reduced left ventricular ejection fraction	ejection fraction	left ventricular ejection fraction decreased

7.4 Discussion

An evaluation of the output of our method confirmed that it is able to handle a variety of types of term variation, as summarised in Table 7.4.

Table 7.4 Types of term variation in the PhenoCHF corpus

Type of variability	PhenoCHF mentions	UMLS term
Orthographic variation	light-headness	lightheadness
Morphological variation	Hyperkalemia	hyperkalemic
Roman-Arabic variation	type II diabetes	type 2 diabetes
Differing numbers of words	mild mitral regurgitation	mitral regurgitation
Synonyms	worsening renal function	decreased renal function
Different internal structure of terms	jugular venous pressure is elevated	elevated jugular venous pressure
spelling mistakes	left ventricular hypertrophi	left ventricular hypertrophy

Comparison of concepts used in the literature and EHRs

Analysis of the UMLS concepts which are linked to the phenotype annotations in PhenoCHF corpus by our method revealed that a total of 835 UMLS concepts are mentioned in PhenoCHF as a whole. Of these concepts, 184 occur in both EHRs and literature articles, as shown in Figure 7.2.

The fact that the remaining concepts only appear in either EHRs or the literature articles (but not both) provides strong evidence of the complementary nature of the information contained within the two text types. However, our finding that there is a significant overlap in the concepts that occur in both parts of the corpus provides evidence that common types of information are reported in the two text types. The shared concepts between the two parts represent a link between EHRs and the literature articles. It can also be used to support the provision of personalised healthcare, by linking patients' clinical records to the new findings and discoveries in the literature.

Manual analysis of the concepts that are shared between EHRs and the literature articles revealed that there is variability between the two text types in terms of expressing the same phenotypic concept. In literature articles, standardised forms are most commonly used to refer to phenotypic concepts, whereas mentions of the same concepts in the EHRs are generally more descriptive, and can consist of long phrases with different syntactic structures. To illustrate this, Table 7.5 includes some examples of the diversity in expressing the same phenotypic concepts in EHRs and literature articles. These examples highlight the importance of employing sophisticated normalisation methods to map these concept mentions to standardised forms, to allow the links between the two information sources to be established, based on their commonly mentioned concepts, and thus to allow complementary information to be integrated.

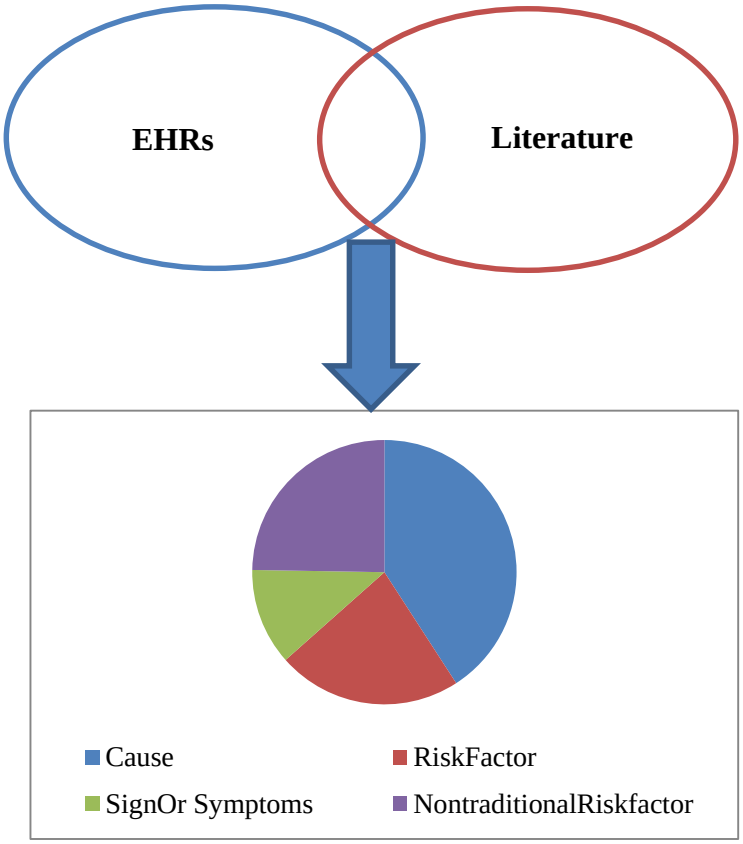


Figure 7.2 The overlap between EHRs and articles phenotypic concepts

Table 7.5 Difference in expression of the same phenotypic concepts in EHRs and the literature

Type of variability	**PhenoCHF corpus**	
	EHR mentions	**Article mentions**
Synonymy	sodium overload drop in blood pressure	hypernatremia hypotension
Syntactic structure	left ventricular is dilated mild mitral calcification	left ventricular dilatation calcification of mitral valve
Word ordering	cardiac output decreased	decreased cardiac output
Spelling variation	Hyperkalemic	hyperkalemia
Additional word	moderate left ventricular enlargement	left ventricular enlargement

7.5 Evaluation

To demonstrate the wider applicability of our PhenoNorm method, we have also evaluated it on four different annotated data sets (ShARe/CLEF task1 [28], NCBI disease corpus [158], heart failure and pulmonary embolism annotations [239]). In

each case, we firstly process the gold standard annotations using the pre-processing steps that were described in Section 7.3.

To provide a fair comparison of the PhenoNorm method with other normalisation methods that have previously been applied to the same data sets, we used different metrics for evaluation (i.e., accuracy, precision, recall and F-measure). For the ShARe/CLEF task, participating systems were evaluated using the accuracy metric. Therefore, we evaluate the output of PhenoNorm on this corpus in terms of accuracy. However, for the three other corpora evaluated, i.e. the NCBI disease corpus, the heart failure corpus and the pulmonary embolism corpus, we used precision, recall and F-measure, to allow our results to be more easily compared to other normalisation methods reported for the disease corpus, and with IAA agreement for the heart failure and pulmonary embolism annotations.

7.5.1 ShARe/CLEF data set

As described in Section 2.2.1, ShARe/CLEF task 1 data set is a collection of 300 clinical records annotated for disorder mentions and linked to UMLS CUIs.

Our method achieved an accuracy of 0.83 when applied to this data set, and was able to address the variability in expressing the same concept. The main reason for the incorrectly predicted mappings was due to ambiguous abbreviations, which are very prevalent in this corpus. This is because it was designed specifically to address two different tasks, i.e., both disorder recognition and normalisation of acronyms/abbreviations to UMLS concepts. Accordingly, it includes a large number of challenging abbreviations and acronyms, which sometimes proved to be problematic for our method. For example, our method mapped the abbreviation "*3VD"* to the UMLS concept '*three vessel disease'*. However, the correct concept in the gold standard is '*triple vessel disease of the heart'*. Another source of error was that our method searches for the most similar concept in UMLS. However, in creating the gold standard links to UMLS in the ShARe /CLEF corpus, the annotators considered the textual context to determine the most relevant mapping. This is problematic for our method, since it considers only the annotation, rather than the surrounding context, when carrying out the mapping. As an example, the ShARe/CLEF corpus contains the annotated span "*recurrent ventral hernia"* which exactly matches the UMLS concept '*recurrent ventral hernia'*, and hence this is the mapping chosen by our method. However, presumably according to textual context,

the actual concept assigned in the ShARe/CLEF gold standard corpus is the more specific '*recurrent ventral incisional hernia'*. Table 7.6 provides examples that show the differences between the mappings produced by the PhenoNorm method and the gold standard annotation of ShARe/CLEF corpus.

Table 7.6 Differences between the UMLS mappings in the ShARe/CLEF gold standard annotations and those produced by PhenoNorm

ShARe /CLEF mentions	**ShARe /CLEF annotations**	**PhenoNorm mappings**
three-vessel coronary artery disease	triple vessel disease of the heart	multi vessel coronary artery disease
heart rhythm abnormality	cardiac arrhythmia	irregular heart rhythm
recurrent ventral hernia	recurrent ventral incisional hernia	recurrent ventral hernia
3VD	triple vessel disease of the heart	three vessel disease

7.5.2 NCBI disease corpus

The NCBI disease corpus [158] consists of 793 PubMed abstracts annotated for 6,892 disease mentions and mapped to 790 unique disease concepts in the MEDIC vocabulary [261] (which merges OMIM into the disease branch of MeSH).

We evaluated PhenoNorm in terms of its ability to normalise disease mentions in the test subset of the NCBI corpus to the most similar disease concept in the MEDIC database and the results are shown in Table 7.7, in which a comparison is made with the normalisation methods applied by Leaman et al., [262] to the same data sets. Our method is broadly comparable to several of the other methods applied, i.e.., those that use lexical normalisation (MetaMap and Norm[4] which is a tool for addressing the problem of name variation by normalising case, plurals, inflections and word order) and those that apply string similarity metrics i.e.., inference and cosine similarity methods. The inference method [263] works by applying a combination of rules that use the O (ND) difference string similarity algorithm [264]. Briefly, this algorithm works by finding the smallest number of edits needed to transform one string into another. However, our method cannot be compared to DNorm, since it is ML-based and it uses pairwise learning to learn the level of similarity between the disease

[4] https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/2013/docs/userDoc/tools/norm.html

mentions and MEDIC vocabularies. PhenoNorm outperforms all other methods apart from DNorm, and the recall of PhenoNorm is almost the same as for DNorm.

Table 7.7 Micro-averaged performance comparison of PhenoNorm against other normalisation approaches

Methods	P	R	F-measure
PhenoNorm	0.65	0.75	0.69
MetaMap	0.50	0.66	0.57
Norm	0.21	0.68	0.33
Inference method	0.53	0.66	0.59
Cosine similarity	0.64	0.71	0.67
DNorm	0.80	0.76	0.78

Analysing the results of PhenoNorm when applied to the NCBI corpus showed that it was able to handle term variations and make the correct decision most of the time. For example, PhenoNorm correctly mapped the disease mention "*familial neurohypophyseal diabetes insipidus"* to the correct concept in MEDIC, i.e., '*Diabetes Insipidus, Neurogenic'*.

The NCBI disease annotation guidelines instruct human annotators to assign disease mentions that could correspond to a complete family of more specific diseases to the more general concept in MEDIC. For example, according to the gold standard NCBI disease mappings, instances of "*complement deficiency"* mentions are mapped to the more general concept '*Immunologic Deficiency Syndromes'*. However, PhenoNorm mapped disease mentions to the most similar disease concepts in MEDIC and therefore, mentions of "*complement deficiency"* are mapped to '*C9 Deficiency'*.

As was the case for the PhenoCHF corpus, the pre-processing step of PhenoNorm that splits up coordinated noun phrases is useful in many cases. For example, it is able to correctly identify the four different types of cancer mentioned in the phrase "*breast, brain, prostate and kidney cancer"* and map them to four separate MEDIC concepts, i.e., '*breast neoplasms'*, '*brain neoplasms'*, '*prostatic neoplasms'* and '*kidney neoplasms'*. When applied to such phrases, PhenoNorm can sometimes produce mappings that are more correct than those produced by the best-performing DNorm method, which incorrectly mapped the coordinated phrase to a single concept, i.e., *'prostate cancer/brain cancer susceptibility'*. Despite the fact that the application of the pre-processing step to split up coordinated phrases is usually advantageous, it can

lead to errors in a small number of cases, since some mentions of single diseases contain coordinating conjunctions. This is the case, for example, for "*breast and ovarian cancer syndrome",* which corresponds to a concept in MEDIC. However, PhenoNorm incorrectly splits this coordinated phrase condition into two separate phrases, i.e., "*breast cancer syndrome"* and "*ovarian cancer syndrome",* which are incorrectly mapped to separate MEDIC concepts, i.e. *'breast cancer'* and '*ovarian neoplasms'*.

7.5.3 Annotations from heart failure and pulmonary embolism ontologies

Wang et al. [239] developed ontologies for phenotypic information (i.e., causes, sign or symptoms, diagnostic tests and treatments) pertaining to heart failure, rheumatoid arthritis and pulmonary embolism. The ontologies were curated manually using a semi-automatic annotation approach. Specifically, documents were automatically pre-annotated with a dictionary compiled using previously adjudicated annotations to annotate three corpora (i.e., one corpus for each of the three conditions introduced above). With these pre-annotated documents, annotators could modify or delete pre-annotations, or add missed occurrences of terms. Each corpus is compiled from different types of textual sources including textbooks, evidence-based online resources, practice guidelines and journal articles. The heart failure corpus includes 2588 annotations, while the rheumatoid arthritis and pulmonary embolism corpora are annotated with 193 and 425 mentions, respectively. To address lexical variations in concept mentions between the different sources (e.g., *alcohol consumption* and *alcoho*l *intake*), the annotations were mapped to UMLS concepts using MetaMap followed by manual verification and correction by one annotator. The accuracy of the mapping was evaluated by comparing the mappings assigned for 237 randomly selected terms against mappings produced by a second annotator. The mapping agreement between the two annotators, in terms of F-score, was 0.84.

We exploited the heart failure and pulmonary embolism ontologies to evaluate the ability of PhenoNorm to map concept mentions relating to heart failure and pulmonary embolism to the corresponding concepts in the ontologies; the results are shown in Table 7.8.

To the best of our knowledge, our work constitutes the first attempt to use these data sets to normalise phenotypic mentions relating to heart failure and pulmonary

embolism to concepts in the corresponding ontologies. As such, we cannot compare our results with any other normalisation approaches. However, for reference, we compare our results with the IAA results mentioned above. As can be observed in Table 7.8, the performance of PhenoNorm in normalising concept mentions relating to heart failure is almost the same as human levels of agreement.

Table 7.8 Results of applying PhenoNorm to the heart failure and pulmonary embolism data sets

Method	**Corpus**	**P**	**R**	**F**
PhenoNorm	**Pulmonary embolism**	0.75	0.77	0.76
PhenoNorm	**Heart failure**	0.82	0.86	0.83
IAA between MetaMap followed by manual corrections	**Randomly selected 237 phenotypic mentions**	-	-	0.84

As can be observed in Table 7.8, PhenoNorm achieved high recall for both corpora and in most cases it was able to handle term variations and associate mentions with the correct concept in the corresponding ontology. The pre-processing step of splitting coordinated phrases into two or more phrases was very useful for this data set and was able to help to map annotated phrases containing coordinations, such as "*stable or unstable angina"*, to appropriate separate concepts, in this case *'stable angina'* and *'unstable angina'*.

However, specific features of the PhenoNorm method, in combination with the mapping strategy applied by annotators in these corpora, resulted in some mapping errors. Firstly, as has been mentioned previously, PhenoNorm tends to assign a higher score to concepts that share the greatest number of words with the mention in question. Secondly, the heart failure and pulmonary embolism ontologies were manually curated and the guidelines instruct annotators to map each phenotypic mention to the concept that conveys the mention's specific meaning within the context of the original sentence. According to the above, our method caused a number of FNs and FPs in both the pulmonary embolism and heart failure corpora. For example, PhenoNorm maps the mention "*continuous blood pressure monitoring"* to '*blood pressure monitoring'* instead of *'Continuous Sphygmomanometers'*. Another source of incorrect decisions made by PhenoNorm was due to the strategy of mapping mentions

to concepts with the smallest edit distance. For example, it maps "*haemoglobin"* to '*haemoglobin low'* instead of to '*haemoglobin measurement'*. It should be noted, however, that the above mentioned features of PhenoNorm were highly advantageous in other cases, e.g., in facilitating the correct mapping of "*permanent pacemaker implantation"* to *'Implantation of permanent intravenous cardiac pacemaker'*.

7.6 Summary

In this chapter, we have described our development of a novel method, PhenoNorm, that allows variant mentions of different types of phenotypic entities to be mapped effectively to concepts in the UMLS Metathesaurus. This constitutes an important first step towards the automatic integration of information from these heterogeneous sources. Our method shows encouraging performance in mapping mentions with a variety of internal structures to the most appropriate concepts in the Metathesaurus. The range of different internal structures of concept mentions includes simple noun phrases (e.g. *progressive renal failure*), coordinated noun phrases (e.g., *increased chest pain and fatigue*), noun phrases followed by prepositional phrases (e.g., *increasing dyspnea on exertion*) and complete clauses or sentences (e.g., *jugular venous pressure is elevated*). The different mentions also exhibit a range of different term variation patterns such as different orderings of words, different forms of words and the use of semantically related words.

The expert evaluation/correction of the automatic mappings produced by our method adds value to the PhenoCHF corpus and allows it to be used as a gold standard to compare our method with state-of-art normalisation techniques such as hybrid string similarity methods (i.e., SoftTFIDF) and dictionary-based methods (i.e., MetaMap).

To demonstrate the wider applicability and the utility of our method, we have applied it to the task of normalising concepts in a number of different annotated corpora, whose concepts include gold standard mappings to concepts in different knowledge resources. Specifically, we have applied PhenoNorm to the ShARe/CLEF task 1 corpus, the NCBI disease corpus and the heart failure and pulmonary embolism datasets. We have demonstrated that our method can achieve competitive performance in carrying out mapping of concept mentions in all of these corpora, compared to the results achieved for PhenoCHF, despite the differing parameters and complexity of the

tasks. For the NCBI disease corpus, we also showed that PhenoNorm could outperform a number of different lexical and string similarity based methods when applied to the task of normalising disease names, and that the PhenoNorm method could produce comparable recall to the more sophisticated DNorm method, which is based on machine learning.

Chapter 8 Conclusion and future work

Biomedical text, including both literature articles and EHRs, constitutes a rich source of disease-phenotypic information. Since each of these text types contains different types of valuable information, combining details from each source can be important in uncovering new disease-phenotypic associations that may be of interest to clinicians and which can provide useful information to assist in clinical decision making and evidence-based health care. However, integrating information can be problematic if the same concepts are expressed in different ways in EHRs and articles.

Therefore, there is an emerging need for methods that automatically extract and integrate phenotypic information from EHRs and biomedical literature. However, developing such methods is reliant on the availability of corpora of clinical records and literature articles that are annotated for phenotypic information. However, acquiring corpora of clinical text can be particularly challenging, mainly due to privacy and confidentiality concerns that hamper ready access to clinical records.

In the following sections, we summarise the work described in this thesis and provide an outline of further research directions. Firstly, we revisit the research objectives that we established in Section 1.4.3 and explain how we fulfilled each one. Subsequently, we review the contributions of this study and summarise the main findings described in the preceding chapters. Finally, we conclude with a discussion of the main areas of future work.

8.1 Evaluation of Research Objectives

To obtain knowledge about the state-of-the-art in NER, we established objective O1 .

Objective O1

O1	To conduct a comprehensive review of existing resources, annotated corpora and approaches for clinical NER.

As an initial step towards achieving this objective, we analysed existing lexical resources, and presented our findings in Section 2.2. This enabled us to select the UMLS Metathesaurus as the most appropriate domain-specific knowledge resource

for our requirements, along with the associated MetaMap concept recognition tool, which recognises instances of Metathesaurus concepts that occur in text.

By examining a wide variety of biomedical and clinical corpora in section 2.2.1, we discovered that there are many publicly available corpora of scientific biomedical literature which are annotated for biological entities and/or their interactions. However, corpora containing clinical text drawn from EHRs are much rarer, which is mainly due to privacy and confidentiality concerns. Recently, however, a small number of annotated corpora containing clinical text have been made available, mainly in the context of shared task challenges. These corpora vary both in terms of the types of reports that have been drawn from EHRs for inclusion in the corpora (e.g., discharge summaries, progress notes and radiology reports) and in terms of the levels, types and granularity of the annotations that have been added to corpora. Despite the opportunities that such corpora offer in terms of advancing the state-of-the-art in clinical information extraction, very few of the above corpora (either clinical text or scientific literature) are annotated with the types of entities and relationships that are relevant to the study of phenotypic information.

In response to the above, we have created a new annotated corpus (PhenoCHF), to stimulate research into the automatic extraction of phenotypic information from free text. The PhenoCHF corpus integrates two text types, i.e., literature articles and discharge summaries from EHRs. The major part of the PhenoCHF corpus consists of discharge summaries from EHRs on the subject of CHF, drawn from the data released for the *i2b2 obesity and its co-comorbidities challenge*. The second part of the corpus consists of 10 full-text articles on the subject of CHF retrieved from the PubMed Central Open Access database. PhenoCHF has been annotated with various types of information relating to phenotype-disease associations by two domain experts (doctors). To our knowledge, the corpus is unique, both in terms of the level of detail of the phenotypic information annotated, and in that it integrates both literature articles and unstructured text reports from EHRs.

By reviewing previously reported approaches to clinical NER in Section 2.1.2.1, we determined that ML approaches have become one of the most commonly used approaches to NER. In particular, we found that previous solutions based on the CRF algorithm have frequently produced results that are superior to those obtained using other ML algorithms, especially when coupled with the use of a rich set of linguistic features.

Objective O2

O2	To apply NER techniques at a large scale to extract phenotypic information from both EHRs and literature articles.

We have achieved this objective by applying a common set of NER methods to extract phenotypic information from EHRs and from full-text articles, as described in Section 5.1.1.

To demonstrate the utility of PhenoCHF corpus in training ML-based phenotype extraction systems, we presented a comparative evaluation of different types of NER methods, i.e., rule-based (using the CAFETIERE system), dictionary-based (using MetaMap) and various ML approaches (i.e., CRF, MEMM and HMM). For the ML methods, we employed different sets of features that have previously been employed successfully in other NER efforts in the biomedical and clinical domains. These features included syntactic features (e.g., POS and chunk tags) and morphological features (i.e., prefixes and suffixes). We carried out an assessment of the contribution of different feature sets toward the performance of different ML algorithms when applied to the task of phenotypic NER.

While the rule-based method achieved the best results, the results obtained by certain ML algorithms and feature sets compare extremely favourably with the results achieved by the rules. ML-based methods also present the advantage of being far less time consuming than manually derived rules.

For literature articles, the discrepancy between the results achieved by the rule-based method and ML-based methods is considerably greater than for clinical records. This can be explained by the smaller size of the article subset, and the scarcity of its annotations, compared to those in the clinical records. This means that ML models trained only on the literature articles had fewer observations from which to learn how to recognise entities accurately.

As the CRF algorithm achieved the best performance amongst the different ML algorithms applied to both the discharge summary and article subsets of PhenoCHF, we employed CRF in carrying out further experiments. Specifically, we demonstrated that a CRF ML tagger trained to recognise phenotypic information in one type of text (i.e. clinical records) is fairly robust when applied to another type of text (i.e.,

literature articles). We also explored the use of a pooled corpus of heterogeneous textual sources (i.e., EHRs and literature articles), annotated according to a common set of guidelines, to train a single classifier. Our results showed that such a trained model is robust to variations in text type and exhibits superior performance to models trained only on a single text type.

The portability and the robustness of the pooled model to different text types, together with the superiority of CRF models in this context, were further reinforced through our application of PhenoCHF-trained pooled model to related annotated corpora (i.e., ShARe/ CLEF 2013, HD risk factor and COPD phenotype), for which encouraging results were obtained.

Objective O3

O3 To conduct a comprehensive review of existing corpora and approaches for clinical relation extraction.

In section 2.2.1, a review of clinical corpora annotated for relations (e.g., i2b2 and CLEF) revealed that, although a variety of binary relations has previously been annotated, none of the currently available corpora is annotated for *n*-ary relations (where n > 2). Our review thus determined that there is a research gap in terms of extracting complex (*n*-ary) relations from clinical records.

A review of existing approaches to the automatic extraction of binary relations from texts in the clinical domain was also presented in Section 2.1.2.2. Whilst earlier work focussed on rule-based approaches and co-occurrence of pairs of entities, more recent work has explored ML approaches.

A straightforward way to extend binary relation extraction to *n*-ary relation extraction is to factorise the *n*-ary relations into binary relations and then to apply binary classification methods to extract the factorised relations. However, an issue in directly applying this method to the extraction of *n*-ary relations is that, as *n* increases, the factorisation will lead to a large increase in the number of candidate binary relations. This in turn will result in low performance when extracting *n*-ary relations.

In our context, it is very important to facilitate joint detection and semantic categorisation of all *n* related arguments, and also to determine whether the relation is negated.

Unlike methods for binary relation extraction, event extraction systems (e.g., EventMine), reviewed in Section 2.1.2.3, have achieved state-of-the-art results in extracting complex biomedical events, and they are capable of capturing many different types of associations, in which an arbitrary number of entities is semantically characterised through the assignment of a variety of semantic roles. Furthermore, event extraction systems are able to detect different aspects of meta-knowledge, i.e., how events should be interpreted according to their textual contexts, as a subtask of the event extraction process. Such meta-knowledge includes the detection of when events are negated.

The above findings were used to help us to accomplish our research objective O4**.**

.

Objective O4

O4 To adapt TM tools currently used to extract relations and events from full papers and abstracts and make them suitable for extracting the relations between phenotypic entities in EHRs.

This objective was also established at the beginning of this research, with the aim of exploring how *n*-ary clinical relations could be extracted efficiently from clinical records.

We accomplished this objective by adapting EventMine, coupled with the application of a number of post-processing rules, which use the outputs of the Enju and GDep parsers to refine trigger detection, as described in Chapter 6.

We compared the performance achieved by EventMine to the performance of state-of-the-art supervised ML methods for binary relation extraction, using a rich set of features. Our results demonstrated that EventMine (with post-processing rules), was able to outperform the binary relation extraction methods, proving that it is a useful and successful means of detecting complex relations within medical records.

To the best of our knowledge, our work constitutes the first research effort to explore and evaluate this methodology as a solution for *n*-ary relation extraction.

Objective O5

O5 To develop a novel method to normalise phenotypic concept mentions from heterogeneous textual sources (EHRs and literature articles) and to map them to UMLS concepts.

One of the major outcomes of this work, geared towards the fulfilment of objective O5 , was the development of a novel method (i.e., PhenoNorm). As described in Chapter 7, PhenoNorm is able to normalise/map different textual mentions/variations of a given phenotypic concept to an appropriate concept in the UMLS Methesaurus. Given that concepts can be expressed in various ways in different types of text, the normalisation process constitutes an important first step towards bridging the gap between the information contained within the two different text types that we have considered (i.e., EHRs and the biomedical literature).

PhenoNorm is a hybrid method that integrates token-based, character-based and semantic similarity measures to allow variant mentions of different phenotypic concepts to be mapped effectively to concepts in the UMLS Metathesaurus. The results of the automatic mapping carried out by PhenoNorm were manually evaluated by a domain expert. The evaluation confirmed that PhenoNorm is able to handle a variety of types of term variations with different internal structures.

The expert evaluation/correction of the automatic mappings produced by our methods adds value to the PhenoCHF corpus, since the corrected mappings may be used in future as a gold standard for the training and evaluation of ML normalisation methods. We used the expert-produced gold standard mappings as a means to compare the performance of the PhenoNorm method with other state-of-the-art methods for normalisation, i.e., dictionary-based (MetaMap) and string similarity (SoftTFIDF) methods. PhenoNorm achieves higher accuracy than the compared methods, and shows encouraging performance in terms of its ability to map phenotype mentions with a variety of internal structures, and which exhibit a range of different term variation patterns, to appropriate UMLS concepts.

To further demonstrate the utility of our normalisation method, we applied it to the tasks of normalising disorder mentions in the ShARe/CLEF 2013 corpus, disease mentions in the test set of NCBI disease corpus and phenotypic mentions related to heart failure and pulmonary embolism in the corpora used to develop heart failure and

pulmonary embolism ontologies. Despite the differing parameters and levels of complexity of these tasks, PhenoNorm achieved results that compare favourably both to the results obtained when the method is applied to the PhenoCHF corpus, and to results of other normalisation methods that have been previously applied to the NCBI corpus. These results help to prove the wider applicability and utility of our method in linking a wide range of biomedical concept mentions to appropriate ontology concepts.

Objective O6

O6 To integrate information extracted from both text types using the normalisation approach.

We accomplished this objective through the application of our novel PhenoNorm normalisation method. Phenotypic mentions from both text types (i.e., EHRs and the literature articles) were mapped to unique concepts in UMLS using this method. Based upon the output of the method, we found that the PhenoCHF corpus mentions 835 unique phenotypic concepts relating to CHF. Of these concepts, 184 occur in both EHRs and literature articles. Thus, the normalisation process allows the identification of concepts that are shared between the two text types. These shared concepts represent a link which can be used to bridge the gap between the different kinds of information that are present in the two document types contained within the PhenoCHF corpus. Examining instances of these shared concepts in the different text sources can represent a first step towards understanding how complementary information within the different resources can be integrated in an effective manner.

The findings and results summarised above demonstrate that we have successfully fulfilled our research objectives. In turn, the results obtained through completing our research objectives serve to prove the research hypotheses formulated at the beginning of this thesis:

H1 Existing text mining techniques can be adapted to extract phenotypic information from the overwhelming volume of information in the literature

and EHRs, and to discover hidden knowledge and associations that may occur across texts of different types.

H2 *N*-ary relations between phenotypic entities can be cast as events, and they can be extracted using an event-based approach.

H3 Normalising various types of phenotypic information that appear in both EHRs and literature articles can act as a first step towards the automatic integration of knowledge that is dispersed within these two text types.

8.2 Future work

There is evidence of a 13−17 year gap between the point at which research findings are reported in the literature and the time when they are put into practice in the context of clinical care [16]. This reality suggests that the current methods of making use of scientific results within actual clinical care are severely lacking. As a result of this time lag, evidence-based treatments derived from research are often out-of-date by the time they achieve widespread use, and do not always account for real-world variation. These factors can significantly impede the effective implementation of treatments based on research findings.

The work in this thesis presents the first step towards the automatic integration of phenotypic information occurring in clinical records and the biomedical literature. Such integration can fill in the knowledge gaps that result from the discrepancies between the types of information that are present in EHRs on the one hand, and in the biomedical literature on the other. Through the integration of the complementary information contained within these two textual sources, the valuable information in EHRs can be further augmented with relevant information and findings contained within the vast amount of published biomedical literature, in multiple ways. For example, the integration can be helpful in identifying diseases and their associated symptoms, potential treatments, genes responsible for the disease and in determining adverse drug events. Discovery of all of these types of associations is unlikely to be possible by considering only one of the text types in isolation.

Furthermore, multiple threads of novel research can be carried out, using the outcomes of our work as a basis. Below, we review some of the main areas of future work.

As a result of carrying out the work described in this thesis, we have developed novel resources, i.e., the annotated PhenoCHF corpus and the models trained using the annotations within the corpus. PhenoCHF is richly annotated for CHF phenotypes that are linked to canonical forms in the UMLS Metathesaurus. By making these fine-grained phenotype annotations and their links to ontological concepts available in a computable form, they are suitable for use in computerised applications such as CDSS and clinical research. For example, the application of the models to large amounts of patient data can facilitate the extraction of phenotypic information associated with different diseases which, in turn, can be profiled and used to construct the backbone of CDSS, driven by live data based on the actual population. Phenotypic information associated with different diseases will further enable researchers to identify patient cohorts based on particular sets of phenotypic features that are shared amongst particular groups of patients. In turn, the identification of these cohorts will facilitate clinical trial enrolment and support clinical decision support.

The ability to obtain detailed phenotypic information for different diseases can help to build up a clearer picture about individual patients' drug reactions. This in turn can facilitate phenotype-driven treatment of diseases and pave the way for personalised healthcare.

Understanding diseases at the phenotypic level will also help to guide improvements in diagnosis, prevention and discovery of the origins of diseases. For instance, an interesting extension of this thesis would be to link information in EHRs with information in the biomedical literature by applying a text mining approach to literature-based discovery, targeted specifically at understanding phenotype-genotype associations that may be dispersed across different types of biomedical text.

Another obvious extension of this thesis is to broaden the scope of phenotypes that can be recognised automatically (e.g., treatments and tests could additionally be included), as well as to extend the range of detected relations that can hold between phenotypes. Examples of further relevant relation types could include *cure* and *diagnose*. Although we designed our methods with a particular focus on the CHF disease, they are sufficiently general to allow their application to other diseases, as long as suitably annotated corpora are available. We have already partially

demonstrated the generality of our methods by successfully applying them to corpora whose scopes only partially overlap with the PhenoCHF corpus. The proposed methods for extracting phenotypic information in this thesis can be incorporated with other text mining tools within interoperable Web-based text mining platforms (e.g., Argo [45]) to generate semi-automatic annotation workflows; these can help to accelerate the manual annotation process when producing further annotated corpora.

Besides the integration of information contained within clinical records and biomedical literature, it would be useful to investigate the additional integration of information contained within various social media channels, such as Facebook and Twitter [265, 266], as well as other social networking web sites such as PatientsLikeMe [267]. These data sources are highly valuable, since they are able to provide the most up-to-date information about diseases, as they occur in the real world. Within these fora, users frequently supply first-hand information regarding their health status, adverse effects of medication they are taking, etc. Taking into account information contained within social media is becoming increasingly important, given the recent steep increase in the volume of potentially useful information that is publicly shared via these channels. We envisage that future confluence of details contained within social networks and EHRs will open up new ways to manage diseases and treatments. Nevertheless, it is important to bear in mind that the variable reliability of the first-hand information provided within social media channels is an important issue that must be taken into account when determining the most effective strategies for combining this information with that originating from health professionals.

Chapter 9 References

1. Chen, H., Fuller, S.S., Friedman, C., and Hersh, W. *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*. Vol. 8. 2006: Springer Science & Business Media.
2. Richesson, R. and Smerek, M. *Electronic health records-based phenotyping.* Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials. Acessed: April 2016.
3. Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W., and Shen, B. *Biomedical text mining and its applications in cancer research.* Journal of Biomedical Informatics, 2013. **46**(2): p. 200–211.
4. Kohane, I.S. *The twin questions of personalized medicine: who are you and whom do you most resemble.* Genome Medicine, 2009. **1**(1): p. 4.
5. Zhou, L. and Xu, H. *Biomedical scientific textual data types and processing*, in *Encyclopedia of Database Systems*. 2009, Springer. p. 233–236.
6. Erhardt, R.A., Schneider, R., and Blaschke, C. *Status of text-mining techniques applied to biomedical text.* Drug Discovery Today, 2006. **11**(7): p. 315–325.
7. Cohen, A.M. and Hersh, W.R. *A survey of current work in biomedical text mining.* Briefings in Bioinformatics, 2005. **6**(1): p. 57–71.
8. *U.S. National Library of Medicine. Yearly Citation Totals from 2015 MEDLINE/PubMed Baseline. Accessed April 2016.*
9. Zweigenbaum, P., Demner-Fushman, D., Yu, H., and Cohen, K.B. *Frontiers of biomedical text mining: current progress.* Briefings in Bioinformatics, 2007. **8**(5): p. 358–375.
10. Friedman, C., Alderson, P.O., Austin, J.H., Cimino, J.J., and Johnson, S.B. *A general natural-language text processor for clinical radiology.* American Medical Informatics Association, 1994. **1**(2): p. 161–174.
11. Friedman, C., Knirsch, C., Shagina, L., and Hripcsak, G. *Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries.* American Medical Informatics Association 1999: p. 256.

12. Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., and Setzer, A. *Building a semantically annotated corpus of clinical texts.* Biomedical Informatics, 2009. **42**(5): p. 950–966.
13. Gundersen, M.L., Haug, P.J., Pryor, T.A., van Bree, R., Koehler, S., Bauer, K., and Clemons, B. *Development and evaluation of a computerized admission diagnoses encoding system.* Computers and Biomedical Research, 1996. **29**(5): p. 351–372.
14. Wright, A., Chen, E.S., and Maloney, F.L. *An automated technique for identifying associations between medications, laboratory results and problems.* Biomedical Informatics, 2010. **43**(6): p. 891–901.
15. Lin, R., Lenert, L., Middleton, B., and Shiffman, S. *A free-text processing system to capture physical findings: Canonical Phrase Identification System (CAPIS)*. In *Proceedings of the Annual Symposium on Computer Applications in Medical Care*. 1991. American Medical Informatics Association.p. 843.
16. Bennett, C. and Doub, T. *Data mining and electronic health records: selecting optimal clinical treatments in practice*. In *Proceedings of the 6th International Conference on Data Mining*. 2011.p. 313–318.
17. Lippincott, T., Séaghdha, D.Ó., and Korhonen, A. *Exploring subdomain variation in biomedical language.* BMC Bioinformatics, 2011. **12**(1): p. 212.
18. Ananiadou, S. and McNaught, J. *Text Mining for Biology and Biomedicine*. 2006: Artech House Boston, London.
19. Shatkay, H. and Feldman, R. *Mining the Biomedical Literature in the Genomic Era: an Overview.* Computational Biology, 2003. **10**(6): p. 821–855.
20. Simpson, M.S. and Demner-Fushman, D. *Biomedical text mining: A survey of recent progress*, in *Mining Text Data*. 2012, Springer. p. 465–517.
21. Settles, B. *ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text.* Bioinformatics, 2005. **21**(14): p. 3191–3192.
22. Sasaki, Y., Tsuruoka, Y., McNaught, J., and Ananiadou, S. *How to make the most of NE dictionaries in statistical NER.* BMC Bioinformatics, 2008. **9**(Suppl 11): p. S5.
23. Sætre, R., Sagae, K., and Tsujii, J.i. *Syntactic features for protein-protein interaction extraction*. In *Proceedings of Languages in Biology and Medicine (Short Papers)*. 2007.

24. Hunter, L., Lu, Z., Firby, J., Baumgartner, W.A., Johnson, H.L., Ogren, P.V., and Cohen, K.B. *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, 2008. **9**(1): p. 78.

25. Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., and Salakoski, T. *Extracting contextualized complex biological events with rich graph-based feature set.* Computational Intelligence, 2011. **27**(4): p. 541–557.

26. Miwa, M., Thompson, P., and Ananiadou, S. *Boosting automatic event extraction from the literature using domain adaptation and coreference resolution.* Bioinformatics, 2012. **28**(13): p. 1759–1765.

27. Uzuner, Ö., Luo, Y., and Szolovits, P. *Evaluating the state-of-the-art in automatic de-identification.* American Medical Informatics Association, 2007. **14**(5): p. 550–563.

28. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Pradhan, S., South, B.R., Mowery, D.L., and Jones, G.J. *Overview of the ShARe/CLEF eHealth Evaluation Lab 2013*. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*. 2013. Springer.p. 212–231.

29. Pradhan, S., Elhadad, N., Chapman, W., Manandhar, S., and Savova, G. *Semeval-2014 task 7: Analysis of clinical text.* SemEval 2014, 2014. **199**(99): p. 54.

30. Uzuner, Ö., Goldstein, I., Luo, Y., and Kohane, I. *Identifying patient smoking status from medical discharge records.* American Medical Informatics Association, 2008. **15**(1): p. 14–24.

31. Uzuner, Ö. *Recognizing obesity and comorbidities in sparse data.* American Medical Informatics Association, 2009. **16**(4): p. 561–570.

32. Uzuner, Ö., South, B.R., Shen, S., and DuVall, S.L. *2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text.* American Medical Informatics Association, 2011. **18**(5): p. 552–556.

33. Bodenreider, O. *The Unified Medical Language System (UMLS): Integrating Biomedical Terminology.* Nucleic Acids Research, 2004. **32**(suppl 1): p. D267–D270.

34. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. *Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders.* Nucleic Acids Research, 2005. **33**(suppl 1): p. D514–D517.

35. Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., and Campbell, J. *The Human Phenotype Ontology Project: Linking Molecular Biology and Disease through Phenotype Data.* Nucleic Acids Research, 2013: p. gkt1026.

36. yam Khordad, M., Mercer, R.E., and Rogan, P. *A machine learning approach for phenotype name recognition*. In *Proceedings of Conference on Computational Linguistics*. 2012.

37. *World Health Organization. Cardiovascular diseases (CVDs).* http://www.who.int/cardiovascular_diseases/en/. Accessed April 2016.

38. *World Health Organization. The top 10 causes of death.* http://www.who.int/mediacentre/factsheets/fs310/en/. Accessed April 2016.

39. Shah, A.S., Griffiths, M., Lee, K.K., McAllister, D.A., Hunter, A.L., Ferry, A.V., Cruikshank, A., Reid, A., Stoddart, M., and Strachan, F. *High sensitivity cardiac troponin and the under-diagnosis of myocardial infarction in women: prospective cohort study.* BMJ, 2015. **350**: p. g7873.

40. Shiba, N. and Shimokawa, H. *Chronic Kidney Disease and Heart Failure—Bidirectional Close Link and Common Therapeutic Goal.* Cardiology, 2011. **57**(1): p. 8–17.

41. Chan, E.J. and Dellsperger, K.C. *Update on cardiorenal syndrome: a clinical conundrum.* Adv Perit Dial, 2011. **27**: p. 82–86.

42. Swanson, D.R. *Two medical literatures that are logically but not bibliographically connected.* American Society for Information Science, 1987. **38**(4): p. 228–233.

43. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., and Hurdle, J.F. *Extracting information from textual documents in the electronic health record: a review of recent research.* Yearbook of Medical Informatics, 2008. **35**: p. 128–44.

44. Kano, Y., Baumgartner, W.A., McCrohon, L., Ananiadou, S., Cohen, K.B., Hunter, L., and Tsujii, J.i. *U-Compare: share and compare text mining tools with UIMA.* Bioinformatics, 2009. **25**(15): p. 1997–1998.

45. Rak, R., Rowley, A., Black, W., and Ananiadou, S. *Argo: an integrative, interactive, text mining-based workbench supporting curation.* Database, 2012. **2012**: p. bas010.
46. Leaman, R. and Gonzalez, G. *BANNER: an executable survey of advances in biomedical named entity recognition*. In *Pacific Symposium on Biocomputing*. 2008. World Scientific.p. 652–663.
47. Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., and Tsujii, J.i. *Developing a robust part-of-speech tagger for biomedical text*, in *Advances in Informatics*. 2005, Springer. p. 382–392.
48. Seckman, C. *Electronic health records and applications for managing patient care*. 2013, Elsevier and typesetter Toppan Bes.
49. Patrick, J., Wang, Y., and Budd, P. *Automatic mapping clinical notes to medical terminologies*. In *Australasian Language Technology Workshop*. 2006.
50. Gundlapalli, A., South, B., Phansalkar, S., Kinney, A., Shen, S., Delisle, S., Perl , T., and Samore, M. *Application of natural language processing to VA electronic health records to identify phenotypic characteristics for clinical and research purposes.* American Medical Informatics Association, 2008: p. 36–40.
51. Birman-Deych, E., Waterman, A.D., Yan, Y., Nilasena, D.S., Radford, M.J., and Gage, B.F. *Accuracy of ICD-9-CM Codes for Identifying Cardiovascular and Stroke Risk Factors.* Medical Care, 2005. **43**(5): p. 480–485.
52. Li, L., Chase, H.S., Patel, C.O., Friedman, C., and Weng, C. *Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study.* American Medical Informatics Association, 2008. **2008**: p. 404.
53. Sager, N., Friedman, C., Chi, E., Macleod, C., Chen, S., and Johnson, S. *The Analysis and Processing of Clinical Narrative*. In *Proceedings of the Fifth Conference on Medical Informatics (MedInfo)*. 1986.p. 1101–5.
54. Sager, N., Friedman, C., and Lyman, M.S. *Medical language processing: computer management of narrative data*. 1987: Addison Wesley Longman Publishing Company. 57–65.
55. Krauthammer, M. and Nenadic, G. *Term identification in the biomedical literature.* Biomedical Informatics, 2004. **37**(6): p. 512–526.

56. Savova, G., Masanz, J., Orgen , P., Zheng, J., Shon, S., Kipper-Schuler, K., and Chute, C. *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, Component Evaluation and Applications.* American Medical Association, 2010. **17**: p. 507–513.
57. Ferrucci, D. and Lally, A. *UIMA: an architectural approach to unstructured information processing in the corporate research environment.* Natural Language Engineering, 2004. **10**(3-4): p. 327–348.
58. Savova, G.K., Fan, J., Ye, Z., Murphy, S.P., Zheng, J., Chute, C.G., and Kullo, I.J. *Discovering peripheral arterial disease cases from radiology notes using natural language processing*. In *Proceedings of American Medical Informatics Association*. 2010.p. 722.
59. Savova, G.K., Ogren, P.V., Duffy, P.H., Buntrock, J.D., and Chute, C.G. *Mayo Clinic NLP system for patient smoking status identification.* American Medical Informatics Association, 2008. **15**(1): p. 25–28.
60. Zeng, Q.T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S.N., and Lazarus, R. *Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system.* BMC Medical Informatics and Decision Making, 2006. **6**(1): p. 30.
61. Cunningham, H., Humphreys, K., Gaizauskas, R., and Wilks, Y. *GATE–a TIPSTER-based general architecture for text engineering*. In *Proceedings of the TIPSTER Text Program (Phase III)*. 1997.
62. Demner-Fushman, D., Chapman, W.W., and McDonald, C.J. *What can natural language processing do for clinical decision support?* Biomedical Informatics, 2009. **42**(5): p. 760–772.
63. Aronson Alan. *Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program.* American Medical Informatics Association:AMIA, 2001: p. 17-21.
64. Pratt, W. and Yetisgen-Yildiz, M. *A study of biomedical concept identification: MetaMap vs. people*. In *Proceedings of the American Medical Informatics Association*. 2003.p. 529.
65. Weeber, M., Klein, H., Aronson, A.R., Mork, J.G., De Jong-van Den Berg, L., and Vos, R. *Text-based discovery in biomedicine: the architecture of the DAD-system*. In *Proceedings of the American Medical Informatics Association*. 2000.p. 903.

66. Meystre, S. and Haug, P.J. *Natural language processing to extract medical problems from electronic clinical documents: performance evaluation.* Journal of Biomedical Informatics, 2006. **39**(6): p. 589–599.

67. Friedman, C., Shagina, L., Socratous, S.A., and Zeng, X. *A WEB-based version of MedLEE: A medical language extraction and encoding system.* American Medical Informatics Association, 1996: p. 938.

68. Friedman, C., Shagina, L., Lussier, Y., and Hripcsak, G. *Automated encoding of clinical documents based on natural language processing.* American Medical Informatics Association, 2004. **11**(5): p. 392–402.

69. Lussier, Y. and Friedman, C. *BiomedLEE: a natural-language processor for extracting and representing phenotypes, underlying molecular mechanisms and their relationships.* Intelligent Systems for Molecular Biology, 2007.

70. Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A. *GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles.* Bioinformatics, 2001. **17**(suppl 1): p. S74–S82.

71. Mihaila, C., Batista-Navarro, R., Alnazzawi, N., Kontonatsios, G., Korkontzelos, I., Rak, R., Thompson, P., and Ananiadou, S. *Mining the Biomedical Literature*. Healthcare Data Analytics. Vol. 36. 2015. 251.

72. Roberts, A., Gaizauskas, R.J., Hepple, M., and Guo, Y. *Combining terminology resources and statistical methods for entity recognition: an evaluation*. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 2008.

73. Dehghan, A., Keane, J., and Nenadic, G. *Challenges in clinical named entity recognition for decision support*. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*. 2013. IEEE.p. 947–951.

74. Nadeau, D. and Sekine, S. *A survey of named entity recognition and classification.* Lingvisticae Investigationes, 2007. **30**(1): p. 3–26.

75. Gaizauskas, R., Hepple, M., Davis, N., Guo, Y., Harkema, H., Roberts, A., and Roberts, I. *AMBIT: Acquiring medical and biological information from text*. In *Proceedings of the UK e-Science All Hands Meeting*. 2003. Nottingham, UK.

76. Gaizauskas, R., Demetriou, G., Artymiuk, P.J., and Willett, P. *Protein structures and information extraction from biological texts: the PASTA system.* Bioinformatics, 2003. **19**(1): p. 135–143.

77. Yang, H. *Automatic extraction of medication information from medical discharge summaries.* American Medical Informatics Association, 2010. **17**(5): p. 545–548.
78. Uzuner, Ö., Solti, I., and Cadag, E. *Extracting medication information from clinical text.* American Medical Informatics Association, 2010. **17**(5): p. 514–518.
79. Childs, L.C., Enelow, R., Simonsen, L., Heintzelman, N.H., Kowalski, K.M., and Taylor, R.J. *Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data.* American Medical Informatics Association, 2009. **16**(4): p. 571–575.
80. Li, D., Kipper-Schuler, K., and Savova, G. *Conditional Random Fields and Support Vector Machines for disorder named entity recognition in clinical texts*. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. 2008. Association for Computational Linguistics.p. 94–95.
81. Nadeau, D. *Semi-supervised named entity recognition*. 2007, University of Ottawa
82. Yeh, A., Morgan, A., Colosimo, M., and Hirschman, L. *BioCreAtIvE task 1A: gene mention finding evaluation.* BMC Bioinformatics, 2005. **6**(Suppl 1): p. S2.
83. Shen, D., Zhang, J., Zhou, G., Su, J., and Tan, C.-L. *Effective adaptation of a Hidden Markov Model-based named entity recognizer for biomedical domain*. In *Proceedings of the Association for Computational Linguistics*. 2003.p. 49–56.
84. de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J., and Zhu, X. *Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010.* American Medical Informatics Association, 2011. **18**(5): p. 557–562.
85. Kazama, J.i., Makino, T., Ohta, Y., and Tsujii, J.i. *Tuning support vector machines for biomedical named entity recognition*. In *Proceedings of the Association for Computational Linguistics*. 2002.p. 1–8.
86. Settles, B. *Biomedical named entity recognition using conditional random fields and rich feature sets*. In *Proceedings of the International Joint*

Workshop on Natural Language Processing in Biomedicine and its Applications. 2004. Association for Computational Linguistics.p. 104–107.

87. Pathak, P., Goswami, R., Joshi, G., Patel, P., and Patel, A. *CRF-based Clinical Named Entity Recognition using clinical NLP*. In *Proceedings of International Conference on Natural Language Processing*.

88. Tang, B., Wu, Y., Jiang, M., Denny, J.C., and Xu, H. *Recognizing and encoding discorder concepts in clinical text using machine learning and vector space model*. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab 2013*. 2013.

89. Jiang, M., Chen, Y., Liu, M., Rosenbloom, S.T., Mani, S., Denny, J.C., and Xu, H. *A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries.* American Medical Informatics Association, 2011. **18**(5): p. 601–606.

90. Denny, J.C., Miller, R.A., Johnson, K.B., and Spickard III, A. *Development and evaluation of a clinical note section header terminology.* American Medical Informatics Association.

91. Wang, C. and Akella, R. *A hybrid approach to extracting disorder mentions from clinical notes.* American Medical Informatics Association, 2015.

92. Chen, E.S., Hripcsak, G., Xu, H., Markatou, M., and Friedman, C. *Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study.* Journal of the American Medical Informatics Association, 2008. **15**(1): p. 87–98.

93. Mallory, E.K., Zhang, C., Ré, C., and Altman, R.B. *Large-scale extraction of gene interactions from full-text literature using DeepDive.* Bioinformatics, 2015: p. btv476.

94. Lee, L.C., Horn, F., and Cohen, F.E. *Automatic extraction of protein point mutations using a graph bigram association.* PLOS Computational Biology, 2007. **3**(2): p. e16.

95. Chang, D.T.-H., Weng, Y.-Z., Lin, J.-H., Hwang, M.-J., and Oyang, Y.-J. *Protemot: Prediction of Protein Binding Sites with Automatically Extracted Geometrical Templates.* Nucleic Acids Research, 2006. **34**(suppl 2): p. W303–W309.

96. Chun, H.-W., Tsuruoka, Y., Kim, J.-D., Shiba, R., Nagata, N., Hishiki, T., and Jun'ichi, T. *Extraction of gene-disease relations from Medline using domain dictionaries and machine learning.* 2006. **11**: p. 4–15.

97. Czarnecki, J., Nobeli, I., Smith, A.M., and Shepherd, A.J. *A text-mining system for extracting metabolic reactions from full-text articles.* BMC Bioinformatics, 2012. **13**(1): p. 1.

98. Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Duboué, P.A., Weng, W., and Wilbur, W.J. *GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data.* Biomedical Informatics, 2004. **37**(1): p. 43–53.

99. Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y., and Friedman, C. *PheneGo: Assigning phenotypic context to gene ontology annotations with natural language processing.* Pacific Symposium on Biocomputing., 2006: p. 64–75.

100. Krallinger, M., Leitner, F., Rodriguez-Penagos, C., and Valencia, A. *Overview of the protein-protein interaction annotation extraction task of BioCreative II.* Genome Biology, 2008. **9**(Suppl 2): p. S4.

101. Jonnalagadda, S. *An effective approach to biomedical information extraction with limited Training Data*. 2011, Arizona State University.

102. Grishman, R., Huttunen, S., and Yangarber, R. *Information extraction for enhanced access to disease outbreak reports.* Biomedical Informatics, 2002. **35**(4): p. 236–246.

103. Roberts, A., Gaizauskas, R., and Hepple, M. *Extracting clinical relationships from patient narratives*. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. 2008. Columbus, Ohio: Association for Computational Linguistics.p. 10–18.

104. Manning, C.D. and Schütze, H. *Foundations of statistical natural language processing*. Vol. 999. 1999: Cambridge: MIT Press.

105. Wang, X., Chase, H., Markatou, M., Hripcsak, G., and Friedman, C. *Selecting information in electronic health records for knowledge acquisition.* Biomedical Informatics, 2010. **43**(4): p. 595–601.

106. Šarić, J., Jensen, L.J., Ouzounova, R., Rojas, I., and Bork, P. *Extraction of regulatory gene/protein networks from Medline.* Bioinformatics, 2006. **22**(6): p. 645–650.

107. Fundel, K., Küffner, R., and Zimmer, R. *RelEx—Relation extraction using dependency parse trees.* Bioinformatics, 2007. **23**(3): p. 365–371.

108. Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T., and Tsujii, J.i. *Semantic retrieval for the accurate identification of relational concepts in massive textbases*. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. 2006.p. 1017–1024.

109. Rindflesch, T.C. and Fiszman, M. *The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text.* Biomedical Informatics, 2003. **36**(6): p. 462–477.

110. Rindflesch, T.C., Kilicoglu, H., Fiszman, M., Rosemblat, G., Shin, D., Kilicoglu, H., Fiszman, M., Rosemblat, G., and Shin, D. *Semantic MEDLINE: An advanced information management application for biomedicine.* Information Services & Use, 2011. **31**(1-2): p. 15–21.

111. Hristovski, D., Dinevski, D., Kastrin, A., and Rindflesch, T.C. *Biomedical question answering using semantic relations.* BMC Bioinformatics, 2015. **16**(1): p. 6.

112. Fiszman, M., Rindflesch, T.C., and Kilicoglu, H. *Abstraction summarization for managing the biomedical research literature*. In *Proceedings of the Human Language Technologies-North American Chapter of the Association for Computational Linguistics: Workshop on Computational Lexical Semantics*. 2004. Association for Computational Linguistics.p. 76–83.

113. Hristovski, D., Friedman, C., Rindflesch, T.C., and Peterlin, B. *Exploiting semantic relations for literature-based discovery*. American Medical Informatics Association, 2006. **2006**: p. 349.

114. Bejan, C.A. and Denny, J.C. *Learning to identify treatment relations in clinical text.* American Medical Informatics Association, 2014. **2014**: p. 282.

115. Nguyen, N.T., Miwa, M., Tsuruoka, Y., Chikayama, T., and Tojo, S. *Wide-coverage relation extraction from MEDLINE using deep syntax.* BMC Bioinformatics, 2015. **16**(1): p. 107.

116. Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., and Salakoski, T. *BioInfer: a corpus for information extraction in the biomedical domain.* BMC Bioinformatics, 2007. **8**(1): p. 50.

117. Leser, U. and Hakenberg, J. *What makes a gene name? Named entity recognition in the biomedical literature.* Briefings in Bioinformatics, 2005. **6**(4): p. 357–369.
118. Grover, C., Haddow, B., Klein, E., Matthews, M., Nielsen, L.A., Tobin, R., and Wang, X. *Adapting a relation extraction pipeline for the BioCreAtIvE II task*. In *Proceedings of the BioCreAtIvE II Workshop*. 2007.p. 273–286.
119. Bundschus, M., Dejori, M., Stetter, M., Tresp, V., and Kriegel, H.-P. *Extraction of semantic biomedical relations from text using conditional random fields.* BMC Bioinformatics, 2008. **9**(1): p. 207.
120. Giuliano, C., Lavelli, A., and Romano, L. *Exploiting shallow linguistic information for relation extraction from biomedical literature*. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. 2006.p. 401–408.
121. Katrenko, S. and Adriaans, P. *Learning relations from biomedical corpora using dependency trees*, in *Knowledge Discovery and Emergent Complexity in Bioinformatics*, K. Tuyls, et al., Editors. 2006, Springer. p. 61–80.
122. Rink, B., Harabagiu, S., and Roberts, K. *Automatic extraction of relations between medical concepts in clinical texts.* American Medical Informatics Association, 2011. **18**(5): p. 594–600.
123. Abacha, A.B. and Zweigenbaum, P. *A hybrid approach for the extraction of semantic relations from MEDLINE abstracts*, in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Editor. 2011, Springer. p. 139–150.
124. Kim, J.-D., Ohta, T., and Tsujii, J.i. *Corpus annotation for mining biomedical events from literature.* BMC Bioinformatics, 2008. **9**(1): p. 10.
125. Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., and Tsujii, J.i. *Extracting bio-molecular events from literature—the bionlp'09 shared task.* Computational Intelligence, 2011. **27**(4): p. 513–540.
126. Kim, J.-D., Wang, Y., and Yasunori, Y. *The genia event extraction shared task, 2013 edition-overview*. In *Proceedings of the BioNLP Shared Task 2013 Workshop*. 2013.p. 8–15.
127. Pyysalo, S., Ohta, T., Rak, R., Sullivan, D., Mao, C., Wang, C., Sobral, B., Tsujii, J.i., and Ananiadou, S. *Overview of the infectious diseases (ID) task of*

BioNLP Shared Task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*. 2011. Association for Computational Linguistics.p. 26–35.

128. Pyysalo, S., Ohta, T., and Ananiadou, S. *Overview of the cancer genetics (CG) task of bionlp shared task 2013*. In *Proceedings of the BioNLP Shared Task 2013 Workshop*. 2013.p. 58–66.
129. Ohta, T., Pyysalo, S., Rak, R., Rowley, A., Chun, H.-W., Jung, S.-J., Jeong, C.-h., Choi, S.-p., and Ananiadou, S. *Overview of the pathway curation (PC) task of BioNLP shared task 2013.* 2013.
130. Miwa, M., Thompson, P., Korkontzelos, I., and Ananiadou, S. *Comparable Study of Event Extraction in Newswire and Biomedical Domains.* Computational Linguistics, 2014: p. 2270–2279.
131. Wattarujeekrit, T., Shah, P.K., and Collier, N. *PASBio: predicate-argument structures for event extraction in molecular biology.* BMC Bioinformatics, 2004. **5**(1): p. 155.
132. Ananiadou, S., Pyysalo, S., Tsujii, J.i., and Kell, D.B. *Event extraction for systems biology by text mining the literature.* Trends in Biotechnology, 2010. **28**(7): p. 381–390.
133. Ananiadou, S., Thompson, P., Nawaz, R., McNaught, J., and Kell, D.B. *Event-based text mining for biology and functional genomics.* Briefings in Functional Genomics, 2014: p. elu015.
134. Miwa, M., Sætre, R., Kim, J.-D., and Tsujii, J.i. *Event extraction with complex event classification using rich features.* Journal of Bioinformatics and Computational Biology, 2010. **8**(01): p. 131–146.
135. Dan Klein, C. and Manning, C.D. *Fast Exact Inference with a Factored Model for Natural Language Parsing*. In *Advances in Neural Information Processing Systems 15*. 2003. Cambridge, MA: MIT press.p. 3–10.
136. Fellbaum, C., ed. *WordNet: An electronic lexical database*. 1998, MIT press Cambridge, MA.
137. Riedel, S. and McCallum, A. *Robust biomedical event extraction with dual decomposition and minimal domain adaptation*. In *Proceedings of the BioNLP Shared Task 2011 Workshop*. 2011. Association for Computational Linguistics.p. 46–50.

138. McClosky, D., Riedel, S., Surdeanu, M., McCallum, A., and Manning, C.D. *Combining joint models for biomedical event extraction.* BMC Bioinformatics, 2012. **13**(Suppl 11): p. S9.
139. Riedel, S. and McCallum, A. *Fast and robust joint models for biomedical event extraction*. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2011. Association for Computational Linguistics.p. 1-12.
140. McClosky, D., Surdeanu, M., and Manning, C.D. *Event extraction as dependency parsing.* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 2011: p. 1626–1635.
141. Thompson, P., Batista-Navarro, R.T., Kontonatsios, G., Carter, J., Toon, E., McNaught, J., Timmermann, C., Worboys, M., and Ananiadou, S. *Text Mining the History of Medicine.* PLOS ONE, 2016. **11**(1): p. e0144717.
142. Wilbur, W.J., Rzhetsky, A., and Shatkay, H. *New directions in biomedical text annotation: definitions, guidelines and corpus construction.* BMC Bioinformatics, 2006. **7**(1): p. 356.
143. Denny, J.C. *Chapter 13: mining electronic health records in the genomics era.* PLOS Computational Biology, 2012. **8**(12): p. e1002823.
144. Wang, Y. *Annotating and recognising named entities in clinical notes*. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*. 2009. Association for Computational Linguistics.p. 18–26.
145. Tateisi, Y. and Tsujii, J.i. *Part-of-Speech annotation of biology research abstracts*. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 2004.
146. Tateisi, Y., Yakushiji, A., Ohta, T., and Tsujii, J.i. *Syntax annotation for the GENIA corpus*. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-2005), Jeju Island, Korea, October*. 2005.p. 11–13.
147. Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J.i. *GENIA corpus—a semantically annotated corpus for bio-textmining.* Bioinformatics, 2003. **19**(suppl 1): p. i180–i182.

148. Su, J., Yang, X., Hong, H., Tateisi, Y., and Tsujii, J.i. *Coreference resolution in biomedical texts: a machine learning approach*. In *Dagstuhl Seminar Proceedings*. 2008. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

149. Pyysalo, S., Ohta, T., Kim, J.-D., and Tsujii, J.i. *Static relations: a piece in the biomedical information extraction puzzle*. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. 2009. Association for Computational Linguistics.p. 1–9.

150. Ohta, T., Pyysalo, S., Kim, J.-D., and Tsujii, J.i. *A re-evaluation of biomedical named entity–term relations.* Bioinformatics and Computational Biology, 2010. **8**(05): p. 917–928.

151. Kim, J.-D., Ohta, T., and Tsujii, J.i. *Corpus annotation for mining biomedical events from literature*. BMC Bioinformatics, 2008. **9**(1): p. 1.

152. Vincze, V., Szarvas, G., Farkas, R., Móra, G., and Csirik, J. *The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes.* BMC Bioinformatics, 2008. **9**(Suppl 11): p. S9.

153. Kulick, S., Bies, A., Liberman, M., Mandel, M., McDonald, R., Palmer, M., Schein, A., Ungar, L., Winters, S., and White, P. *Integrated annotation for biomedical information extraction*. In *Proceedings of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*. 2004.p. 61–68.

154. Taylor, A., Marcus, M., and Santorini, B. *The Penn treebank: an overview*, in *Treebanks*, A. Abeillé Editor. 2003, Springer. p. 5–22.

155. Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. *Overview of BioCreAtIvE: critical assessment of information extraction for biology.* BMC Bioinformatics, 2005. **6**(Suppl 1): p. S1.

156. Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L., and Valencia, A. *Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge.* Genome Biology, 2008. **9**(Suppl 2): p. S1.

157. Nédellec, C. *Learning language in logic—genic Interaction extraction challenge*. In *Proceedings of the 4th Learning Language in Logic Workshop*.p. 1–7.

158. Doğan, R.I., Leaman, R., and Lu, *Z. NCBI disease corpus: a resource for disease name recognition and concept normalization.* Biomedical Informatics, 2014. **47**: p. 1–10.

159. Ananiadou, S., Kell, D.B., and Tsujii, J.-i. *Text mining and its potential applications in systems biology*. Trends in Biotechnology, 2006. **24**(12): p. 571–579.

160. Cohen, K.B., Verspoor, K., Fort, K., Funk, C., Bada, M., Palmer, M., and Hunter, L. *The Colorado Richly Annotated Full Text (CRAFT) corpus: Multi-model annotation in the biomedical domain*. Handbook of Linguistic Annotation, ed. N. Ide, Pustejovsky, James. 2015: Springer.

161. Wagholikar, K., Torii, M., Jonnalagadda, S., and Liu, H. *Pooling annotated corpora for clinical concept extraction.* Biomedical Semantics, 2013. **4**(1): p. 3.

162. Saeed, M., Villarroel, M., Reisner, A.T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T.H., Moody, B., and Mark, R.G. *Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database.* Critical Care Medicine, 2011. **39**(5): p. 952.

163. Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J., and South, B.R. *Evaluating the state of the art in coreference resolution for electronic medical records.* American Medical Informatics Association, 2012. **19**(5): p. 786–791.

164. Uzuner, Ö., Solti, I., Xia, F., and Cadag, E. *Community annotation experiment for ground truth generation for the i2b2 medication challenge.* American Medical Informatics Association, 2010. **17**(5): p. 519–523.

165. Frunza, O. and Inkpen, D. *Identifying and classifying semantic relations between medical concepts in clinical data (I2b2 Challenge)*. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. 2010.p. 98.

166. Savova, G.K., Chapman, W.W., Zheng, J., and Crowley, R.S. *Anaphoric relations in the clinical narrative: corpus creation.* American Medical Informatics Association, 2011. **18**(4): p. 459–465.

167. Goeuriot, L., Jones, G., Kelly, L., Leveling, J., Hanbury, A., Müller, H., Salanterä, S., Suominen, H., and Zuccon, G. *ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when*

reading clinical reports, in *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, P. Forner, et al., Editors. 2013.

168. Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Setzer, A., and Roberts, I. *Semantic annotation of clinical text: The CLEF corpus*. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC).* 2008.p. 19–26.

169. Ogren, P.V., Savova, G., Buntrock, J.D., and Chute, C.G. *Building and evaluating annotated corpora for medical NLP systems*. In *Proceedings of the American Medical Informatics Association*. 2006.p. 1050.

170. Mihăilă, C., Ohta, T., Pyysalo, S., and Ananiadou, S. *BioCause: Annotating and analysing causality in the biomedical domain.* BMC Bioinformatics, 2013. **14**(1): p. 2.

171. Alnazzawi, N., Thompson, P., and Ananiadou, S. *Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature*. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL*. 2014.p. 69–74.

172. Fu, X., Batista-Navarro, R., Rak, R., and Ananiadou, S. *A strategy for annotating clinical records with phenotypic information relating to the chronic obstructive pulmonary disease*. In *Proceedings of the 22nd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2014)*. 2014.p. 1–8.

173. Fu, X., Batista-Navarro, R., Rak, R., and Ananiadou, S. *Supporting the annotation of chronic obstructive pulmonary disease (COPD) phenotypes with text mining workflows.* Biomedical Semantics, 2015. **6**(1): p. 1.

174. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.-W.W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W.A. *Disease Ontology: a Backbone for Disease Semantic Integration.* Nucleic Acids Research, 2012. **40**(D1): p. D940–D946.

175. Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E., and Haendel, M.A. *Uberon, an integrative multi-species anatomy ontology.* Genome Biology, 2012. **13**(1): p. R5.

176. Batista-Navarro, R., Carter, J., and Ananiadou, S. *Semi-automatic curation of chronic obstructive pulmonary disease phenotypes using Argo*. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*. 2015.p. 403–408.

177. Xia, F. and Yetisgen-Yildiz, M. *Clinical corpus annotation: challenges and strategies*. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*. 2012.

178. Yetisgen-Yildiz, M., Solti, I., Xia, F., and Halgrim, S.R. *Preliminary experience with Amazon's Mechanical Turk for annotating medical named entities*. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 2010. Association for Computational Linguistics.p. 180–183.

179. Ogren, P.V. *Knowtator: a Protégé plug-in for annotated corpus construction*. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations*. 2006. Association for Computational Linguistics.p. 273-275.

180. Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., and Gorrell, G. *GATE Teamware: a web-based, collaborative text annotation framework*. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 2013.p. 1007–1029.

181. Day, D.S., McHenry, C., Kozierok, R., and Riek, L. *Callisto: A configurable annotation workbench*. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 2004.

182. Petasis, G., Karkaletsis, V., Paliouras, G., and Spyropoulos, C.D. *Using the ellogon natural language engineering infrastructure*. In *Proceedings of the Workshop on Balkan Language Resurces and Tools, 1st Balkan Conference in Informatics (BCI 2003)*. 2003.

183. Morton, T. and LaCivita, J. *WordFreak: an open tool for linguistic annotation*. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations-Volume 4*. 2003. Association for Computational Linguistics.p. 17–18.

184. South, B.R., Shen, S., Leng, J., Forbush, T.B., DuVall, S.L., and Chapman, W.W. *A prototype tool set to support machine-assisted annotation*. In

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. 2012. Association for Computational Linguistics.p. 130–139.

185. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J.i. *BRAT: a web-based tool for NLP-assisted text annotation*. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012. Association for Computational Linguistics.p. 102–107.
186. Thompson, P., McNaught, J., Montemagni S., Calzolari, N., del Gratta, R., Lee, V., Marchi, S., Monachini, M., Pezik, P., Quochi, V., Rupp, C.J., Sasaki, Y., Venturi, G., Rebholz-Schuhmann, D., and Ananiadou, S. *The BioLexicon: a large-scale terminological resource for biomedical text mining.* BMC Bioinformatics, 2011. **12**: p. 397.
187. Hahn, U., Romacker, M., and Schulz, S. *How knowledge drives understanding—matching medical ontologies with the needs of medical language processing.* Artificial Intelligence in Medicine, 1999. **15**(1): p. 25–51.
188. National Library of Medicine. *SPECIALIST Lexicon and Lexical Tools.*, in *Chapter 6 of UMLS® Reference Manual*. 2009, Bethesda, MD : U.S. National Library of Medicine.
189. Patrick, J., Wang, Y., and Budd, P. *An automated system for conversion of clinical notes into SNOMED clinical terminology*. In *Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68*. 2007. Australian Computer Society, Inc.p. 219–226.
190. Mooney, R.J. and Bunescu, R.C. *Subsequence kernels for relation extraction*. In *Advances in neural information processing systems*. 2005.p. 171–178.
191. Giuliano, C., Lavelli, A., and Romano, L. *Simple information extraction (SIE): A portable and effective IE system*. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*. 2006.p. 9–16.
192. Biber, D. and Gray, B. *Challenging stereotypes about academic writing: Complexity, elaboration, explicitness.* Journal of English for Academic Purposes, 2010. **9**(1): p. 2–20.
193. Harris, Z. *A Grammar of English on Mathematical Principles*. 1983: Wiley.

194. Friedman, C., Kra, P., and Rzhetsky, A. *Two biomedical sublanguages: a description based on the theories of Zellig Harris.* Biomedical Informatics, 2002. **35**(4): p. 222–235.

195. Dominiczak, A.F., Negrin, D.C., Clark, J.S., Brosnan, M.J., McBride, M.W., and Alexander, M.Y. *Genes and hypertension from gene mapping in experimental models to vascular gene transfer strategies.* Hypertension, 2000. **35**(1): p. 164–172.

196. Mihăilă, C., Batista-Navarro, R., Alnazzawi, N., Kontonatsios, G., Korkontzelos, I., Rak, R., Thompson, P., and Ananiadou, S. *Mining the Biomedical Literature,* in *Healthcare Data analytics* C.K. Reddy and C.C. Aggarwal, Editors. 2015, CRC Press.

197. Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W.A., Cohen, K.B., Verspoor, K., and Blake, J.A. *Concept annotation in the CRAFT corpus.* BMC Bioinformatics, 2012. **13**(1): p. 161.

198. Thompson, P., Iqbal, S., McNaught, J., and Ananiadou, S. *Construction of an annotated corpus to support biomedical information extraction.* BMC Bioinformatics, 2009. **10**(1): p. 349.

199. Prasad, R., McRoy, S., Frid, N., Joshi, A., and Yu, H. *The biomedical discourse relation bank.* BMC Bioinformatics, 2011. **12**(1): p. 188.

200. South, B.R., Shen, S., Jones, M., Garvin, J., Samore, M.H., Chapman, W.W., and Gundlapalli, A.V. *Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease.* BMC Bioinformatics, 2009. **10**(Suppl 9): p. S12.

201. Ogren, P.V., Savova, G.K., and Chute, C.G. *Constructing evaluation corpora for automated clinical named entity recognition*. In *Proceedings of the 12th World Congress on Health (Medical) Informatics*. 2007.p. 2325.

202. Artstein, R. and Poesio, M. *Inter-coder agreement for computational linguistics.* Computational Linguistics, 2008. **34**(4): p. 555–596.

203. Hripcsak, G. and Heitjan, D.F. *Measuring agreement in medical informatics reliability studies.* Biomedical Informatics, 2002. **35**(2): p. 99–110.

204. Cohen, J. *A coefficient of agreement for nominal scales.* Educational and Psychological Measurement, 1960. **20**(1): p. 37–46.

205. Hripcsak, G. and Rothschild, A.S. *Agreement, the F-measure, and reliability in information retrieval.* American Medical Informatics Association, 2005. **12**(3): p. 296–298.

206. Alnazzawi, N., Thompson, P., Batista-Navarro, R., and Ananiadou, S. *Using text mining techniques to extract phenotypic information from the PhenoCHF corpus.* BMC Medical Informatics and Decision Making, 2015. **15**(Suppl 2): p. S3.

207. Zeng, Q., Goryachev, S., Weiss, S., Sordo, M., Murphy, S., and Lazarus, R. *Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system.* BMC Medical Informatics and Decision Making, 2006. **6**(1): p. 30.

208. Khordad, M., Mercer, R.E., and Rogan, P. *Improving phenotype name recognition*, in *Advances in Artificial Intelligence*, C. Butz and P. Lingras, Editors. 2011, Springer. p. 246–257.

209. Rabiner, L.R. and Juang, B.-H. *An introduction to Hidden Markov Models.* IEEE ASSP Magazine, 1986. **3**(1): p. 4–16.

210. McCallum, A., Freitag, D., and Pereira, F.C. *Maximum Entropy Markov Models for Information Extraction and Segmentation*. In *International Conference on Machine Learning*. 2000.p. 591–598.

211. Lafferty, J., McCallum, A., and Pereira, F.C. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In *Proceedings of the Eighteenth International Conference on Machine Learning*. 2001.p. 282–289.

212. Okazaki, N. *CRFsuite: a fast implementation of Conditional Random Fields (CRFs).* http://www.chokkan.org/software/crfsuite. 2007. Accessed April 2016.

213. McCallum, A.K. *MALLET: A Machine Learning for Language Toolkit.* http://mallet.cs.umass.edu. 2002. Accessed April 2016.

214. Ponomareva, N., Rosso, P., Pla, F., and Molina, A. *Conditional Random Fields vs. Hidden Markov Models in a biomedical named entity recognition task*. In *Proceedings of Int. Conf. Recent Advances in Natural Language Processing, RANLP*. 2007.p. 479–483.

215. Zhou, G. and Su, J. *Named entity recognition using an HMM-based chunk tagger*. In *Proceedings of the 40th Annual Meeting on Association for*

Computational Linguistics. 2002. Association for Computational Linguistics.p. 473–480.

216. LingPipe. *http://alias-i.com/lingpipe. Accessed April 2016.* 4.1.0.

217. Zhang, J., Shen, D., Zhou, G., Su, J., and Tan, C.-L. *Enhancing HMM-based biomedical named entity recognition by studying special phenomena.* Journal of Biomedical Informatics, 2004. **37**(6): p. 411–422.

218. Patrick, J., Wang, Y., and Budd, P. *Automatic mapping clinical notes to medical terminologies*. In *Proc. Of the 2006 Australian Language Technology Workshop*. 2006.p. 75-82.

219. Wagholikar, K.B., Torii, M., Jonnalagadda, S., and Liu, H. *Pooling annotated corpora for clinical concept extraction.* Biomedical Semantics, 2013. **4**: p. 3.

220. Patrick, J. and Li, M. *High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge.* American Medical Informatics Association, 2010. **17**(5): p. 524–527.

221. Zhou, D., Zhong, D., and He, Y. *Biomedical Relation Extraction: From Binary to Complex.* Computational and Mathematical Methods in Medicine, 2014. **2014**.

222. McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y., and White, P. *Simple algorithms for complex relation extraction with applications to biomedical IE*. In *Proceedings of Association for Computational Linguistics*. 2005.p. 491–498.

223. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. *The WEKA data mining software: an update.* ACM SIGKDD Explorations Newsletter, 2009. **11**(1): p. 10–18.

224. Miyao, Y. and Tsujii, J.i. *Feature forest models for probabilistic HPSG parsing.* Computational Linguistics, 2008. **34**(1): p. 35–80.

225. Bunescu, R.C. and Mooney, R.J. *A shortest path dependency kernel for relation extraction*. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 2005. Association for Computational Linguistics.p. 724–731.

226. Sagae, K. and Tsujii, J.i. *Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles*. In *Proceedings of Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2007.p. 1044–1050.

227. Hirschman, L., Colosimo, M., Morgan, A., and Yeh, A. *Overview of BioCreAtIvE task 1B: normalized gene lists.* BMC Bioinformatics, 2005. **6**(Suppl 1): p. S11.

228. Lu, Z. and Wilbur, W.J. *Overview of BioCreative III gene normalization*. In *Proceedings of the BioCreative III Workshop*. 2010. Citeseer.p. 24–45.

229. Lu, Z., Kao, H.-Y., Wei, C.-H., Huang, M., Liu, J., Kuo, C.-J., Hsu, C.-N., Tsai, R.T., Dai, H.-J., and Okazaki, N. *The gene normalization task in BioCreative III.* BMC Bioinformatics, 2011. **12**(Suppl 8): p. S2.

230. Dogan, R.I., Murray, G.C., Névéol, A., and Lu, Z. *Understanding PubMed® user search behavior through log analysis.* Database, 2009. **2009**: p. bap018.

231. Kang, N., Singh, B., Afzal, Z., van Mulligen, E.M., and Kors, J.A. *Using rule-based natural language processing to improve disease normalization in biomedical text.* American Medical Informatics Association, 2013. **20**(5): p. 876–881.

232. Névéol, A., Kim, W., Wilbur, W.J., and Lu, Z. *Exploring two biomedical text genres for disease recognition*. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. 2009. Association for Computational Linguistics.p. 144–152.

233. Bos, L. and Donnelly, K. *SNOMED-CT: The advanced terminology and coding system for eHealth.* Studies in Health Technology and Informatics, 2006. **121**: p. 279–290.

234. Collier, N., Oellrich, A., Groza, T., Verspoor, K., and Shah, N. *Phenotype Day 2015*. In *Proceedings of the 23rd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2015)*. 2015.

235. Collier, N., Oellrich, A., Groza, T., Verspoor, K., and Shah, N. *Phenotype Day 2014*. In *Proceedings of the 22nd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2014)*. 2014.

236. Vasant, D., Neff, F., Gormanns, P., Conte, N., Fritsche, A., Staiger, H., Kim, J.-H., Malone, J., Raess, M., and de Angelis, M.H. *DIAB: An Ontology of Type 2 Diabetes Stages and Associated Phenotypes*. In *Proceedings of the 23rd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2015)*. 2015.p. 24–27.

237. Vasant, D., Chanas, L., Malone, J., Hanauer, M., Olry, A., and Jupp, S. *ORDO: An Ontology Connecting Rare Disease*. In *Proceedings of the 22nd*

Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2014). 2014.p. 49–52.

238. Wang, l., Bray, B.E., Shi, J., and Haug, P. *Can we acquire a complete heart-failure vocabulary from textual knowledge sources for building reference disease ontology?* . Proceedings of the 22nd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2014), 2014: p. 67.
239. Wang, L., Bray, B.E., Shi, J., Del Fiol, G., and Haug, P.J. *A method for the development of disease-specific reference standards vocabularies from textual biomedical literature resources.* Artificial Intelligence in Medicine, 2016: p. 47–57.
240. Winnenburg, R. and Bodenreider, O. *Coverage of phenotypes in standard terminologies*. In *Proceedings of the 22nd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2014)*. 2014.p. 41–44.
241. Hettne, K., Kaliyaperumal, R., van der Horst, E., Thompson, M., Hoen, P.t., and Roos, M. *Genotype-phenotype knowledge discovery using the Concept Profile Analysis Web Services*. In *Proceedings of the 23d Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2015).* 2015.p. 36.
242. Bello, S. and Eppig, J. *Inferring Gene-to-Phenotype and Gene-to-Disease Relationships: Challenges and Solutions*. In *Proceedings of the 23rd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2015)*. 2015.p. 16–19.
243. Orly, A., Prévot, C., and Jaramillo, C. *Indexation of rare diseases with HPO terms: A new Orphanet service to refine phenotype genotype correlations*. In *Proceedings of the 23rd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2015)*. 2015.p. 35.
244. Leaman, R., Khare, R., and Lu, Z. *Challenges in clinical natural language processing for automated disorder normalization.* Journal of Biomedical Informatics, 2015. **57**: p. 28-37.
245. Hersh, W.R. and Hickam, D.H. *Information retrieval in medicine: the SAPHIRE experience.* Association for Information Science and Technology, 1995. **46**(10): p. 743–747.

246. Oellrich, A., Collier, N., Smedley, D., and Groza, T. *Generation of silver standard concept annotations from biomedical texts with special relevance to phenotypes.* PLOS ONE, 2015. **10**(1): p. e0116040.

247. Jonquet, C., Shah, N., and Musen, M. *The open biomedical annotator*. In *AMIA Summit on Translational Bioinformatics*. 2009.p. 56–60.

248. Nunes, T., Campos, D., Matos, S., and Oliveira, J.L. *BeCAS: biomedical concept recognition services and visualization.* Bioinformatics, 2013: p. btt317.

249. Levenshtein, V.I. *Binary codes capable of correcting deletions, insertions, and reversals.* Soviet Physics Doklady, 1966. **10**: p. 707–710.

250. Jaro, M.A. *Probabilistic linkage of large public health data files.* Statistics in Medicine, 1995. **14**(5-7): p. 491–498.

251. Winkler, W.E. *The state of record linkage and current research problems*. *Statistical Research Division, SDR of the US Census Bureau*. 1999.

252. Das, J. and Choong, P.L. *Resolving partial name mentions using string metrics*. 2007, Defence Science and Technology Organisation, Edinburgh (Australia), Command, Control, Communications, and Intelligence Division. p. 40.

253. Cohen, W., Ravikumar, P., and Fienberg, S. *A comparison of string metrics for matching names and records*. In *KDD workshop on Data Cleaning and Object Consolidation*. 2003.p. 73–78.

254. Jaccard, P. *The distribution of the flora in the alpine zone. 1*. New Phytologist, 1912. **11**(2): p. 37–50.

255. Cohen, W.W., Ravikumar, P., and Fienberg, S.E. *A Comparison of String Distance Metrics for Name-Matching Tasks.* KDD workshop on Data Cleaning and Object Consolidation, 2003.

256. Camacho, D., Huerta, R., and Elkan, C. *An evolutionary hybrid distance for duplicate string matching*. 2008, Technical report, Universidad Autonoma de Madrid.

257. Moreau, E., Yvon, F., and Cappé, O. *Robust similarity measures for named entities matching*. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. 2008. Association for Computational Linguistics.p. 593–600.

258. Cheatham, M. and Hitzler, P. *The role of string similarity metrics in ontology alignment.* 2013.

259. Baumgartner Jr, W.A., Lu, Z., Johnson, H.L., Caporaso, J.G., Paquette, J., Lindemann, A., White, E.K., Medvedeva, O., Cohen, K.B., and Hunter, L. *Concept recognition for extracting protein interaction relations from biomedical text.* Genome Biology, 2008. **9**(Suppl 2): p. S9.

260. Cohen, W., Ravikumar, P., and Fienberg, S. *Secondstring: An open source java toolkit of approximate string-matching techniques.* Project web page: http://secondstring.sourceforge.net, 2003.

261. Davis, A.P., Thomas C, W., Rosenstein, M.C., and Mattingly, C.J. *MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database.* Database, 2012. **2012**: p. bar065.

262. Leaman, R., Islamaj Doğan, R., and Zhiyong, L. *DNorm: disease name normalisation with pairwise learning to rank.* Bioinformatics, 2013: p. btt474.

263. Islamaj Doğan, R. and Lu, Z. *An inference method for disease name normalization*. In *2012 AAAI Fall Symposium Series*. 2012.

264. Myers, E.W. *An O (ND) difference algorithm and its variations.* Algorithmica, 1986. **1**(1-4): p. 251–266.

265. Paul, M.J., Sarker, A., Brownstein, J.S., Nikfarjam, A., Scotch, M., Smith, K.L., and Gonzalez, G. *Social media mining for public health monitoring and surveillance*. In *Pacific Symposium on Biocomputing*. 2016.p. 468.

266. Denecke, K. and Nejdl, W. *How valuable is medical social media data? Content analysis of the medical web.* Information Sciences, 2009. **179**(12): p. 1870–1880.

267. Pearson, J.F., Brownstein, C.A., and Brownstein, J.S. *Potential for electronic health records and online social networking to redefine medical research.* Clinical chemistry, 2011. **57**(2): p. 196–204.

Appendix A

A. PhenoCHF corpus

A.1 Annotation guidelines

Overview:

More than one third of patients with chronic kidney disease (CKD) develop symptoms of heart failure. Congestive heart failure (CHF) is also a common contributor to the progression of CKD. Thus, a vicious circle exists between these two diseases. Therefore, renal failure or renal insufficiency may be more than a marker for heart failure severity and instead may play a causative role in the progression of heart failure.

Annotation is marking up piece of text with class to describe it. In this task we will use BRAT in which it displays the annotations by highlighting the text with different colours according to the chosen class.

The Data:

The corpus consists of 300 clinical records obtained from the i2b2 obesity challenge. The medical records are filtered to include only the discharge summaries for the patients with CHF and kidney failure (renal insufficiency). The second part of the corpus consists of the 10 most recent (at corpus collection time) full-text articles retrieved from the PubMed Central Open Access database, using the query "Heart failure" OR "Congestive Heart Failure" OR "Heart failure Clinical presentation" OR "Heart failure Clinical features" OR "Heart failure Symptoms" OR "Heart failure clinical manifestation" OR "Heart failure clinical picture" AND ("Chronic renal failure" OR "Renal failure" OR "Chronic Renal insufficiency" OR "Renal insufficiency" OR "Kidney failure" OR "CRF" OR "CRI").

Aim of current annotation :

This annotation task consists of two stages: i) to mark-up the mentions of medical terms that denote phenotypic information of CHF to highlight the entire characteristics of CHF such a: causes, risk factors , clinical signs/symptom and also to

identify non-traditional risk factors (uremia-related) of heart failure to investigate the extent to which renal failure can worsen the condition of CHF, ii) the relationships annotations link two annotated entities to reflects the interactions between them.

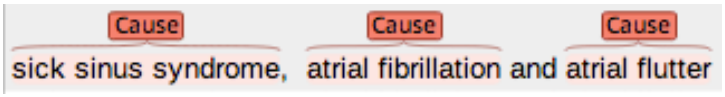
Annotation task:

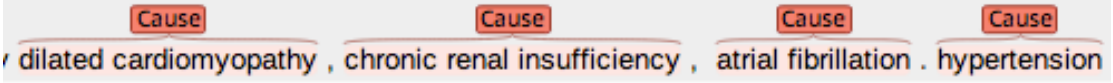
1- Terms annotation (Entities annotation):

The basic task of annotation is entity in which stretches of text (medical terms) that denote phenotypic information is marked up with the most relevant class.

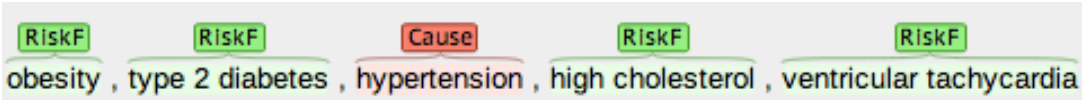
CHF Phenotypic information is defined in three general classes as follows:

1.1) **Cause** is any medical problem that contributes to the occurrence of CHF. It could be any disease (i.e. coronary artery disease, renal insufficiency) or disordered activity of body systems, organ or tissues (i.e. atrial fibrillation) that cause heart failure.

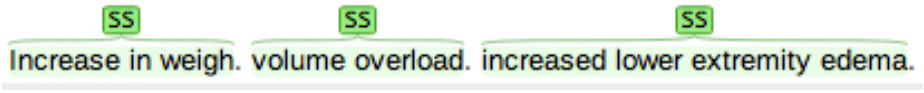




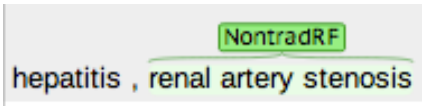
1.2) **Risk factor** (RiskF) is a medical or social condition that increases the risk of CHF disease or it may contribute to put the patient in higher risk of developing the causes of CHF such as: bad life style (i.e. physical inactivity , smoking , being overweight) or it could be drug to control other diseases i.e. amikacin. Also, any family history of cardiovascular disease.

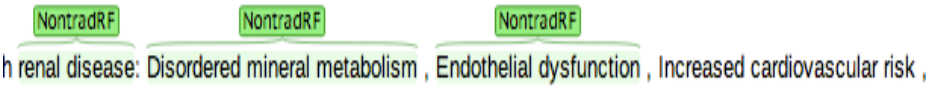


1.3) Sign or Symptom is any observable manifestation of CHF disease which either experienced by a patient and reported to the doctor or found as a result by the doctor's examination. For example, fatigue , decreased exercise tolerance, shortness of breath, sweating, irregular rhythm , murmur , rub and gallop sounds, low cardiac output.



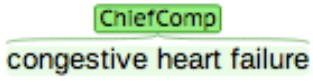
1.4) Non-traditional risk factor: (NontradRF) is the medical term that denotes the complication associated with abnormalities in the kidney functions that put the patient in a higher risk to progress "signs/symptoms" and causes of CHF, it could be disease (i.e. anemia) , electrolyte imbalance (i.e. hyperkalemia , hypokalemia and increased creatinine).



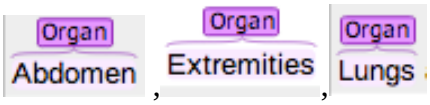


Other classes that highlight important and relevant information to the task:

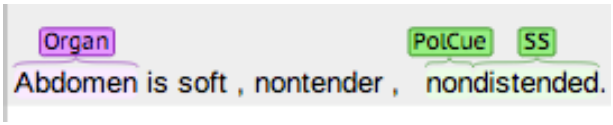
a)Chief complaint(Chiefcomp) is congestive heart failure in this annotation task.



b)Organ is any body part. For example,



c)Polarity clue (polcue) to highlight the negation modifier that negate medical condition such as no , without , denies etc. Polarity clue refers to any words that denote negation or absence of medical conditions meaning that the patient does not have this condition. Only annotate negation when it is related to medical condition.



Only annotate negation modifier when it is negated medical condition related to CHF .

General guidelines for term annotations:

1-Only annotate the correct span (as much information as required).

2- The phenotypic term could be expressed in any syntactic structure it can be:

- a noun phrase
  - *pedal edema*
  - *diastolic dysfunction*
  - *orthopnea*
- prepositional phrase.
  - *Pain in chest*
  - *Shortness of breath*
- adjective phrases

- The patient was *hypertensive.*
- The patent was *anemic*
- The patient becomes *hypercalemic.*

3- If the medical condition is preceded by modifier of multi words phrase. Annotate the whole phrase (except for negation refer to polarity clue annotation) for example:

- *Increased potassium.*
- *increased shortness of breath*
- *left atrial enlargement*

4- There should be only one annotation per mentioned disorder.

6- Annotate all abbreviations and acronyms that refer to phenotypic information. For example:

- JVP refers to jugular venous pressure
- CABG stand for coronary artery bypass graft
- *AF stands for Atrial fibrillation*

- *A-fib stands for Atrial fibrillation*

Do not annotate the following:

1- Normal condition for example if the information describe normal function i.e. EKG showed normal sinus rhythm , regular chest etc.
2- anything you infer from the text only annotate the explicitly mentioned entity. For example consider the following lab result:

Laboratory data: INR of 1.6, BUN of 110, creatinine 3, potassium 5.5, white blood cell count of 11.7 and a hematocrit of 27.9.

Do not annotate anything in the previous example even though you felt that, the patient have high level of creatinine which suggest that the patient have impaired kidney function or the patient have hyperkalemia which is uremia related risk factor for heart failure.

Another example is: Echo showed ejection fraction of 10-15%

Do not annotate anything in the previous sentence, even though an ejection fraction of 10-15% indicates sign of CHF.

2-Relations Annotation:

The aim of this task is to annotate the existing relationships on the top of the annotated entities. It usually links two annotated concepts (arguments) within the boundaries of a single sentence. Relationships help to identify:

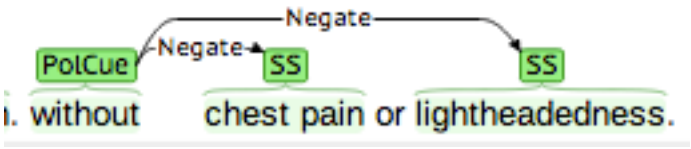
- Which medical condition causes the other?
- Which negation modifies which sign or symptoms?

There are three types of relationships and each type constrains to link two specific and predefined pair of arguments.

2.1) Negate

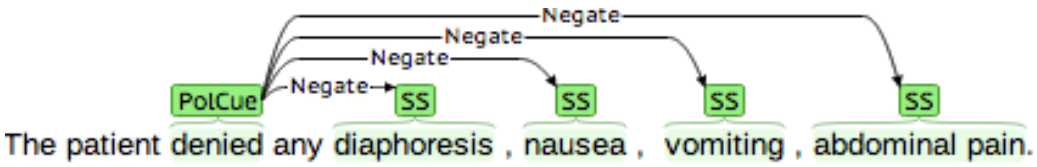
It is one-way relation to relate negation attribute to the condition that it is denied.

For example:



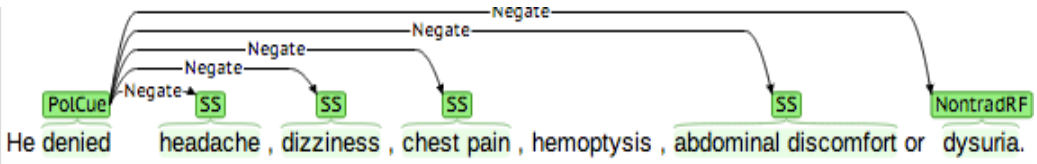
The above mentioned example illustrates the negate relationship as follows the modifier *without* is used to explicitly negate the following phenotypic information *chest pain* and *lightheadness.*

- If the negation modifier refer to several medical condition create different negate relation for each negated medical condition.



The above mentioned example illustrates the negate relationship as follows the modifier *denied* is used to explicitly negate the following phenotype information *diaphoresis, nausea*, *vomiting* and *abdominal pain*.

Another example is:



-Only annotate Negate relationship on the top of the annotated concepts and, DO NOT annotate *negate* relationship for concepts out the scoop of this annotation task for example:

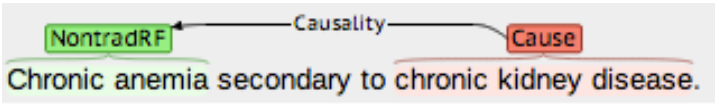
- The patient does not have diarrhea or constipation.
- No visual change.
- Patient was found without mental changes.

In the above mentioned examples *negate* relationship is not annotated because the underlined medical conditions are not associated with CHF.

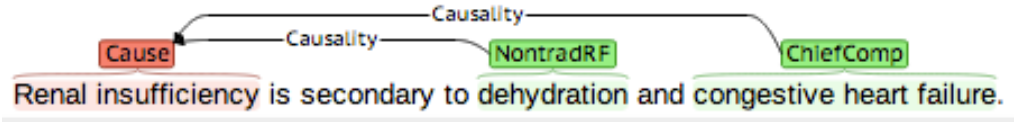
2.2)Causality:

This relationship links two concepts in which one argument causes the other.

For example, in the following sentence the *chronic kidney disease* causes the *chronic anemia*.

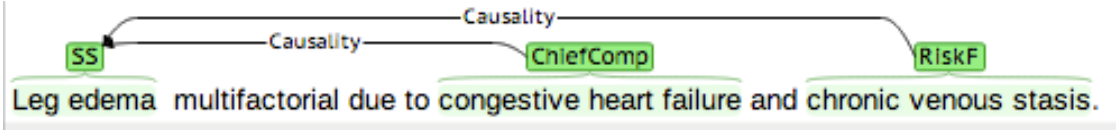


Another example,



In the above mentioned example there are two *causality* relations:

- First *causality* relationship associating *dehydration* to cause *renal insufficiency*.
- Second *causality* relationship associating *congestive heart failure* to cause *renal insufficiency*.



In the above mentioned example there are two causality relations:

- First *causality* relationship associating *congestive heart failure* to cause *leg*

edema.

- Second *causality* relationship associating *chronic venous stasis* to cause *leg edema*.

➢ Do not annotate the relationship causality whenever the relationship does not contribute to the progression of CHF. For example,

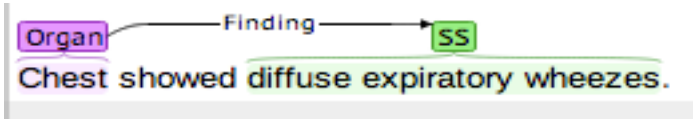
She had a urinary tract infection per report secondary to E.Coli resistant to Levaquin and gentamicin.

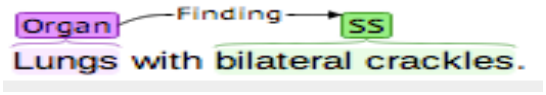
In the above mentioned example neither the relationship nor the medical concepts are annotated because they are irrelevant to this annotation task.

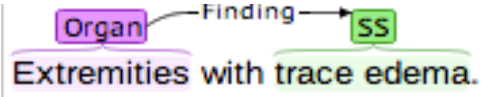
2.3) Finding

This relationship links the organ to the manifestation or abnormal variation that is observed during the diagnoses process.

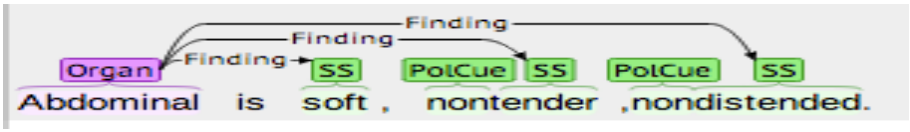
For example,

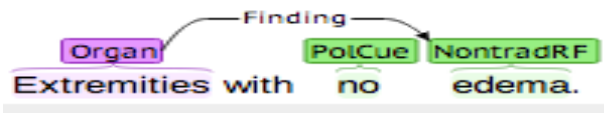






➢ Annotate the finding relationship even if the signs or symptoms of CHF are negated in the records, and annotate the negation relationship.



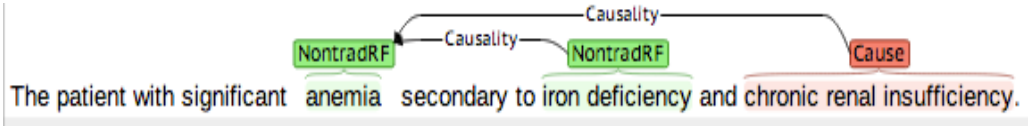


General guidelines for relationships annotations:

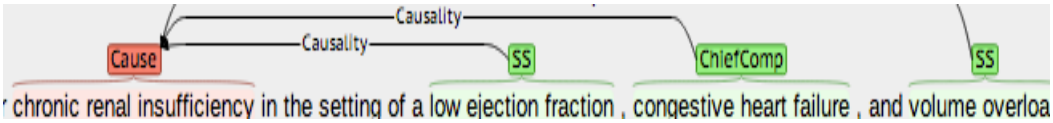
1- Relationship may be created based on medical knowledge:

2-If the relationship has many arguments create different causality relation for each negated medical condition.

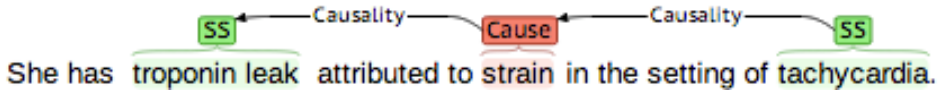
for example:



- First *causality* relationship associating *iron deficiency* to cause *anemia*.
- Second *causality* relationship associating *chronic renal insufficiency* to cause *anemia*.



3-Some argument causing other argument



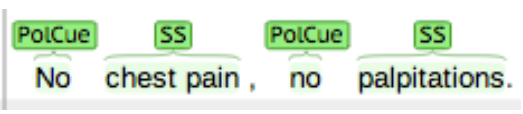
4- At least one relation exits for each annotated negation modifier

General Guidelines for the annotation task:

1- Complete this task independently. Do not discuss your annotations with anyone else.

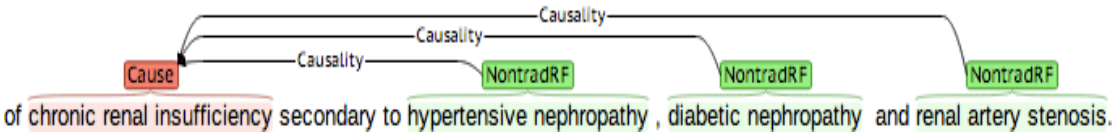
2- Read the whole patient's record first to get understanding about the patient medical case.

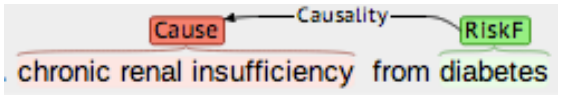
3- Read the document again and annotate the medical terms with correct class and in parallel annotate the negation modifier (polarity clue) where they are found.



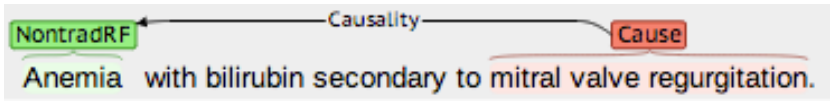
4- Go to each annotated terms and look if it is related to any other annotated term in the same sentence. For example consider the following :

- For each Cause decide whether it was caused by any non-traditional (uremia) risk factor or risk factor.

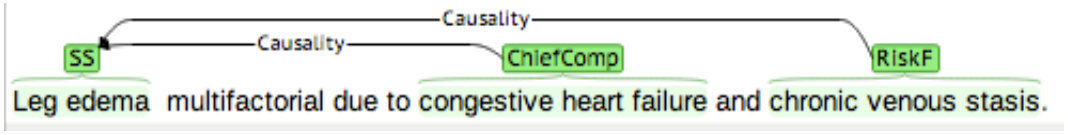


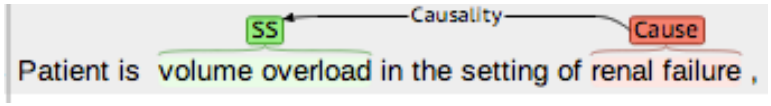


- For each non-traditional risk factor decide whether it was caused by any cause.

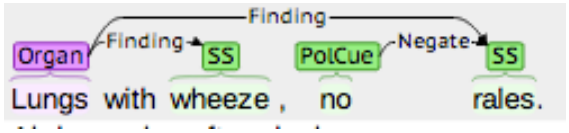


- For each sign or symptom decide whether it was caused by any risk factor or cause.





- For each organ check if it is associated with any sign or symptom.



5- Per each record you have annotated please record any comments you might think it is important to improve the guidelines for example you could record comments about:
   - The clarity and applicability of the guidelines.
   - Adding any important information that is not covered by the guidelines.

Appendix B

B. Phenotypic resources

B.1 List of phenotypic affixes

	Size 2	Size 3	Size 4	Size 5
Prefix	Ju,hy	Lig,Ble,deh,ede ,Cig,kes,chr ,mit,Bib,Dia ,jug,hyp	ligh,righ ,blee,dehy ,crac,whee ,atria,myoc ,coro,chro ,mitr,arrh ,pleu,biba ,jugu,dece, tric,hype	dyspn,Light, hypon, crack,wheez, ,myoca,globa ,coron,jugul ,chron,mitra ,palpi,aorti diast, arrhy ,pleur,biba ,hypoc,hyper
Suffix	Ep	eep,dia,ism	dism,lure ,tite,rdia ,sure,tter ,lism,emia,mnia,okes,sion	jugul,perip ,nsion,emia ,hemia,lemia ,nuria,rokes ,cytic, ,ilure,opnea ,spnea,litus ,etite,ardia ,ssure,utter ,olism,temia, iency ,ality