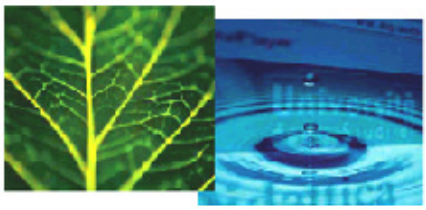
PhD Dissertation



International Doctorate School in Information and Communication Technologies

DIT - University of Trento

Improving the Effectiveness of Information Extraction from Biomedical Text

Md. Faisal Mahbub Chowdhury

Advisor:

Alberto Lavelli

Senior Researcher

HLT research unit, FBK-irst, Italy

April 2013

Members of the Examination Committee:

Prof. Eric Gaussier
Université Joseph Fourier - Laboratoire d'informatique de Grenoble (LIG)

Prof. Massimo Poesio
University of Essex - School of Computing and Electronic Engineering
University of Trento - Centro interdipartimentale Mente/Cervello (CIMEC)

Prof. Pierre Zweigenbaum
Université de Paris-Sud - LIMSI-CNRS, ILES Group
INALCO - Centre de Recherche en Ingénierie Multilingue (CRIM)

Abstract

Information extraction (IE) is the task which aims at automatically extracting specific target information from texts by means of various natural language processing (NLP) and Machine Learning (ML) techniques. The huge amount of available biomedical and clinical texts is an important source of undiscovered knowledge and an interesting domain where IE techniques can be applied. Although there has been a considerable amount of work for IE on other genres of text (such as newspaper articles), results of the state-of-the-art approaches for some of the IE tasks show there is still the need of improvement. Moreover, when these IE approaches are directly applied on biomedical/clinical data, the performance drops considerably. Customization of the IE approaches with biomedical/clinical genre specific features and pre/post-processing techniques does improve the results (with respect to applying the approaches directly) but the situation is still not completely satisfactory. There are many ways to accomplish this goal (e.g. exploitation of scope of negations, discourse structure, semantic roles, etc) which are yet to be fully harnessed for the improvement of IE systems. Additional challenges come from the usage of machine learning (ML) techniques themselves. Imbalance in data distribution is quite common in many NLP (including IE) tasks. Previous studies have empirically shown that unbalanced datasets lead to poor performance for the minority class.

In this PhD research, we aim to address the open issues outlined above. We focus on three core IE tasks which are crucial for text mining: named entity recognition (NER), coreference resolution (CoRef), and relation extraction (RE).

For NER, we propose an approach for the recognition of disease entity mentions which achieves state-of-the-art performance and is later exploited as a component in our RE system. Our NER system achieves results on par with the state of the art also for other bio-entity types such as genes/proteins, species and drugs. Since the creation of manually annotated training data is a costly process, we also investigate the practical usability of automatically annotated corpora for NER and propose how to automatically improve the quality of such corpora.

CoRef, which is naturally the next step after NER, is often deemed as one of the stumbling blocs for other IE tasks such as RE. We propose a greedy and constrained CoRef approach that achieves high results in clinical texts for each individual entity mention type and for each of the four different evaluation metrics usually computed for assessing systems' performance.

As for RE, one of the fundamental characteristics of our approach is that we propose to exploit other NLP areas such as scope of negations, elementary discourse units and semantic roles. We propose a novel hybrid kernel that not only takes advantage of different types of information (syntactic, semantic, contextual, etc) but also of the different ways they can be represented (i.e. flat structure, tree, graph). Our approach yields significantly better results than the previous state-of-the-art approaches for drug-drug interaction and protein-protein interaction extraction tasks.

In each of the above tasks, we concentrate to develop pro-active IE approaches to automatically get rid of unnecessary training/test instances even before training ML models and using those models on test data. This enables better performance because of the reduction of less skewed data distribution as well as faster runtime.

We tested our NER and RE approaches on other genres of text such as newspaper articles and automatically transcribed broadcast news. The

results show that our approaches are largely domain independent.

Keywords

Information extraction, biomedical text mining, named entity recognition, coreference resolution, relation extraction.

Acknowledgements

Time flies so fast. Exactly 3.5 years ago when I started my PhD, I had no idea what a mysterious journey was waiting for me. Within the first few months I became completely confused. I wanted to do something cool but I did not know what exactly my PhD topic would be. My advisor, Alberto Lavelli, took a monk like stance and encouraged me to do whatever I find interesting. I used to report him during the weekly meetings what I had been reading, and he patiently used to listen my complains – "biomedical domain is difficult", "people have already investigated whatever problem I can think of", etc etc. He used to tell me that it is completely normal having such anxiety, confusion and difficulty at the beginning, I have to stick through. I did. And here I am, at the end of this journey full of emotions – boundless joy whenever a paper got accepted, sheer anger towards the (usually 3rd) reviewer(s) whenever a paper got rejected, hair pulling in late night whenever my idea produced something completely different than what I expected, jealous feeling towards Yashar (my office mate and one year senior PhD student) for his outstanding record on publishing regularly in *ACL conferences, and, among many other things, extreme desire to break my laptop whenever it shut down by becoming hot and I had not saved something that I had been working for half an hour!

I am so excited not just because I am finishing but also because of what I have learnt during this PhD research. There are a number of people to whom I am indebted for their help and support during this period to achieve whatever I was able to do.

First and foremost, I am grateful to my advisor, Alberto Lavelli. I cannot thank him enough for being such a wonderful supervisor. He never forced me for anything and was so nice to me. If I said anything during discussion that does not make any sense, he would say something like

"probably, I would do it in this other way". I always had free access to him if I wanted to discuss or show something, and he would listen patiently to my non-stop crazy ideas. His approach of letting me to do whatever I like allowed me to learn new things, face different types of problems and become a much better researcher than what I used to be. I also thank him for his helps and advices in non-academic issues.

I am also thankful to Bernardo Magnini for his advices, support and encouragement. He made sure that I do not face funding problems to attend conferences, and also introduced me to some well known researchers.

I would like to thank Pierre Zweigenbaum, mentor of my PhD internship. His expertise and breath of knowledge in biomedical domain and linguistics helped me to gather many insightful knowledge. He took a lot of trouble to complete the bureaucratic processes for the internship and to find a place for me and my wife to stay. I cannot thank him enough for that.

I would like to thank the members of my PhD examination committee – Eric Gaussier, Massimo Poesio and Pierre Zweigenbaum – for their insightful feedback.

Some other people whom I would like to thank include Lorenza Romano, Claudio Giuliano, Matteo Negri, Asma Ben Abacha, Sara Tonelli, Alessandro Moschitti, Yashar Mehdad, Kateryna Tymoshenko, Marco Guerini, Dasha Bogdanova, and Stefan Rigo. I would also like to thank all the remaining members of the HLT unit of FBK-irst for being so helpful and supportive in these years, and the members of the ILES group of the LIMSI-CNRS for being so kind to me during my internship.

I would like to thank my wife, Safwana. Her love and caring have an inestimable influence in preserving the happy outlook of my life through the good and rough times during these years.

Most importantly, I want to thank my parents, my sisters, my brother-

in-laws, and my nephew and nieces, for all they did for me, for their love, and for always being there for me. I am blessed to have such a wonderful family.

Thanks also to all of my friends (near and far, old and new) for being so helpful and supportive, especially to Rahat.

Also, thanks to Facebook, VoipWise, Rynga, Skype, Gmail and Yahoo! for their wonderful technologies which allowed me always to be in touch with my family and friends, and to Trento for being such a friendly place to live.

Lastly, I express my gratitude to the almighty for standing with me in the toughest times, and for giving me patience and strength.

Thank you all very much.

— Faisal

Trento, April 2013

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> *"What information consumes is rather obvious: it consumes the attention of its recipients. Hence, a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."*
>
> – Herbert A. Simon
> Turing Award (1975), Nobel Prize (1978)

1.1 Background

The massive volume of biomedical text, partly due to the exponential growth of biomedical literature in recent years, has made the development of Biomedical Text Mining (BioTM) solutions indispensable. As Zweigenbaum et al. (2007) argued, it has become extremely difficult for biologists to keep up with the relevant publications in their own discipline, let alone publications in other, related disciplines. The rapidly growing amount of clinical texts (e.g. Electronic Health Records or EHR, the parallel growth of narrative data of telemedicine in electronic form, etc) also requires the usage of text mining techniques for improving the quality of care and reducing medical errors.

Clinical texts and biomedical literature are sources of authentic medical knowledge which is crucial for eHealth[1] applications. These applications have a huge commercial prospect. According to the US National Center for Health Statistics[2], 51% of USA adults had used internet for health information in 2009. This potential commercial prospect has led to the launch of several public and commercial health related websites[3]. Online eHealth applications that directly interact with patients can help early identification of disease symptoms, adverse drug affects and even first aid.

Biological researches such as the human genome project and the (upcoming decade-long) project to examine the workings of the human brain and build a comprehensive map of its activity[4] are also very much dependent on the effective use of the knowledge dumped inside the sheer volume of biomedical literature. Medline[5] contains more than 22 millions abstracts from medicine, biology, biochemistry, etc, as of now. Approximately 90% of these abstracts are in English. According to the literature statistics[6], only in the year 2008, there were 751,387 biomedical research publications in English. Other languages also have a considerable number of yearly publications (e.g. French: 10,200; German: 7,586; Italian: 1,678, etc) although far less than English. Pharmaceutical companies invest in various researches of literature-based discovery since it has been often proved to be a potential source of promising hypotheses.

All the above are some of the key reasons that contributed in the surge of

[1]E-health (or eHealth) is the process of providing health care via electronic means, in particular over the Internet. It can include teaching, monitoring (e.g. physiologic data), and interaction with health care providers, as well as interaction with other patients afflicted with the same conditions. (Reference: Robert Pretlow, URL: http://www.ehealthnurse.com/ehealthi.html [accessed on December 10, 2010])

[2]http://www.cdc.gov/nchs/data/hestat/healthinfo2009/healthinfo2009.htm

[3]E.g. www.healthcentral.com, www.healthline.com, etc

[4]Information source: http://www.nytimes.com/2013/02/18/science/project-seeks-to-build-map-of-human-brain.html

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

[6]http://dan.corlan.net/medline-trend.html

an increasing interest of the natural language processing (NLP) community for BioTM. Due to this growing interest, a considerable effort has been devoted to the development of various linguistic resources (e.g. ontologies and corpora), which are pre-requisites for performing various BioTM tasks.[7]

Information extraction (IE) is a text mining task whose goal is to automatically extract specific target information from machine-readable human language texts by means of various NLP techniques. The aim is to identify the information contained inside the text data in a structured form that is more amenable to database or data mining algorithms (Grishman, 2003; Bunescu, 2007).

While there exist keyword-based search systems such as Entrez[8] for biomedical domain, IE-based solutions are much more desirable. The structured output of IE approaches can be used by end user applications having different purposes such as computerized clinical decision support, bio-surveillance, personalized medicine, comparative effectiveness studies, automatic terminology management, question answering, summarization, statistical analysis and many more, and even for semantic web or for uncovering hidden, indirect links to propose potential scientific hypotheses[9].

Though a few well-defined IE tasks, such as gene/protein mention recognition, have achieved a sufficient level of maturity, solutions for most of the biomedical IE tasks are far from being robust and practically usable. Biomedical texts are substantially different from other texts such as newspaper articles. Ranging from the terminology and sentence construction to the valency and semantics of verbs, these texts show an inherently complex structure. Authors of these texts often do not follow proposed standard-

[7]For example, see the lists in http://www2.informatik.hu-berlin.de/~hakenber/links/benchmarks.html or http://www.ims.uni-stuttgart.de/~jasmin/corpora.html

[8]http://www.ncbi.nlm.nih.gov/sites/gquery

[9]One such example is the discovery of relation between fish oil and Raynaud's disease that was hypothesized by Swanson (1986) in his seminal paper after he linked information of several biomedical literature.

ized names or formats, which complicates things further. As there is the possibility of introducing serious health-related risks due to the provision of any wrong information, it is very critical to provide information to the end user/system with the maximum possible accuracy.

In recent years, shared tasks/challenges such as TREC Genomics track[10], KDD cup[11], LLL05 challenge task[12], BioCreAtIvE[13], BioNLP shared tasks[14,15,16,17], CMC medical NLP challenge[18], i2b2 shared tasks[19] and CALBC challenges[20] have been pushing the limits of the biomedical/clinical IE research. These challenges are based on different visions and approaches. For example, the BioCreAtIvE challenge emphasizes the development of text mining techniques to help database curation, whereas the i2b2 challenge focuses on identifying concepts, co-references among concepts and relations in clinical notes. In any case, all of these shared tasks emphasize the improvement of the state of the art of some fundamental IE tasks. This is important since more advanced BioTM tasks (e.g. automatic discovery of protein pathways) are difficult to be accomplished unless the fundamental IE tasks reach a sufficient maturity.

1.2 Tasks Investigated

In the context of this thesis, we investigate three core IE tasks. The first is **Named Entity Recognition (NER)**, i.e. the task of locating the

[10]http://ir.ohsu.edu/genomics/
[11]http://www.sigkdd.org/kddcup/
[12]http://www.cs.york.ac.uk/aig/lll/lll05/
[13]http://www.biocreative.org/
[14]http://www.nactem.ac.uk/tsujii/GENIA/ERtask/report.html
[15]http://www.nactem.ac.uk/tsujii/GENIA/SharedTask/
[16]https://sites.google.com/site/bionlpst/
[17]http://2013.bionlp-st.org/
[18]http://www.computationalmedicine.org/challenge/index.php
[19]https://www.i2b2.org/NLP/TemporalRelations/PreviousChallenges.php
[20]http://www.calbc.eu/

boundaries of the entity mentions in a text and tagging them with their corresponding semantic type (e.g. person, location, protein, disease, ...). For example, consider the following sentence:

> *The most common symptom of coronary artery disease is angina or "angina pectoris", also known simply as chest pain.*

A named entity recognizer that is trained for recognizing disease and symptom mentions should identify the four named entities (present in the given text) shown in the upper part of Table 1.1.

Entity mention name (*Entity type*):
"coronary artery disease" *(disease)*, "angina pectoris" *(symptom)*, "angina" *(symptom)*, "chest pain" *(symptom)*
Coreference type *{list of coreferreing mentions}*:
Coreferring symptom mentions *{angina, angina pectoris, chest pain}*
Relation type (*arg1, arg2*):
symptomOf *(angina, coronary artery disease)*
symptomOf *(angina pectoris, coronary artery disease)*
symptomOf *(chest pain, coronary artery disease)*

Table 1.1: An example of information extraction from the following given text: *'The most common symptom of coronary artery disease is angina or "angina pectoris", also known simply as chest pain'.*

Assuming that the relevant named entities have been correctly identified, a further step is to find if different textual expressions/mentions denote the same real world entity. This task is known as **co-reference resolution**, which is the second IE problem that we address in this PhD research. In the example shown in Table 1.1, *angina*, *angina pectoris* and *chest pain* – all refer to the same real world entity.

The third IE problem that we address is called **relation extraction (RE)**. This is the *main focus* of this PhD research. The goal of RE is to identify instances of pre-defined semantic relation types that exist between

pairs of entity mentions in a given text. Intuitively, in a real scenario, RE is the next step after NER and co-reference resolution. As an example, if a RE system is trained for extracting symptoms of medical conditions from text, and if the sentence of Table 1.1 is provided as input to that RE system, it is expected to find the relations shown in the lower part of Table 1.1.

In the following chapters of this thesis, we will more elaborately describe the above tasks as well as summarize the state of the art in the different fields.

1.3 Primary Research Goals

One of our primary research goals is to develop robust IE approaches. For RE, none of the existing RE approaches tested on multiple biomedical benchmarks is shown to be consistently better than the other approaches. A somewhat similar problem also exists for NER and coreference resolution. With regard to biomedical NER (BNER), most of the research is focused on gene/protein mention identification. As a result, it is usually difficult to assess whether these BNER systems would work as well for identifying other bio-entity mentions. In case of coreference resolution, the problem is that there are multiple evaluation metrics and there is hardly any approach which fares well for all of these metrics.

Another research goal is to harness the benefits that various linguistic aspects can bring. IE research has come a long way. The addition of new techniques (e.g., the usage of word clustering for RE (Sun et al., 2011)) and the usage of rich semantic resources (e.g. Yago (Suchanek et al., 2007)) are some of the key advancements in the recent years. Yet there remain a lot of unexplored options, particularly in exploiting the outcomes produced in other computational linguistics fields. Bearing this in mind, one of our

goals is to exploit some of such linguistic phenomena, such as the scope of negation cues, the discourse units of sentences and the semantic roles of entity mentions. In addition, we also envisaged to propose and use new linguistic features and linguistically motivated rules.

Last but not least, we want to develop pro-active IE approaches. No matter how good a ML algorithm is and the features it uses are, the final performance of the ML-based system would depend on several other issues as well. One of such issues is the imbalance in the number of instances per class (for binary cases, positive and negative instances) which is quite common in many NLP (including IE) tasks. Previous studies have empirically shown that unbalanced datasets lead to poor performance for the minority class (Weiss and Provost, 2001). Apart from some exceptions, the number of negative instances is usually higher than that of the positive instances. As Gliozzo et al. (2005) argued, in most cases the error rate of a classifier trained on a skewed dataset is typically very low for the majority class and this results in biased estimation (Kotsiantis and Pintelas, 2003) and suboptimal classification performance (Chawla et al., 2004).

Some ML techniques have built-in mechanisms to deal with the skewness in somewhat limited scope[21]. But this does not guarantee to completely overcome the impact of skewness. Some ML algorithms (e.g. kNN) do instance pruning during training while maintaining the generalization accuracy. However, the main drawback of such techniques is the increased time complexity, which is generally quadratic in the data set size, without any guarantee of performance improvement (Gliozzo et al., 2005).

We believe that, for IE classification tasks (and also in other NLP tasks), the traditional approach of using ML-based classifiers for training on annotated data (we will refer to these classifiers as *objective classifiers*) and

[21]E.g., SVM allows to provide a cost-factor by which training errors on positive instances outweigh errors on the negatives.

then using them to predict class labels on test data is not fully adequate. Such approaches should be complemented with additional layers, where separate rule-based or ML-based classifiers (we will refer to them as *expert classifiers*) are deployed.

The task of (some of) the expert classifiers should be to determine which of the training instances will be kept (and which others will be discarded) for the training of the objective classifier(s). If the expert classifiers are highly accurate in filtering out less informative instances and reducing the imbalance in data distribution, then this will guarantee not only faster training and test time, but also enable to train more accurate and focused objective classifiers.

Similarly, expert classifiers should also focus on identifying as many true negative instances (of the test data) as possible with no or very few mistakes, before using the objective classifiers. The identification of such instances would limit the focus of the objective classifiers on the remaining "relatively hard" test instances. It is practically impossible for a ML-based (objective) classifier to identify correct class labels for every test instance of an NLP task. Despite of being trained with diverse and rich set of features, objective classifiers often make silly mistakes for some apparently obvious instances which can be easily dealt with expert knowledge. Hence, the highly accurate filtering of true negatives using expert classifiers might reduce the number of false positives to be identified by the objective classifiers.

As noted above, the issue of imbalance in data distribution has been already addressed, even if partially. For example, in the context of named entity recognition (NER), stopword filtering is used to reduce the number of candidate tokens to be considered as target entity mentions (Gliozzo et al., 2005; Giuliano et al., 2006b). However, for some of the tasks, such as RE, to the best of our knowledge, there is no study showing that the

reduction of skewed distribution can lead to better results.

1.4 Thesis Contributions

1.4.1 Named Entity Recognition

While most of the *biomedical named entity recognition (BNER)* work is focused on protein/gene mention tagging, other entities (e.g. disease) have not received enough attention (Jimeno et al., 2008). BNER for protein/gene has already achieved a sufficient level of maturity (Torii et al., 2009). However, the lack of availability of adequately annotated corpora has hindered the progress of BNER research for other semantic types (Jimeno et al., 2008; Leaman et al., 2009). Even if annotated corpora are available in some cases, they are often small in size. To overcome the shortage of annotated corpora for training ML systems, recently there has been an initiative under the European project CALBC[22] to create a huge, so called, **silver standard corpus (SSC)** using harmonized annotations done by multiple BNER annotation systems (Rebholz-Schuhmann et al., 2010b; Rebholz-Schuhmann et al., 2010a). Such annotation systems are expected to take advantage of existing dictionaries, lexicons, etc, to provide reliable annotations in a huge unannotated dataset.

In the context of BNER, our research contribution is twofold.

A state-of-the-art BNER approach for disease and other bio-entity types

Firstly, we propose a ML-based approach which is able to identify different types of bio-entity mentions (such as diseases, chemicals, species and proteins/genes) with high performance. Specifically, it obtains state-of-the-art

[22]http://www.ebi.ac.uk/Rebholz-srv/CALBC/project.html

results for the recognition of disease entity mentions.

Assessment of the practical usability of a machine annotated huge corpus

Secondly, we also investigate how and to what extent can a silver standard corpus be exploited for BNER. Since the creation of manually annotated training data is a costly process, we propose how to automatically improve the quality of SSC corpora that can enable the training of more accurate ML classifiers. In this process, we propose a very simple approach, the usage of *condensed corpus* (defined in Chapter 2) instead of full training corpus, to reduce imbalance in instance distribution.

1.4.2 Coreference Resolution

One of the open issues regarding coreference resolution is the usage of different evaluation metrics and the disparity in the scores obtained by the systems. So, our main research objective is to investigate whether it is possible to minimize the differences among the scores obtained for different metrics, at the same time maintaining their unweighted average as high as possible.

A greedy and constrained supervised coreference resolution approach

We propose a greedy and constrained supervised coreference resolution approach that not only achieves high results in clinical texts for each individual entity mention type, but also for four different evaluation metrics usually computed for assessing systems' performance. Our proposed approach combines a series of syntactically and semantically motivated constraints that control the generation of less-informative/sub-optimal training and

test instances, and also some aggressive greedy strategies during the chain clustering.

1.4.3 Relation Extraction

A significant portion of biomedical RE research has been conducted on protein-protein interaction (PPI) extraction, due to the importance of the task. None of the previous biomedical approaches that have been tested on the 5 widely used PPI benchmark corpora consistently outperform other approaches. Besides, their performances are not high in most of these corpora.

In the context of RE, our research contribution is manifold.

A novel hybrid kernel-based RE method tested on multiple RE tasks and corpora

We propose a novel hybrid kernel-based RE approach that outperforms the previous approaches in 4 out of the 5 benchmark PPI corpora. In addition, our result is very close to the state of the art on the other corpus.

Our proposed approach also outperforms previous best results on two benchmark drug-drug interaction (DDI) corpora, i.e. on a separate biomedical RE task. Moreover, the results of our proposed approach on a benchmark news domain corpus, which consist of completely different genres of texts, are on a par with the state of the art, too.

Exploitation of negation scopes, discourse units and semantic roles

We propose a self-supervised technique to exploit the scope of negations for RE without using any corpus specifically annotated with the scope of negations. This technique can be exploited to filter less informative sentences which would allow to reduce imbalance in data distribution. We

also propose how to exploit knowledge accumulated using a data driven technique with already known common knowledge to reduce imbalance in data distribution. Our proposed data driven technique for knowledge collection is based on indirect exploitation of semantic role labelling. In addition, we propose an approach to exploit elementary discourse units to filter negative test instances. We show that the reduction of skewed distribution by exploiting the information provided by the above mentioned linguistic information could lead to better results.

A linguistically informed approach

We propose new structures (namely, mildly extended dependency tree and reduced graph) to separate important part of a sentence (with respect to a pair of candidate entity mentions) to extract target relation. A number of linguistically motivated rules were also proposed for extracting a variety of features as well as for preprocessing the input data.

1.5 Outline of the Thesis

Below is a summary of the remaining chapters in this thesis, with reference to the relevant publications:

- **Chapter 2 – Biomedical Named Entity Recognition:** This chapter starts with a discussion of the related work on biomedical NER. This is followed by our proposed approach and the corresponding empirical results (Chowdhury and Lavelli, 2010a). We also present our findings regarding how machine annotated corpora can be maximally exploited (Chowdhury and Lavelli, 2011a; Rebholz-Schuhmann et al., 2011).
- **Chapter 3 – Coreference Resolution:** At first, we describe the existing state of the art and the issues regarding coreference resolution.

Then, we present our proposed approach for coreference resolution on clinical text and make a comparison of our results with those of previous work (Chowdhury and Zweigenbaum, 2013; Zweigenbaum et al., 2013). Finally, we discuss the limitations and possible future extension of our approach.

- **Chapter 4 – Relation Extraction:** We start with a brief overview of the current state of RE research. Then, we propose our novel RE approach (Chowdhury et al., 2011; Chowdhury and Lavelli, 2011b; Chowdhury et al., 2011c; Chowdhury and Lavelli, 2012b; Chowdhury and Lavelli, 2012a) and its further extensions (Chowdhury and Lavelli, 2012c; Chowdhury and Lavelli, 2013a; Chowdhury and Lavelli, 2013b), accompanied with discussion on experimental results. This is followed by suggestions for future expansions.

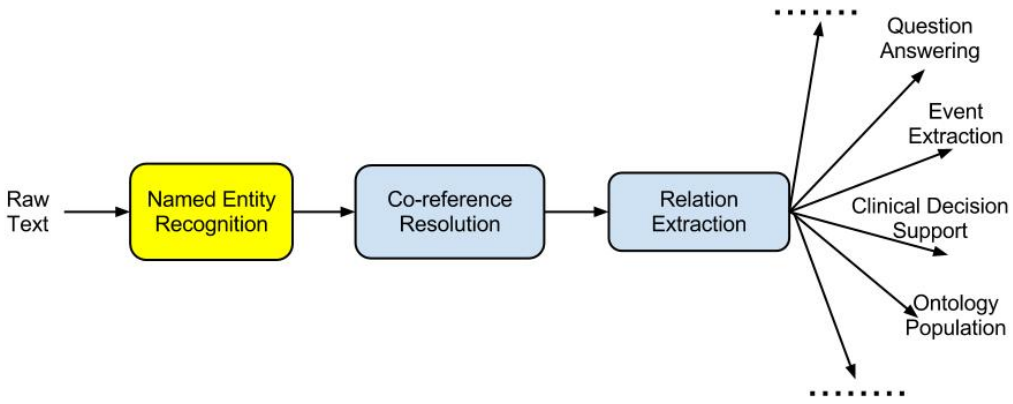
- **Chapter 5 – Conclusion:** We conclude with a summary of this PhD thesis and possible directions for future research.

Chapter 2

Biomedical Named Entity Recognition

"The beginning of wisdom is to call things by their right names."

– A Chinese proverb



Named Entity Recognition is the task of locating entity mentions in texts and recognizing their appropriate semantic types. It is usually the first step towards making full use of the information contained inside texts. Our interest for biomedical named entity recognition (BNER) was primarily driven by the need of developing a high performance NER system that

we could later use as a preliminary step for relation extraction, which is the main focus of this PhD research. In the context of BNER, we had two research goals.

Firstly, we worked on an approach that could be applicable to different types of biomedical entities obtaining state-of-the-art results. In this regard, we focused our attention particularly on recognizing disease mentions, a type of biomedical entities which did not attract much attention when this PhD started. The resulting BNER system has been later used as a preliminary step for relation extraction (more details in Chapter 4).

Secondly, we wanted to investigate a recently emerging topic, called silver standard annotation, in the context of BNER. The goal of such annotation is to automatically annotate large corpora using different automatic systems (instead of human experts), and then to learn models/classifiers from such corpora. This is a rather unconventional approach which requires assessing whether such corpora can really enable to train high performing systems.

In the subsequent sections, we will describe our approach for each of the above stated goals. Each section begins with a brief literature review of the related work, followed by a detailed discussion of the contributions of this thesis.

2.1 Related Work on Biomedical NER

Most of the work on NER has initially focused on news domain. However, the features, pre-processing and post-processing used in these work are not equally effective on biomedical text, unless domain specific knowledge and techniques are incorporated. Biomedical texts are substantially different from other genres of text (such as newspaper articles). Ranging from the terminology and sentence construction to the valency and semantics of

verbs, these texts show an inherently complex structure. New bio-entity names are created continuously. Besides, authors of biomedical texts often do not follow proposed standardized names or formats and prefer to use abbreviations or other forms depending on personal inclination (Bodenreider, 2004; Dai et al., 2010). Because of their limited length, such abbreviations/acronyms are sometimes identical to other words or symbols which increases the ambiguity. For instance, it was reported that 80% of the abbreviations listed in the UMLS have ambiguous representation in MEDLINE (Liu et al., 2002). Sometimes the same name is shared by different types of bio-entity types. For example, "C1R" is a cell line, but there exists a gene (SwissProt P00736) that has the same name. Usage of digits and other non-alphabetic characters inside bio-entity names is also common. Compound names further complicate the situation. Locating the beginning and ending of such names within a sentence is not so straightforward since verbs and adjectives are often embedded in such names. Due to these complexities, BNER attracted a huge amount of research interests. A number of shared tasks/challenges such as BioNLP/NLPBA 2004, BioCreative, CALBC, etc provided benchmarks to compare and showcase the advancement in this field.

State-of-the-art BNER approaches use various ML algorithms. These include hidden Markov model (HMM), support vector machine (SVM), maximum entropy Markov model, conditional random fields (CRFs), ... Among these algorithms, CRFs appear to be the most popular choice.

One common characteristic in many of these systems is the combination of results from multiple classifiers (e.g. see Torii et al. (2009)). Apart from that, there is a substantial agreement among the feature sets used by these systems, most of which are actually various orthographic features.

Most of the work to date on BNER is focused on genes/proteins. The state-of-the-art gene/protein mention recognition systems achieve F-scores

around 88%, which is quite high. These systems often use either gene/protein specific features (e.g. Greek alphabet matching) or post-processing rules (e.g. extension of the identified mention boundaries to the left when a single letter with a hyphen precedes them (Torii et al., 2009)) which might not be equally effective for other bio-entity type identification (more in Section 2.3.4). More efforts should be devoted to take advantage of contextual clues and features.

One of the important bio-entity types which did not receive as much attention (when this PhD started) as gene/protein is disease, which is a particular topic of interest in this thesis. In the last few years, some disease annotated corpora have been released. However, they have been annotated primarily to serve the purpose of relation extraction and, for different reasons, most of them are not suitable for the development of ML based disease mention recognition systems (Leaman et al., 2009). For example, the BioText (Rosario and Hearst, 2004) corpus has no specific annotation guideline and contains several inconsistencies, while the PennBioIE (Kulick et al., 2004) is very specific to a particular sub-domain of diseases. Among other disease annotated corpora, the EBI disease corpus (Jimeno et al., 2008) is not annotated with disease mention boundaries which makes it unsuitable for BNER evaluation for diseases. Recently, an annotated corpus, named Arizona Disease Corpus (AZDC) (Leaman et al., 2009), has been released which has adequate and suitable annotation of disease mentions by following specific annotation guidelines.

There has been some work on identifying diseases in clinical texts, especially in the context of CMC medical NLP challenge and i2b2 challenge. However, as noted by Meystre et al. (2008), there are a number of reasons that make clinical texts different from texts of biomedical literature, e.g. composition of short, telegraphic phrases, use of implicit templates and pseudo-tables, ... Hence, the strategies adopted for NER on clinical texts

are not the same as the ones practiced for NER on biomedical literature.

Bundschus et al. (2008) relied on a conditional random field (CRF) based approach that uses typical features for gene/protein mention recognition for the purpose of disease, gene and treatment recognition; i.e. there was no feature tailoring for disease recognition. The work has been evaluated on two corpora annotated with those entity mentions that participate in disease-gene and disease-treatment relations. The entity mentions which do not participate in any of these two types of relations are not annotated in those corpora. Hence, their reported F-score for entity mention recognition was not computed for all mentions, regardless of whether a mention is participating in a relation or not. Furthermore, the authors do not indicate which F-score has been specifically achieved for disease recognition. Hence, the reported results are not suitable for comparison.

To the best of our knowledge, the only systematic experimental results reported for disease mention recognition in biomedical literature using ML based approaches are published by Leaman and Gonzalez (2008) and Leaman et al. (2009).[1] They have used a CRF based BNER system named BANNER which basically uses a set of orthographic, morphological and shallow syntactic features (Leaman and Gonzalez, 2008). The system achieves an F-score of 86.43 on the BioCreative II GM corpus[2], which is one of the best results for gene mention recognition task on that corpus.

BANNER achieves an F-score of 54.84 for disease mention recognition on the BioText corpus (Leaman and Gonzalez, 2008). However, as said above, the BioText corpus contains annotation inconsistencies[3]. So, the corpus is not ideal for comparing systems' performance. The AZDC corpus

[1]However, there is some work on disease recognition in biomedical literature using other techniques such as morpho-syntactic heuristic based approach (e.g. MetaMap (Aronson, 2001)), dictionary look-up method and statistical approach (Névéol et al., 2009; Jimeno et al., 2008; Leaman et al., 2009).

[2]As mentioned in `http://banner.sourceforge.net/`

[3]`http://biotext.berkeley.edu/data/dis_treat_data.html`

is much more suitable as it is specifically annotated for benchmarking of disease mention recognition systems. An improved version of BANNER achieves an F-score of 77.9 on the AZDC corpus, which is the state of the art on ML based disease mention recognition in biomedical literature (Leaman et al., 2009).

2.2 Proposed Approach

Previous studies argued that the correct identification of diseases is the most promising candidate for the improvement of various disease-centric knowledge extraction tasks (e.g. drug discovery (Agarwal and Searls, 2008), disease-gene relation extraction (Bundschus et al., 2008)). The identification of disease names can also be useful in other tasks, such as drug-drug interaction extraction, where diseases are not necessarily the entities of interest.[4]

So, our initial effort was concentrated on developing a machine learning based BNER approach that used a feature set specifically tailored for disease mention recognition. We call our system **BioEnEx** (Biomedical Named Entity Extractor). We used conditional random field (CRF) (Lafferty et al., 2001) as the learning algorithm.[5] The main characteristics of our approach are as follows:

- More emphasis on contextual and syntactic features (specifically, we investigated the exploitation of syntactic dependencies which were largely overlooked by previous approaches)
- Extensive segmentation, normalization and simplification of the tokens
- Usage of a single classifier based approach

[4]We will explain this issue in more details in Section 4.3.2.

[5]Our system uses Mallet (McCallum, 2002) to train a first-order CRF model.

  - Most of the state-of-the-art BNER systems multiple classifiers to tag tokens and then later tries to merge the outputs. It is a complex and computational resource intensive approach. Moreover, there are certain difficulties in case of disagreements and overlaps. Furthermore, it is not clear how the classifiers complement each other, which may result in unreliable error analyses.

- Beyond genes/proteins

  - Our system obtains high results for diseases, genes/proteins, drugs and species.

- Exploitation of linguistic structures during post-processing

- Exploitation of **condensed corpus** rather than the full training corpus to reduce the imbalance of positive and negative training instances.

  - A **condensed corpus** is the collection of those training sentences which have at least one (target) entity mention annotation. Usually in the full training corpus, there exist sentences which do not contain any (target) entity mention annotation. The condensed corpus excludes those sentences and, consequently, reduces skewness in the distribution of positive and negative instances.[6]

2.2.1 Description of the Proposed Approach

There are basically three stages in our approach: (i) pre-processing, (ii) feature extraction and model training, and (iii) post-processing.

[6]Note that, for NER, every token is an instance and any token that is not annotated as a part of an (target) entity mention is a negative instance. So, usually the total number of negative instances is significantly higher than that of the positive instances.

Pre-processing

Data pre-processing is an important step which can affect the features to be used by the ML algorithm (and, hence, also the performance) of the NER tagger. After tokenization, POS-tagging and parsing of the sentences[7], our system conducts the generalization and normalization techniques shown in Table 2.1.[8]

All of these techniques have a common goal – minimize (as much as possible) the uninformative dissimilarity among tokens and make them uniform. We selected these techniques from a variety of pre-processing techniques mentioned in the previous BNER literature after doing an in-depth experimental study.[9] At the same time, we kept a separate copy of the original tokens so that no information is lost. Tokens (both original and normalized) are labelled with the corresponding target entity annotations according to the IOB2 format.

1. Each number (both integer and real) inside a token is replaced with *'9'*.
2. Each token is further split if it contains either punctuation characters or both digits and alphabetic characters.
3. All letters are changed to lower case.
4. All Greek letters (e.g. alpha) are replaced with *G* and Roman numbers (e.g. iv) with *R*.
5. Each token is normalized using SPECIALIST lexicon tool[10] to avoid spelling variations.

Table 2.1: Generalization and normalization steps of our BNER system.

[7]We use the Charniak-Johnson reranking parser (Charniak and Johnson, 2005), along with a self-trained biomedical parsing model (McClosky, 2010), for tokenization, POS-tagging and parsing of the sentences. Later, we used the Stanford parser (Klein and Manning, 2003) for the dependency analysis.

[8]In our original paper (Chowdhury and Lavelli, 2010b), we used GeniaTagger for tokenization and PoS tagging. Hence, there is a negligible difference between the results of BioEnEx reported in this thesis and in Chowdhury and Lavelli (2010b).

[9]Results of these particular experiments are not included in this thesis.

[10]`http://lexsrv3.nlm.nih.gov/SPECIALIST/index.html`

Feature extraction and model training

It is implicit that no matter how good the ML algorithm is, if the feature set to be used is not diverse and informative, the NER system will not perform well. The exiting studies already proposed a rich variety of features. We selected the best possible set of features among them after doing ablation experiments. These selected features can be categorized into the following groups:

- general features (Table 2.2)
- orthographic features (Table 2.3)
- contextual features (Table 2.4)
- syntactic dependency features (Table 2.5)
- dictionary lookup features (see Section 2.2.2)

We describe them in more detail in Section 2.2.2. Note that we selected a tailored set of features for the disease mentions, and some additional features for the other types of bio-entity mentions. In other words, we use a generalized feature set for all entities except diseases.

Post-processing

The objective of the post-processing is twofold – firstly, to reduce the number of wrong annotations identified by the NER tagger, and secondly, to include annotations which seem highly probable but are missed by the NER tagger. We propose the following post-processing techniques:

- *Bracket mismatch correction:* If there is a mismatch of brackets in the identified mention, then the immediate following (or preceding) character of the corresponding mention is checked and included inside

the mention if such character is the missing bracket. Otherwise, all the characters from the index where the mismatched bracket exists inside the identified mention are discarded from the corresponding mention.

- *One sense per sentence:* If any instance of a character sequence is identified as a mention of the target entity type, then all the other instances of the same character sequence inside the same sentence are also annotated as such target entity type.

- *Short/long form annotation:* Using the algorithm of Schwartz and Hearst (2003), *"long form (short form)"* instances are detected inside sentences. If the short form is annotated as a mention of the target entity type, then the long form is also annotated and vice versa.[11]

- *Ungrammatical conjunction structure correction:* If an annotated mention contains comma (,) but there is no "and" in the following character sequence (from the character index of that comma) of that mention, then the annotation is splitted into two parts (at the index of the comma). Annotation of the original mention is removed and the splitted parts are annotated as two separate mentions.

- *Short and long form separation:* If both short and long forms are annotated in the same mention, then the original mention is discarded and the corresponding short and long forms are annotated separately.

2.2.2 Features

As we mentioned earlier, our main objective is to develop a system that will have high accuracy for disease mention recognition. There are compelling reasons to believe that various issues regarding the well studied

[11]This post-processing rule was not useful on the AZDC disease mention recognition corpus because it generates a lot of false positives. Our random analysis indicates that many of these false positives are actually true positives but were not annotated inside the AZDC corpus (i.e. missing annotations).

Feature name	Description	Used for diseases?	Used for all the other entities?
PoS	Part-of-speech tag	yes	yes
NormWord	Normalized token (see Section 2.2.1)	yes	yes
Lemma	Lemmatized form	yes	yes
charNgram	3 and 4 character n-grams	yes	yes
Suffix	2-4 character suffixes	yes	yes
Prefix	2-4 character prefixes	yes	yes

Table 2.2: General features for token$_i$

gene/protein mention recognition would not apply to the other semantic types. For example, Jimeno et al. (2008) argue that the use of disease terms in biomedical literature is well standardized, while it is quite the opposite for the gene terms (Smith et al., 2008).

After a thorough study and extensive experiments on various features and their possible combinations, we have selected a feature set specific to the disease mention identification which comprises features shown in Tables 2.2, 2.3, 2.4 and 2.5, and dictionary lookup features. Additional features used for other bio-entities are selected based on the analysis reported by other studies.

Previous studies have shown that dictionary lookup features[12], i.e. name matching against a dictionary of terms, often increase recall (Torii et al., 2009; Leaman et al., 2009). However, an unprocessed dictionary usually does not boost overall performance (Zweigenbaum et al., 2007). So, to reduce uninformative lexical differences or spelling variations, we generalize and normalize the dictionary entries using exactly the same steps followed

[12]If a sequence of tokens in a sentence matches an entry in the dictionary, a feature "B-DB" is added for the leftmost token of that sequence. For each of the remaining tokens of the sequence, features "I-DB" are added. If a token belongs to several dictionary matches, then all the dictionary matches except the longest one are discarded. During dictionary lookup feature extraction, we ignored punctuation characters while matching dictionary entries inside sentences.

Feature name	**Description**	**Used for diseases?**	**Used for all the other entities?**
InitCap	Is initial letter capital	yes	yes
AllCap	Are all letters capital	yes	yes
MixCase	Does contain mixed case letters	yes	yes
SingLow	Is a single lower case letter	yes	yes
SingUp	Is a single upper case letter	yes	yes
Num	Is a number	yes	yes
PuncCharType	Token_i itself, if it is a punctuation character (e.g. - / [ ] : ; % , . etc)	yes	yes
PrevCharAN	Is previous character alphanumeric	yes	yes
Shape of token_i	For example, "Animal" would be mapped to "Aaaaaa". See Collins (2002) for details.	no	yes
Brief shape of token_i	For example, "Animal" would be mapped to "Aa".	no	yes
Nucleoside	Is token_i a Nucleoside name	no	yes
Nucleotide	Is token_i a Nucleotide name	no	yes
AminoAcidLong	Is token_i a long Amino acid name	no	yes
AminoAcidShort	Is token_i a short Amino acid name	no	yes
NucleicAcid	Is token_i a Nucleic acid name	no	yes
ROMAN	Is token_i a Roman number	no	yes
GREEK	Does token_i match a Greek letter	no	yes
HasGREEK	Does token_i contain a Greek letter	no	yes

Table 2.3: Orthographic features for token_i

Feature name	Description	Used for diseases?	Used for all the other entities?
Bi-gram$_{k,k+1}$ for $i-2 \leq k < i+2$	Bi-grams of normalized tokens	yes	yes
Tri-gram$_{k,k+1,k+2}$ for $i-2 \leq k < i+2$	Tri-grams of normalized tokens	yes	yes
CtxPOS$_k$ for $i+1 \leq k \leq i+2$	POS of the following two tokens	yes	yes
CtxLemma$_k$ for $i+1 \leq k \leq i+2$	Lemma of the following two tokens	yes	yes
CtxWord$_k$ for $i-2 \leq k < i+2$	Previous two and following two tokens	yes	yes
Offset conjunctions	New features from all possible conjunctions among features of the tokens from token$_{i-1}$ to token$_{i+1}$. See documentation of Mallet (McCallum, 2002) for details.	yes	yes

Table 2.4: Contextual features for token$_i$

Feature name	Description	Used for diseases?	Used for all the other entities?
obj	Target token(s) to which token$_i$ is an object	yes	yes
subj	Target token(s) to which token$_i$ is a subject	yes	yes
nn	Target token(s) to which token$_i$ is a noun compound modifier	yes	yes

Table 2.5: Syntactic dependency features for token$_i$. For example, in the sentence "Clinton defeated Dole", "Clinton" is the *nsubj* of the *target token* "defeated".

for the pre-processing of sentences (see Section 2.2.1).

To reduce the chances of false and unlikely matches, any entry inside the dictionary having less than 3 characters or more than 10 tokens is discarded.

It might be possible that for some of the bio-entity types the existing gazetteers/dictionaries do not conform to the particular annotation guideline followed for the training data annotation. In such cases, it might be beneficial to automatically build dictionaries from the training data. For this purpose, we propose to automatically construct the following dictionaries from training data:

- *Dictionary of non-entity tokens (***DictAlwaysOtherTok***) :* List of normalized unique tokens which are never annotated as part of any target entity mention (i.e. always labelled as "*O*") inside the training data.

- *Dictionary of entities (***DictAlwaysEntTok***) :* List of normalized unique tokens which are always labelled as part of a target entity mention (i.e. labelled as "*B-*" or "*I-*") inside the training data.

As the descriptions of these dictionaries imply, if there is a token which is annotated as a part of a target entity mention (i.e. labelled as "*B-*" or "*I-*") somewhere in the training data, and also annotated as "O" somewhere else, the system does not consider it for *DictAlwaysOtherTok*. Similar strategy is also followed for the *DictAlwaysEntTok*. Table 2.6 shows the features extracted using these dictionaries.

Feature name	Description
AlwaysTaggedOther	whether the (normalized) token is found in *DictAlwaysOtherTok*
AlwaysTaggedAsPartOfTargetEnt	whether the (normalized) token is found in *DictAlwaysEntTok*

Table 2.6: Global perspective features for token$_i$ extracted from automatically built dictionary from training data.

2.3 Experiments

2.3.1 Data

For our experiments we used the Arizona Disease Corpus (AZDC)[13] (Leaman et al., 2009). The corpus has detailed annotations of diseases including UMLS codes, UMLS concept names, possible alternative codes, and start and end points of disease mentions inside the corresponding sentences. These detailed annotations make this corpus a valuable resource for evaluating and benchmarking text mining solutions for disease recognition. Table 2.7 shows various characteristics of the corpus.

Item name	Total count
Abstracts	793
Sentences	2,783
Total disease mentions	3,455
Disease mentions without overlaps	3,093
Disease mentions with overlaps	362

Table 2.7: Various characteristics of AZDC.

In case of overlapping annotations (e.g. the disease annotation "en-

[13]Downloaded from `http://diego.asu.edu/downloads/AZDC/at5-Feb-2009`

dometrial and ovarian cancers" overlaps with another annotation "ovarian cancers"), we have considered only the larger annotations in our experiments. After resolving overlaps according to the aforementioned criterion, there remain 3,224 disease mentions. We have observed minor differences in some statistics of the AZDC reported by Leaman et al. (2009) with the statistics of the downloadable version[14] (Table 2.7). However, these differences can be considered negligible.

2.3.2 Dictionary

For the purpose of disease mention recognition, like Leaman et al. (2009) we have created a dictionary with the instances of the following nine of the twelve UMLS semantic types from the semantic group "DISORDER"[15] of the UMLS Metathesaurus (Bodenreider, 2004): (i) *disease or syndrome*, (ii) *neoplastic process*, (iii) *congenital abnormality*, (iv) *acquired abnormality*, (v) *experimental model of disease*, (vi) *injury or poisoning*, (vii) *mental or behavioral dysfunction*, (viii) *pathological function* and (ix) *sign or symptom*. We have not considered the other three semantic types (*findings*, *anatomical abnormality* and *cell or molecular dysfunction*) since these three types have not been used during the annotation of Arizona Disease Corpus (AZDC) used in our experiments.

2.3.3 Experimental Setting

We follow an experimental setting similar to the one in Leaman et al. (2009) so that we can compare our results with those of the BANNER

[14]Note that *"Disease mentions (total)"* in the paper of Leaman et al. (2009) actually refers to the *total disease mentions after overlap resolving* (Robert Leaman, personal communication). One other remark is that Leaman et al. (2009) mention 794 abstracts, 2,784 sentences and 3,228 (overlap resolved) disease mentions in the AZDC. But in the downloaded version of AZDC, there is 1 abstract missing (i.e. total 793 abstracts instead of 794). As a result, there is 1 less sentence and 4 less (overlap resolved) disease mentions than the originally reported numbers.

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

system. We performed 10-fold cross validation on AZDC in such a way that all sentences of the same abstract are included in the same fold. The results on all folds are averaged to obtain the final outcome[16], as Leaman et al. (2009) did. All the results are computed based on the exact matching criterion, i.e. partial matches are not rewarded.

2.3.4 Results and discussions

Table 2.8 shows the results of the experiments with different features. As we can see, our approach achieves significantly higher results than that of BANNER, the previous state-of-the-art system for disease recognition. Initially, with only the general and orthographic features the performance is not high. However, once the contextual features are used, there is a substantial improvement in the result. Note that BANNER does not use contextual features (see Leaman and Gonzalez (2008)). In fact, the use of contextual features was also quite limited in other BNER systems that achieved high performance for gene/protein identification (Smith et al., 2008), until recently.

Dictionary lookup features provide a very good contribution in the outcome. This supports the argument of Jimeno et al. (2008) that the use of disease terms in biomedical literature is well standardized. Post-processing and syntactic dependency features also increase performance.

We have computed statistical significance tests for the last four experimental results shown in Table 2.8. For each of such four experiments, the immediate previous experiment is considered as the baseline. The tests have been performed using the approximate randomization procedure (Noreen, 1989). We have set the number of iterations to 1,000 and the confidence level to 0.01. According to the tests, the contributions of contextual

[16]To be more precise, the precision and recall scores of all folds are averaged. F-score is computed from these averaged precision and averaged recall.

features and dictionary lookup features are statistically significant. To our dismay, post-processing rules and syntactic dependency features did not significantly improve the results.[17]

We did some random analysis, particularly, to understand why the proposed post-processing techniques were not effective. Our findings suggest that the AZDC corpus is missing a number of annotations for disease names. For instance, consider the term "VHL" which refers to a disease named "Von Hippel-Lindau". There are at least 8 occurrences where our post-processing module was able to correctly identify that the term refers to a disease name in the corresponding sentences. However, since annotators did not annotate them as diseases, the terms identified by our system were considered as false positives. Hence, we assume our post-processing techniques would perform much better had all the missing annotations been annotated. Below we show an example sentence from the AZDC corpus, where the annotators did not annotate "VHL" as a disease:

> *The age incidence curves for renal cell carcinoma and cerebellar haemangioblastoma in **VHL** disease were compatible with a single mutation model, whereas the age incidence curves for sporadic renal cell carcinoma and cerebellar haemangioblastoma suggested a two stage mutation process.*

Regarding the usage of syntactic dependencies as features, our hypothesis was that there are some cue verbs (e.g. "suffer", "cause" and so on) which could provide indicate whether a certain noun is part of a disease name or not. Our initial assumption was that the three types of syntactic dependencies (reported in Table 2.5) could attach the potential disease names to such cue verbs. But, after doing some random analysis, we came to realize that these three dependency types are not enough. For example, in the following sentence:

[17] On a separate set of experiments, we found that if contextual and dictionary lookup features are not considered, then post-processing rules and syntactic dependency features improve the F-score by 0.34 and 0.58 points respectively.

System	Note	P	R	F-score
BANNER	(Leaman et al., 2009)	80.9	75.1	**77.9**
Our system	Using general and orthographic features	74.00	71.14	72.54
Our system	After adding contextual features	81.51	75.99	78.65
Our system	After adding syntactic dependency features	81.62	75.91	78.66
Our system	After adding dictionary lookup features	83.34	78.52	80.87
Our system	After adding post processing	**83.38**	**78.55**	**80.89**

Table 2.8: 10-fold cross validation results using exact matching criteria on AZDC.

Complement C7 deficiency (C7D) is associated frequently with recurrent bacterial infections, especially meningitis caused by Neisseria meningitidis.

"Neisseria meningitidis" is a disease name whose head word is "meningitidis". The word "meningitidis" is syntactically dependent on the cue word "caused" with dependency type "agent". Further investigation is required to select the most relevant dependency types that can be used to unearth links between potential disease names and cue words, which can be used as features.

Errors due to conjunction structure and abbreviated names

One of the sources of errors concerns the annotations containing conjunction structures. There are 94 disease mentions in the data which contain the word "and". The boundaries of 11 of them have been wrongly identified during experiments, while 39 of them have been totally missed out by our system. Our system also has not performed well for disease annotations that have some specific types of prepositional phrase structures. For example, there are 80 disease annotations having the word "of" (e.g. "deficient activity of acid beta-glucosidase GBA"). Only 28 of them are correctly annotated by our system. The major source of errors, however, concerns abbreviated disease names (e.g. "PNH"). We believe that one

way to reduce this specific error type is to generate a list of possible abbreviated disease names from the long forms of disease names available in databases such as UMLS Metathesaurus.

Why features for diseases and other entities (such as genes/proteins) are not the same

Many of the existing BNER systems, which are mainly tuned for gene/protein identification, use features such as token shapes (also known as word classes and brief word classes), Greek alphabet matching, Roman number matching and so forth. As mentioned earlier, we have done extensive experiments with various feature combinations for the selection of disease specific features. We have observed that many of the features used for gene/protein identification are not equally effective for disease identification. Instead, they hurt performance.

This observation is reasonable because gene/protein names are much more complex than entities such as diseases. For example, they often contain punctuation characters (such as parentheses or hyphen), Greek alphabets and digits which are unlikely to appear in disease names. Ideally, the ML algorithm itself should be able to utilize information from only the useful features and ignore the others in the feature set. But practically, including non-informative features often mislead the model learning. In fact, several surveys have argued that the choice of features matter at least as much as the choice of the algorithm if not more (Nadeau and Sekine, 2007; Zweigenbaum et al., 2007).

One of the interesting trends in gene/protein mention identification is to not utilize syntactic dependency relations (with the exception of Vlachos (2007)). Gene/protein names in biomedical literature are often directly combined (i.e. without being separated by space characters) with other characters which do not belong to the corresponding mentions (e.g.

p53-mediated). Moreover, as mentioned before, gene/protein mentions commonly have very complex structures (e.g. *PKR(1-551)K64E/K296R* or *RXRalphaF318A*). So, it is a common practice to tokenize gene/proten names adopting an approach that split tokens as much as possible to extract effective features (Torii et al., 2009; Smith et al., 2008). But while the extensive tokenization boosts performance, it is often difficult to correctly detect dependency relations for the tokens of the gene/protein names in the sentences where they appear. As a result, use of the syntactic dependency relations is not beneficial in such approaches.[18] In comparison, disease mentions are less complex. So, the identified dependencies for disease mentions are more reliable and hence may be usable as potential features.

The above mentioned issues are some of the reasons why a feature set for the well studied gene/protein focused BNER approaches is not necessarily suitable for other biomedical semantic types such as diseases.

If features from automatically created dictionary from the training data is used

Using the proposed global perspective features, which are extracted from automatically created dictionary from the training data (see Table 2.6), did not improve the results. However, we noticed something interesting in the results which are shown in the Figure 2.1.

Regardless of whether we use the external dictionary (created from UMLS; see Section 2.3.2) or not, the usage of the global perspective feature "AlwaysTaggedOther" increased the number of TPs by at least 100. Moreover, even without using the external dictionary, the number of TPs of our system (the 2nd red bar in Figure 2.1), due to "AlwaysTaggedOther" feature, is higher than that of our system (the 1st blue bar in Figure 2.1)

[18]We have done some experiments on Biocreative II GM corpus with syntactic dependency relations of the tokens, and the results support our argument.

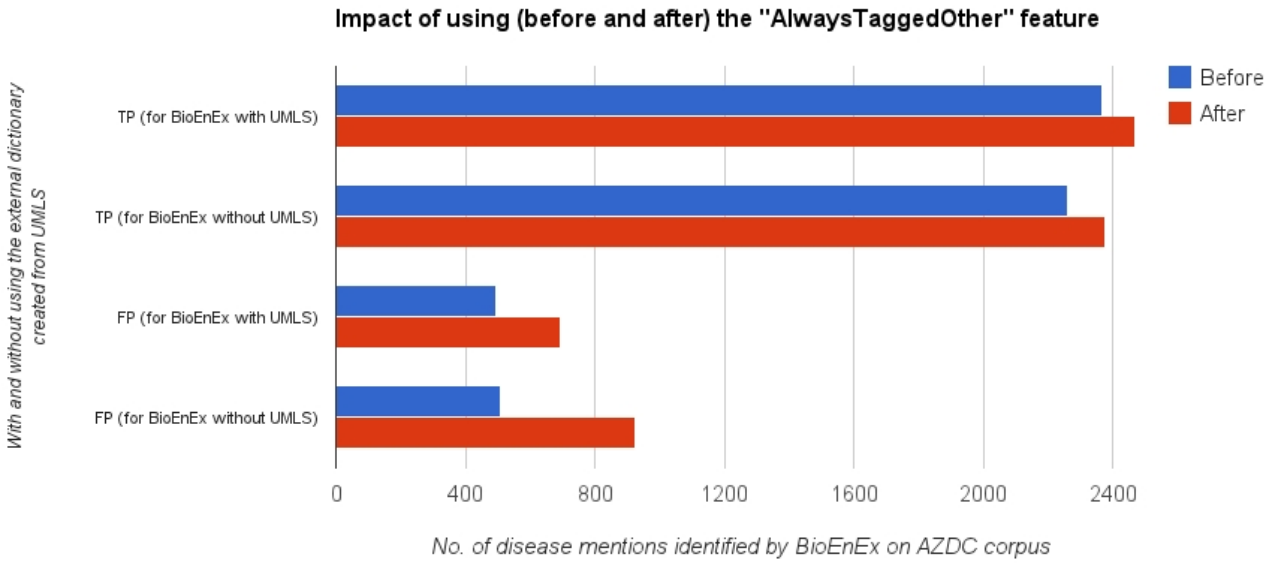


Figure 2.1: Variation in true positives (TPs) and false positives (FPs) due to usage of the global perspective feature "AlwaysTaggedOther".

obtained using the external dictionary (but not the "AlwaysTaggedOther" feature). This observation is important because it indicates that the internal distribution of the non-entity tokens can provide strong clues regarding the tendency of the annotators (or the corresponding guidelines for annotation) for choosing the potential candidate tokens for annotation.

But the problem is that exploiting this distribution as features makes the system too optimistic (because, a corpus cannot contain all possible non-entity tokens), and hence increases FPs. Further investigation is needed to verify whether the proposed automatically created dictionaries can be exploited during post-processing.

It should be noted that in a separate experiment we found the proposed global perspective features quite effective for NER on a benchmark corpus of automatically transcribed broadcast news, a different genre of text. This experiment is discussed in Appendix D.

If only condensed training data is used

Among the 2,783 sentences of AZDC corpus, there are 1,056 sentence which do not contain any disease mention. When we used condensed training data (i.e. only training sentences containing disease mentions) instead full training data, there is a 2 points drop in F-score. Both precision and recall decreased. However, in a later part of this chapter (Section 2.6.4), we will see that this technique has significant impact on performance when the training data does not contain gold annotations.

2.4 From Gold Standard to Silver Standard

The creation of a **gold standard corpus (GSC)** is not only a very laborious task due to the manual effort involved but also a costly and time consuming process. However, the importance of the GSC to effectively train machine learning (ML) systems cannot be underestimated. Researchers have been trying for years to find alternatives or at least some compromise. As a result, self-training, co-training and unsupervised approaches, targeted for specific tasks (such as word sense disambiguation, syntactic parsing, etc), have emerged. In the process of these researches, it became clear that the size of the (manually annotated) training corpus has an impact on the final outcome.

In 2010, the European project CALBC[19] started the development of a large, so called **silver standard corpus (SSC)** using harmonized annotations automatically produced by multiple automatic systems (Rebholz-Schuhmann et al., 2011; Rebholz-Schuhmann et al., 2010a; Rebholz-Schuhmann et al., 2010b). The basic idea is that independent biomedical named entity recognition (BNER) systems annotate a large corpus of biomedical articles without any restriction on the methodology or external resources to be

[19] `http://www.ebi.ac.uk/Rebholz-srv/CALBC/project.html`

exploited. The different annotations are automatically harmonized using some criteria (e.g. minimum number of systems to agree on a certain annotation) to yield a consensus based corpus. This consensus based corpus is called silver standard corpus because, differently from a GSC, it is not created exclusively by human annotators. Several factors can influence the quantity and quality of the annotations during SSC development. These include varying performance, methodology, annotation guidelines and resources of the systems (henceforth **annotation systems**) that would perform the SSC annotation.

The annotation of SSC in the framework of the CALBC project is focused on (bio) entity mentions. However, the idea of SSC creation might also be applied to other types of annotations, e.g. annotation of relations among entities, annotation of treebanks and so on. Hence, if it can be shown that an SSC is a useful resource for the NER task, similar resources can be developed for annotation of information other than entities and utilized for other relevant natural language processing (NLP) tasks.

2.5 BNER using Silver Standard Corpus

We participated in the CALBC I challenge (2010) as part of a wider effort devoted to BNER. One of our motivations was to verify whether our system was robust and portable enough for recognizing other bio-entity types (other than diseases) with high performance.

We used our system, BioEnEx[20], to annotate the following semantic groups: diseases, genes/proteins, species and chemicals. We did not use any external resources such as dictionaries. Understandably, the official training corpus provided for the challenge contains inconsistencies (e.g. incorrect annotations or incorrect boundaries) as it was collaboratively anno-

[20]The earlier version of BioEnEx used GeniaTagger (`http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/`) for tokenization and PoS tagging.

tated by different systems rather than built by human experts. However, we used it (after automatically discarding a few specific types of wrong annotations, e.g. numbers tagged as chemicals) for training our system.[21]

Tables 2.9 shows the official evaluation results of our system on the CALBC SSC I test data, where "*exact*" refers to the exact boundary match and "*cos98*" refers to the relaxed boundary match (i.e. the annotations might differ in uninformative terms such as "the", "a", "acute" etc). Our system obtained the best overall F-score (86.0) for all the 4 entity types among all the participants in this evaluation. More details are available in Rebholz-Schuhmann et al. (2011).

	exact			*cos98*		
	Precision	Recall	F-score	Precision	Recall	F-score
species	91.2%	91.9%	91.6%	92.4%	93.2%	92.8%
diseases	86.0%	87.8%	86.9%	86.5%	88.3%	87.4%
chemicals	82.0%	81.4%	81.7%	82.9%	82.3%	82.6%
genes/proteins	80.0%	79.1%	79.6%	81.7%	80.8%	81.2%

Table 2.9: Official evaluation results of our system on the CALBC SSC I test data.

2.6 Assessing the Practical Usability of SSC

The primary objective of SSC annotation is to compensate the cost, time and manual effort required for a GSC. The procedure of SSC development is inexpensive, fast and yet capable of yielding huge amount of annotated data. These advantages invoke several hypotheses. For example:

- The size of annotated training corpus always plays a crucial role in the performance of ML systems. If the annotation systems have very high

[21]For each of the semantic groups, the system trained separate models. Each of the models was then used to tag mentions of the corresponding semantic group. Finally, all the annotations of different semantic types were combined.

precision and somewhat moderate recall, they would be also able to annotate automatically a large SSC which would have a good quality of annotations. So, one might assume that, even if such an SSC may contain wrong and missing annotations, the larger size of an SSC (15 or 20 times bigger than a GSC) might allow a ML based system to ameliorate the adverse effects of the erroneous annotations.

- Rebholz-Schuhmann et al. (2011) hypothesized that an SSC might serve as an approximation of a GSC.

- In the absence of a GSC, it is expected that ML systems would be able to exploit the harmonised annotations of an SSC to annotate unseen text with reasonable accuracy.

- An SSC could be used to semi-automate the annotation of a GSC. However, in that case, it is expected that the annotation systems would have very high recall. One can assume that converting an SSC into a GSC would be less time consuming and less costly than developing a GSC from scratch.

All these hypotheses are yet to be verified. Once there is an SSC annotated with certain type of information, the main question would be *how this corpus can be maximally exploited* given the fact that it might be created by annotation systems that used different resources and possibly not the same annotation guidelines. This question is directly related to the practical usability of an SSC, which is the focus of this part of the thesis.

Taking the aforementioned hypotheses into account, we wanted to investigate the following research questions which are fundamental to the maximum exploitation of an SSC:

1. How can the annotation quality of an SSC be improved automatically?

2. How would a system trained on an SSC perform if tested on an unseen benchmark GSC?

3. Can an SSC combined with a GSC produce a better trained system?

4. What would be the impact on system performance if *unannotated sentences*[22] are removed from an SSC?

5. What would be the effects of the variation in the size of an SSC on precision and recall?

Our goal is not to judge the procedure of SSC creation, but rather to examine how an SSC can be exploited *automatically* and *maximally* for a specific task. This could provide useful insights to re-evaluate the approach of SSC creation.

2.6.1 Related work with respect to SSC annotation

As mentioned, the concept of SSC has been initiated by the CALBC project (Rebholz-Schuhmann et al., 2010a; Rebholz-Schuhmann et al., 2011). So far, three versions of SSC have been released as part of the project. The CALBC SSC-I has been harmonised from the annotations of the systems provided by the four project partners. Three of them are dictionary based systems while the other is a ML based system. The systems utilized different types of resources such as GENIA corpus (Kim et al., 2003), Entrez Genes[23], Uniprot[24], etc. The CALBC SSC-II and SSC-III corpora have been harmonised from the annotations done by the participants of the 1st and 2nd CALBC challenges and the project partners. Some of the participants have used the CALBC SSC-I and SSC-II versions

[22]For the specific SSC that we use in this work, *unannotated sentences* correspond to those sentences that contain no gene annotation.

[23]`http://jura.wi.mit.edu/entrez_gene/`

[24]`http://www.uniprot.org/`

for training while others used various gene databases or benchmark GSCs such as the BioCreAtIvE II GM corpus.

One of the key questions regarding an SSC would be how close its annotation quality is to a corresponding GSC. On the one hand, every GSC contains its special view of the correct annotation of a given corpus. On the other hand, an SSC is created by systems that might be trained with resources having different annotation standards. So, it is possible that the annotations of an SSC significantly differ with respect to a manually annotated (i.e., gold standard) version of the same corpus. This is because human experts are asked to follow specific annotation guidelines.

Rebholz-Schuhmann and Hahn (2010c) did an intrinsic evaluation of the SSC where they created an SSC and a GSC on a dataset of 3,236 Medline[25] abstracts. They were not able to make any specific conclusion whether the SSC is approaching to the GSC. They were of the opinion that SSC annotations are more similar to terminological resources.

Hahn et al. (2010) proposed a policy where silver standards can be dynamically optimized and customized on demand (given a specific goal function) using a gold standard as an oracle. The gold standard is used for optimization only, not for training for the purpose of SSC annotation. They argued that the nature of diverging tasks to be solved, the levels of specificity to be reached, the sort of guidelines being preferred, ... should allow prospective users of an SSC to customize one on their own and not stick to something that is already prefabricated without a concrete application in mind.

2.6.2 Other approaches similar to SSC annotation

Self-training, co-training and distant supervision are three of the existing approaches that have been used for compensating the lack of a training

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

GSC with adequate size in several different tasks such as word sense disambiguation, semantic role labelling, parsing, relation extraction, etc (Ng and Cardie, 2003; Mihalcea, 2004; McClosky et al., 2006; He and Gildea, 2006; Mintz et al., 2009).

According to Ng and Cardie (2003), self-training is the procedure where a committee of classifiers is trained on the (gold) annotated examples to tag unannotated examples independently. Only those new annotations on which all the classifiers agree are added to the training set and classifiers are retrained. This procedure repeats until a stop condition is met. According to Clark et al. (2003), self-training is a procedure in which "a tagger is retrained on its own labeled cache at each round". In other words, a single classifier is trained on the initially (gold) annotated data and then applied on a set of unannotated data. Those examples meeting a selection criterion are added to the annotated dataset and the classifier is retrained on this new data set. This procedure can continue for several rounds as required.

Co-training is another weakly supervised approach (Blum and Mitchell, 1998). It applies for those tasks where each of the two (or more) sets of features from the initially (gold) annotated training data is sufficient to classify/annotate the unannotated data (Pierce and Cardie, 2001; Mihalcea, 2004; He and Gildea, 2006). As with SSC annotation and self-training, it also attempts to increase the amount of annotated data by making use of unannotated data. The main idea of co-training is to represent the initially annotated data using two (or more) separate feature sets, each called a "view". Then, two (or more) classifiers are trained on those views of the data which are then used to tag new unannotated data. From this newly annotated data, the most confident predictions are added to the previously annotated data. This whole process may continue for several iterations. It should be noted that, by limiting the number of views to one, co-training becomes self-training.

The basic of idea of a distant supervision approach is – given a large database of instances of interest[26] and a large unannotated text corpus, one can use this combination to create a set of positive and negative instances for training a ML based system. Any occurrence of an instance of interest inside the unannotated training data would be automatically considered as a positive instance. It is dubbed as "distant supervision" because the data base provides supervision, but not by direct annotation inside the text data (Mintz et al., 2009).

Like the SSC annotation approach, the multiple classifier approach of self-training and co-training, as described above, adopts the same vision of utilizing automatic systems for producing the annotation. Apart from that, SSC annotation is completely different from both self-training and co-training. For example, classifiers in self-training and co-training utilize the same (manually annotated) resource for their initial training. But SSC annotation systems do not necessarily use the same resource. Both self-training and co-training are weakly supervised approaches where the classifiers are based on supervised ML techniques. In the case of SSC annotation, the annotation systems can be dictionary based or rule based. This attractive flexibility allows SSC annotation to be a completely unsupervised approach since the annotation systems do not necessarily need to be trained.

The distant supervision based approach, as described earlier, can be compared to a single SSC annotation system. But it is not exactly the same as SSC annotation approach, since multiple annotation systems are required to create a harmonized SSC.

[26]For instance, for NER, the database would contain names of target entity type.

2.6.3 Experimental Settings

For our experiments, we use a benchmark GSC called the BioCreAtIvE II GM corpus (Smith et al., 2008) and the CALBC SSC-I corpus (Rebholz-Schuhmann et al., 2010a). Both of these corpora are annotated with genes. The motivation behind the choice of a gene annotated GSC for the SSC evaluation is that ML based BNER for genes has already achieved a sufficient level of maturity. This is not the case for other important bio-entity types, primarily due to the absence of training GSC of adequate size. In fact, for many bio-entity types there exists no GSC. If we can achieve a reasonably good baseline for gene mention identification by maximizing the exploitation of SSC, we might be able to apply almost similar strategies to exploit SSC for other bio-entity types, too.

The training corpus in the BioCreAtIvE II GM corpus has in total 18,265 gene annotations in 15,000 sentences. The test corpus has 6,331 annotations in 5,000 sentences.

Some of the CALBC challenge participants have used the BioCreAtIvE II GM corpus for training to annotate gene/protein in the CALBC SSC-II and SSC-III corpora. We wanted our benchmark corpus and benchmark corpus annotation to be totally unseen by the systems that annotated the SSC to be used in our experiments so that there is no bias in our empirical results. SSC-I satisfies this criterion. So, we use the SSC-I (henceforth, we would refer the CALBC SSC-I as simply the SSC) in our experiments despite the fact that it is smaller than the SSC-II and SSC-III. The SSC has in total 137,610 gene annotations in 316,869 sentences of 50,000 abstracts.

Generally, using a customized dictionary of entity names along with annotated corpus boosts NER performance. However, since our objective is to observe to what extent a ML system can learn from SSC, we avoid the use of any dictionary. We evaluated the performance of BioEnEx on the

BioCreAtIvE II GM test corpus without using any dictionary or lexicon. It achieves comparable results (F-score of 86.22%) to that of the other state-of-the-art systems for gene/protein identification.

One of the complex issues in NER is to come to an agreement regarding the boundaries of entity mentions. Different annotation guidelines have different preferences. There may be tasks where a longer entity mention such as "human IL-7 protein" may be appropriate, while for another task a short one such as "IL-7" is adequate (Hahn et al., 2010).

However, usually evaluation on BNER corpora (e.g., the BioCreAtIvE II GM corpus) is performed adopting exact boundary match. Given that we have used the official evaluation script of the BioCreAtIvE II GM corpus, we have been forced to adopt exact boundary match. Considering a relaxed boundary matching (i.e. the annotations might differ in uninformative terms such as *the*, *a*, *acute*, etc.) rather than exact boundary matching might provide a slightly different picture of the effectiveness of the SSC usage.

2.6.4 Results and analyses

Automatically improving SSC quality for NER

Our hypothesis is that a certain token in the same context can refer to (or be part of) only one entity mention name (i.e. annotation) of a certain semantic group (i.e. entity type). So for overlapping annotations, we kept only the longest ones.[27] After this step, the SSC has 137,604 annotations. We will refer to this version of the SSC as the **initial SSC (ISSC)**.

The next step that we propose for automatically improving annotation quality in the SSC is to remove any word from the SSC which are annotated as genes but cannot constitute a gene name themselves.[28] After this

[27]The CALBC SSC-I corpus has a negligible number of overlapping gene annotations (in fact, only 6).

[28]We construct a list using the lemmatized form of 132 frequently used words that appear in gene

purification step, the total number of annotations is reduced to 133,707. We would refer to this version as the **filtered SSC (FSSC)**.

Then, we use the post-processing module of BioEnEx (please refer to Section 2.2.1), first to further filter out possible wrong gene annotations in the FSSC and then to automatically include potential gene mentions which are not annotated. It has been observed that some of the annotated mentions in the SSC-I span only part of the corresponding token[29]. For example, in the token "IL-2R", only "IL-" is annotated. We extend the post-processing module of BioEnEx to automatically identify all such types of annotations and expand their boundaries when their neighbouring characters are alphanumeric.

Following that, the extended post-processing module of BioEnEx is used to check in every sentence whether there exist potential unannotated mentions[30] which differ from any of the annotated mentions (in the same sentence) by a single character (e.g. "IL-2L" and "IL-2R"), number (e.g. "IL-2R" and "IL-345R") or Greek letter (e.g. "IFN-alpha" and "IFN-beta"). After this step, the total number of gene annotations is 144,375. This means that *we were able to remove/correct some specific types of errors and then further expand the total number of annotations (by including entities not annotated in the original SSC) up to* ***4.92%*** *with respect to the ISSC*. We will refer to this expanded version of the SSC as the **processed SSC (PSSC)**.

When BioEnEx is trained on the above versions of the SSC and tested on the GSC test data, we observed an increase of more than 3% of F-score

names. The words are collected from `http://pir.georgetown.edu/pirwww/iprolink/general_name` and the annotation guideline of GENETAG (Tanabe et al., 2005). These words cannot constitute a gene name themselves. If (the lemmatized form of) all the words in a gene name belong to this list then that gene annotation should be discarded. We use this list to remove erroneous annotations in the ISSC.

[29]By *token* we mean a sequence of consecutive non-whitespace characters.

[30]Any token or sequence of tokens is considered to verify whether it should be annotated or not, if its length is more than 2 characters excluding digits and Greek letters.

because of the filtering and expansion (see Table 2.10). One noticeable characteristic in the results is that the number of annotations obtained (i.e. TP+FP[31]) by training on any of the versions of the SSC is almost half of the actual number annotations of the GSC test data. This has resulted in a low recall. There could be mainly two reasons behind this outcome:

- First of all, it might be the case that a considerable number of gene names are not annotated inside the SSC versions. As a result, the features shared by the annotated gene names (i.e. TP) and unnannotated gene names (i.e. FN) might not have enough influence.
- There might be a considerable number of wrong annotations which are actually not genes (i.e. FP). Consequently, a number of bad features might be collected from those wrong annotations which are misleading the training process.

To verify the above conditions, it would require manual annotation of the large CALBC SSC. This is not feasible because of the cost of human labour and time. Nevertheless, we can try to measure the state of the above conditions roughly by using only *annotated sentences* (i.e. sentences containing at least one annotation) and varying the size of the corpus, which are the subjects of our next experiments.

Impact of annotated sentences and different sizes of the SSC

We observe that only 77,117 out of the 316,869 sentences in the PSSC contain gene annotations. We will refer to the sentences having at least one gene annotation collectively as the **condensed SSC (CSSC)**. Table

[31]TP (true positive) = corresponding annotation done by the system is correct, FP (false positive) = corresponding annotation done by the system is incorrect, FN (false negative) = corresponding annotation is correct but it is not annotated by the system.

	TP	FP	FN	P	R	F-score
ISSC	2,396	594	3,935	80.13	37.85	51.41
FSSC	2,518	557	3,813	81.89	39.77	53.54
PSSC	2,606	631	3,725	80.51	41.16	54.47

Table 2.10: The results of experiments when trained with different versions of the SSC and tested on the GSC test data.

	Total tokens in the corpus	Annotated total genes	TP	FP	FN	P	R	F-score
PSSC	6,955,662	144,375	2,606	631	3,725	80.51	41.16	54.47
100% of CSSC	1,983,113	144,375	3,401	1,161	2,930	74.55	53.72	62.44
75% of CSSC	1,487,823	108,213	3,421	1,070	2,910	76.17	54.04	63.22
50% of CSSC	992,392	72,316	3,265	1,095	3,066	74.89	51.57	61.08
25% of CSSC	494,249	35,984	3,179	1,048	3,152	75.21	50.21	60.22
10% of CSSC	196,522	14,189	2,988	1,097	3,343	73.15	47.20	57.37

Table 2.11: The results of SSC experiments with varying size of the CSSC = condensed SSC (i.e. sentences containing at least one annotation). SSC size = 316,869 sentences. CSSC size = 77,117.

2.11 and Figure 2.2 show the results when we used different portions of the CSSC for training.

There are four immediate observations on the above results:

- Using the full PSSC, we obtain total (i.e. TP+FP) 3,237 annotations on the GSC test data. But when we use only the annotated sentences of the PSSC (i.e. the CSSC), the total number of annotations is 4,562, i.e. there is an increment of 40.93%.

- Although we have a boost in F-score due to the increase in recall using the CSSC in place of the PSSC, there is a considerable drop in precision.

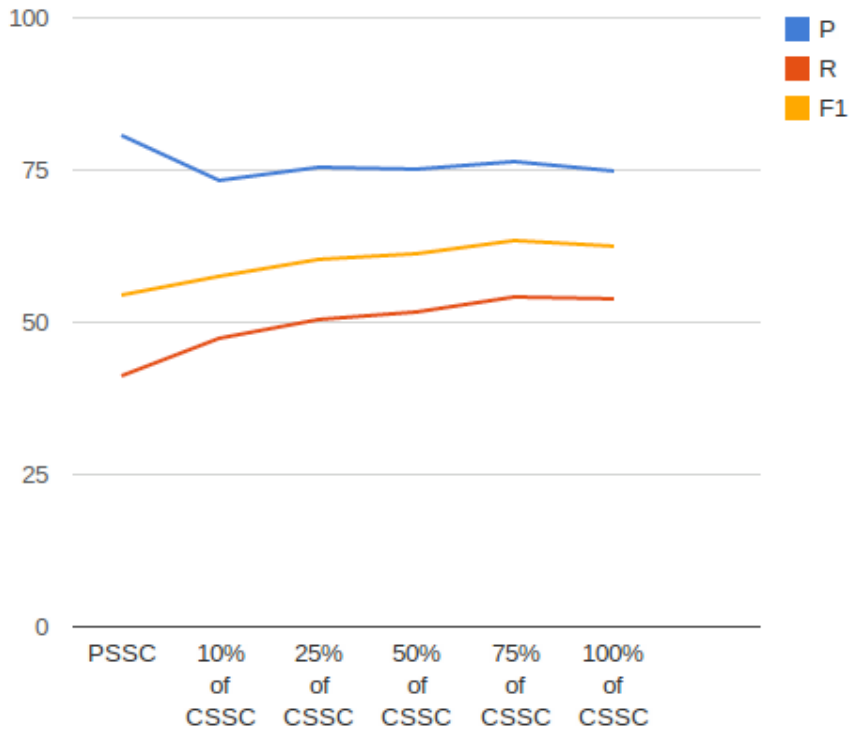


Figure 2.2: Graphical representation of the experimental results with varying size of the CSSC.

- The number of FP is almost the same for the usage of 10-75% of the CSSC.
- The number of FN kept decreasing (and TP kept increasing) for 10-75% of the CSSC.

These observations can be interpreted as follows:

- Unannotated sentences inside the SSC in reality contain many gene annotations; so the inclusion of such sentences misleads the training process of the ML system.
- Some of the unannotated sentences actually do not contain any gene names, while others would contain such names but the automatic annotations missed them. As a consequence, the former sentences contain true negative examples which could provide useful features that

can be exploited during training so that less FPs are produced (with a precision drop using the CSSC). So, instead of simply discarding all the unannotated sentences, we could adopt a filtering strategy that tries to distinguish between the two classes of sentences above.

- The experimental results with the increasing size of the CSSC show a decrease in both precision (74.55 vs 76.17) and recall (53.72 vs 54.04). We plan to run again these experiments with different randomized splits to better assess the performance.

- Even using only 10% of the whole CSSC does not produce a drastic difference with the results when the full CSSC is used. This indicates that perhaps the more CSSC data is fed, the more the system tends to overfit.

- It is evident that the more the size of the CSSC increases, the lower the improvement of F-score, if the total number of annotations in the newly added sentences and the accuracy of the annotations are not considerably higher. It might be not surprising if, after the addition of more sentences in the CSSC, the F-score drops further rather than increasing. The assumption that having a large SSC would be beneficiary might not be completely correct. There might be some optimal limit of the SSC (depending on the task) that can provide maximum benefits.

Training with the GSC and the SSC together

Our final experiments were focused on whether it is possible to improve performance by simply merging the GSC training data with the PSSC and the CSSC. The PSSC has almost 24 times the number of sentences and almost 8 times the number of gene annotations than the GSC. There is a

	TP	FP	FN	P	R	F-score
GSC	5,373	759	958	87.62	84.87	86.22
PSSC + GSC	3,745	634	2,586	85.52	59.15	69.93
PSSC + GSC * 8	4,163	606	2,168	87.29	65.76	75.01
CSSC + GSC * 8	4,507	814	1,824	84.70	71.19	77.36

Table 2.12: The results of experiments by training on the GSC training data merged with the PSSC and the CSSC.

possibility that, when we do a simple merge, the weight of the gold annotations would be underestimated. So, apart from doing a simple merge, we also try to balance the annotations of the two corpora. There are two options to do this – (i) by duplicating the GSC training corpus 8 times to make its total number of annotations equal to that of the PSSC, or (ii) by choosing randomly a portion of the PSSC that would have almost similar amount of annotations as that of the GSC. We choose the 1st option.

Unfortunately, when an SSC (i.e. the PSSC or the CSSC) is combined with the GSC, the result is far below than that of using the GSC only (see Table 2.12). Again, low recall is the main issue partly due to the lower number of annotations (i.e. TP+FP) done by the system trained on an SSC and the GSC instead of the GSC only. As we know, a GSC is manually annotated following precise guidelines, while an SSC is annotated with automatic systems that do not necessarily follow the same guidelines as a GSC. So, it would not have been surprising if the number of annotations were high (since we have much bigger training corpus due to SSC) but precision were low. But in practice, precision obtained by combining an SSC and the GSC is almost as high as the precision achieved using the GSC.

One reason for the lower number of annotations might be the errors that have been propagated inside the SSC. Some of the systems that have been

used for the annotation of the SSC might have low recall. As a result, during harmonization of their annotations several valid gene mentions might not have been included[32].

One other possible reason could be the difference in the entity name boundaries in the GSC and an SSC. We have checked some of the SSC annotations randomly. It appears that in those annotated entity names some relevant (neighbouring) words (in the corresponding sentences) are not included. It is most likely that the SSC annotation systems had disagreements on those words.

When the annotations of the GSC were given higher preference (by duplicating), there is a substantial improvement in the F-score, although still lower than the result with the GSC only.

2.6.5 Summary of the SSC experimental study

The idea of SSC development is simple and yet attractive. Obtaining better results on a test dataset by combining output of multiple (accurate and diverse[33]) systems is not new (Torii et al., 2009; Smith et al., 2008). But adopting this strategy for corpus development is a novel and unconventional approach. Some natural language processing tasks (especially the new ones) lack adequate GSCs to be used for the training of ML based systems. For such tasks, domain experts can provide patterns or rules to build systems that can be used to annotate an initial version of SSC. Such systems might lack high recall but are expected to have high precision. Already available task specific lexicons or dictionaries can also be utilized for

[32]There can be two reasons for this: (i) when a certain valid gene name is not annotated by any of the annotation systems, and (ii) when only a few of those systems have annotated the valid name but the total number of such systems is below than the minimum required number of agreements, and hence the gene name is not considered as an SSC annotation.

[33]A system is said to be accurate if its classification performance is better than a random classification. Two systems are considered diverse if they do not make the same classification mistakes. (Torii et al., 2009)

SSC annotation. Such an initial version of SSC can be later enriched using automatic process which would utilize existing annotations in the SSC.

With this vision in mind, we pose ourselves several questions (see Section 2.6) regarding the practical usability and exploitation of an SSC. In the search of answers, we accumulate several important empirical observations. We have been able to automatically reduce the number of erroneous annotations from the SSC and include unannotated potential entity mentions simply using the annotations that the SSC already provides. Our techniques have been effective for improving the annotation quality as there is a considerable increment of F-score (almost 11% higher when we use CSSC instead of using ISSC; see Tables 2.10 and 2.11).

We also observe that it is possible to obtain more than 80% of precision using the SSC. But recall remains quite low, partly due to the low number of annotations provided by the system trained with the SSC. Perhaps, the entity names in the SSC that are missed by the annotation systems are one of the reasons for that.

Perhaps, the most interesting outcome of this study is that, if only annotated sentences (which we call *condensed* corpus[34]) are considered, then the number of annotations as well as the performance increases significantly. This indicates that many unannotated sentences contain annotations missed by the automatic annotation systems. However, it appears that correctly unannotated sentences influence the achievement of high precision. Future investigation should adopt a more sophisticated approach instead of completely discarding the unannotated sentences, e.g. devising a filter able to distinguish between relevant unannotated sentences (i.e., those that should contain annotations) from non-relevant ones (i.e., those that correctly do not contain any annotation). Measuring lexical similarity between annotated and unannotated sentences might help in this case.

[34]The proposed idea of condensed corpus might be helpful for other NLP tasks as well.

We notice that the size of an SSC affects performance, but increasing it above a certain limit does not always guarantee an improvement of performance (see Figure 2.2). This rejects the hypothesis that having a much larger SSC should allow a ML based system to ameliorate the effect of having erroneous annotations inside the SSC.

Our empirical results show that combining GSC and SSC does not improve results for the particular task of NER, even if GSC annotations are given higher weights (through duplication). We assume that this is partly due to the variations in the guidelines of entity name boundaries[35]. These impact the learning of the ML algorithm. For other NLP tasks where the possible outcome is boolean (e.g. relation extraction, i.e. whether a particular relation holds between two entities or not), we speculate that the results of such combination might be better.

In short, our findings suggest that an automatically pre-processed SSC might already contain annotations with reasonable quality and quantity, since using it we are able to reach more than 62% of F-score. This is encouraging since in the absence of a GSC, a ML system would be able to exploit an SSC to annotate unseen text with a moderate (if not high) accuracy. Hence, SSC development might be a good option to semi-automate the annotation of a GSC.

[35]For example, "human IL-7 protein" vs "IL-7".

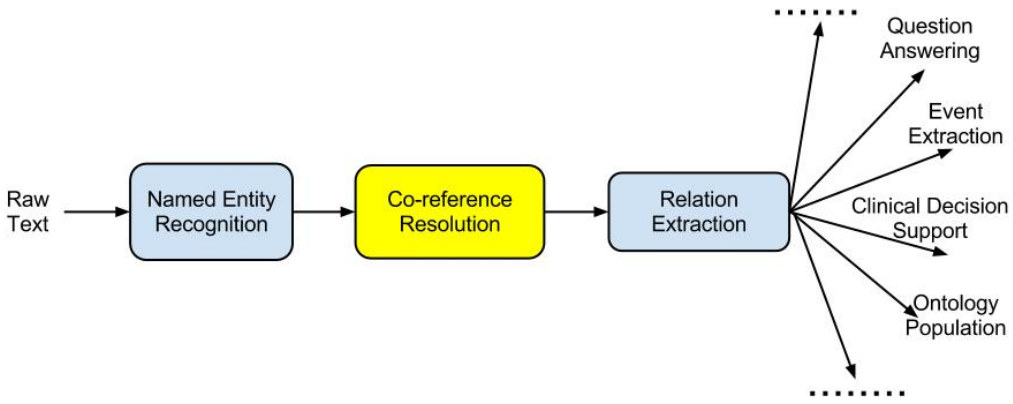
Chapter 3

Coreference Resolution

"Nobody is interested in coreference for its own sake; however it has the potential of enabling higher performance on extraction tasks that are of interest."

– Douglas E. Appelt

"Introduction to Information Extraction", AI Communications, 12(3):161–172. (1999)



Coreference resolution for named entity (NE) mentions is the task of identifying whether a mention refers to the one or more other mentions in a given document in such a way that these mentions denote the same real world entity. The mentions can be either named, nominal or pronominal.

As an example, consider the sentence shown in Figure 3.1. A coreference resolution system is expected to find inside this sentence that both of the two *"her"* pronouns are referring to *"The patient"*, a person, while *"which"* is referring to the problem/medical condition *"Labetalol"*.

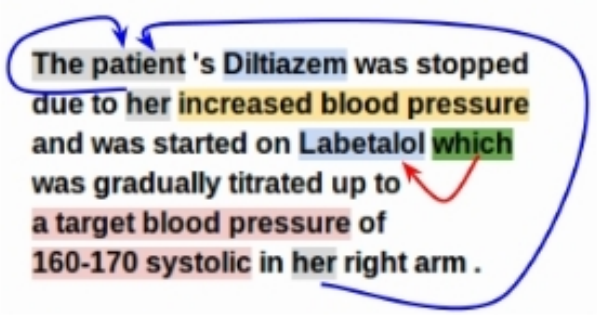


Figure 3.1: An example of co-referring mentions.

This task is important for other natural language processing (NLP) tasks. For example, Stevenson (2006) analysed three IE evaluation corpora from the Message Understanding Conferences and showed that a majority of the cross-sentential relations among entity mentions are due to coreference.

Although significant progress has been made, the problem of coreference resolution is far from being solved for different reasons: partly because of the confusion over different evaluation metrics and partly because the majority of the research done so far are focused on some limited NE types on the newspaper text. Hence, these well-researched existing methodologies with the existing traditional features do not perform as well on new domains such as clinical records. In this chapter, we propose a variant of the influential mention-pair model for coreference resolution. Unlike other related work, when tested on a benchmark clinical corpus it obtains good scores for each of the individual metrics rather than being biased towards a particular metric.[1]

[1]The work described in this chapter was carried out in the context of a research stay by the author of

While all the other studies (regarding NER and RE) reported in this thesis are experimented on biomedical (scientific article) texts, the data used in this chapter are from clinical texts (which belong to a different genre with respect to biomedical texts, even if somehow related). We used the i2b2/VA 2011 challenge corpus for the experiments of this chapter. The corpus was constructed from clinical discharge summaries. Originally, our goal was to extend this work by adapting the approach for biomedical text to create a complete information extraction pipeline. However, we are still working to achieve this goal. Nevertheless, we are confident that some of the key ideas in the proposed approach can be effective for biomedical (or even for newswire) text as well (more details at the end of this chapter). In a broader picture, knowledge embedded in clinical and biomedical texts is complementary to other for evidence based medicine and other related applications. The entity types in these two genres are almost identical and belong to the same hierarchy which is covered by resources such as UMLS. Hence, we believe it worth to conduct this study as part of this thesis.

3.1 Background

As mention above, a significant body of studies has been conducted for coreference resolution in the newswire domain.[2] Recently, applications to the biomedical domain have started to emerge. Until BioNLP shared task 2011 (Nguyen et al., 2011), most of these work was done in a limited context - either using very small datasets or focusing only on pronominal anaphora resolution (Castano et al., 2002; Lin and Liang, 2004; Liang and Lin, 2005; Gasperin and Briscoe, 2008). In the last years, a few studies have been also reported on clinical text (more details in Section 3.2.3), particularly

this PhD thesis in LIMSI-CNRS. It was conducted with the approval of i2b2 and the VA, and under the guidance of Dr. Pierre Zweigenbaum.

[2] An overview of these work can be found in Ng (2010).

in the context of the i2b2/VA 2011 challenge.

The general IE performance measures such as precision, recall and F-score are also used for coreference resolution. But unlike NER and RE (which use true positives, false positives and true negatives), these measures depends on other metrics. The three main metrics are MUC, B^3 and $CEAF$.[3]

A general trend in recent studies is to evaluate a coreference resolver against unweighted average of the scores obtained using multiple of such metrics. This approach has the advantage to indirectly counteract the bias inherent in a particular metric. But such an average risks to give a false impression about the weakness and/or strength of a resolver. As Uzuner et al. (2012) stated, a system that predicts no coreference chains could still achieve an unweighted average F_1 score of 0.541 on the i2b2/VA corpus.

To better understand the current state of the coreference resolution on clinical text, let us consider the results reported by the best system of the i2b2/VA 2011 challenge. It obtained 0.915 as an unweighted average F_1 score for all the four clinical chain types (namely `Person`, `Problem`, `Test` and `Treatment`) of the challenge (Xu et al., 2012). This is a very high score and at a first glance it might appear that the problem of coreference resolution is almost worked out. However, a closer look reveals that the same system achieves only as much as 0.489 MUC score for one of the entity types (`Test`). This highlights that the problem is far from resolved and needs further investigation.

[3]Details about the metrics can be found in Uzuner et al. (2012).

3.2 Related Work

3.2.1 Supervised Coreference Resolution

As opined by Ng (2010), the *mention-pair* model (Aone and Bennett, 1995; McCarthy and Lehnert, 1995) is perhaps the most widely studied and influential learning-based coreference resolution approach. It works as follows. At first, the classifier identifies all the pairs of mentions (i.e. noun phrases) which are co-referent. Once these pairs are identified, they are grouped into co-referent chains by some clustering techniques.

One alternative solution proposed for *mention-pair* model is the *mention-ranking* model (Denis and Baldridge, 2008) which imposes a ranking on all the candidate antecedents of a mention (henceforth, active mention) to determine which candidate antecedent is most probable.

A more complex approach is the *entity-mention* model (Luo et al., 2004; Yang et al., 2004) where a model is trained to determine whether an active mention belongs to a preceding, possibly partially-formed, coreference cluster. The *cluster-ranking* approach is a combination of the mention-ranking and entity-mention models (Rahman and Ng, 2009). To the best of our knowledge, the latter three approaches have been applied on newspaper articles only.

While all the latter three approaches are arguably more natural reformulations of the coreference resolution problem, the relative simplicity of the mention-pair model makes it easier to implement and, perhaps, a more appropriate choice for clinical texts which are quite different from newspaper text and even from biomedical literature text.

3.2.2 Characteristics of Clinical Texts

Meystre et al. (2008) listed a number of peculiarities of clinical texts. For example, sentences in clinical text can be surprisingly short or quite long.

Since they are written mainly for documentation purposes, they are sometimes ungrammatical and composed of short, telegraphic phrases. Moreover, the usage of abbreviations, acronyms, and local dialectal shorthand phrases is quite common. The acronyms are often overloaded (i.e., the same set of letters has multiple expansions) and are highly ambiguous even in the context (Liu et al., 2001).

Misspellings also abound in clinical text. In addition, terms often lack syntactic cues such as definite articles. Furthermore, in order to save time, clinicians often reuse previous notes when writing a new one. Wrenn et al. (2010) showed that on average 78% of the text of discharge notes and 54% of the text in progress notes are copied from previous notes.

Unlike newspaper text, in a clinical text there is only one dominant large chain of *person* (the patient) (Grouin et al., 2011; Bodnari et al., 2012). The other *person* entities are either singletons (i.e. do not have any co-referent) or part of small chains. So, the identification of chains of co-referent *person* mentions in clinical texts is relatively easier.

However, the situation is a bit complicated for other types of mentions because too much emphasis on traditional features (such as string similarities) could end up producing a very low performing system. The same concepts in different spatio-temporal contexts are not necessarily co-referents[4]. Two mentions of the same `Treatment` which have different administration modes (e.g. orally, intravenously, etc) or two `Test` mentions having different results are also non-co-referents. All the above are some of the many peculiarities that make coreference resolution in clinical text different from that in other text genres and hence require a tailored solution.

[4]For example, *Pain* in the head is not co-referential to *pain* in the leg. Also, a `Treatment/Test/Problem` of Monday 4 p.m. and a similar mention of Friday 6 p.m. are not necessarily co-referents.

3.2.3 Coreference Resolution on Clinical Text

One of the first studies for coreference resolution on clinical text that we are aware of was proposed by He (2007). The author proposed a coreference resolver for discharge summaries using a supervised decision-tree classifier and a carefully selected set of features. Zheng et al. (2011) and Zheng et al. (2012) conducted a comprehensive review of various coreference resolution methodologies previously used on other genres of text. They tested these methodologies on the clinical domain. Bodnari et al. (2012) also did a similar study. These studies conclude that coreference resolvers developed for other genres of text perform poorly on clinical text. All these studies are based on the mention-pair model.

The i2b2/VA 2011 challenge was an attempt to push the research on coreference resolution in the clinical domain. A number of teams participated in the challenge and their approaches ranged from rule based and hybrid systems to supervised approaches. A detailed description of the approaches adopted by the participating teams can be found in Uzuner et al. (2012). The best system of the challenge was a supervised approach based on a mention-pair model that uses three different classifiers and a number of domain-specific resources. The system obtained extremely high B^3 scores for all the entities (i.e. it is very good at identifying singletons) as well as very high scores for `Person` type chains. But the results are not as good for MUC scores, especially for the `Test` type.

3.3 Our Proposed Approach

We propose a variant of the influential mention-pair model. The key idea of our approach is to control different phases of the supervised resolution process and achieve the following characteristics:

- Get rid of as many less-informative/sub-optimal *training* instances as

possible before beginning to train a model;

- Discard as many negative *test* instances (i.e. non-co-referent pairs of mentions) as possible even before applying the trained model;
- Form co-referent chains from the pairs based on local and global perspective as well as by considering the confidence scores predicted by the classifier.

The problem caused by the imbalance of negative and positive annotated training instances is a well known issue in machine learning (ML) research. Previous studies had empirically shown that unbalanced datasets lead to poor performance for the minority (i.e. positive) class (Weiss and Provost, 2001). Keeping that in mind, the first characteristic above puts forward the early use of expert-based knowledge to reduce such skewness in the training data, while the second characteristic suggests to use as much expert knowledge as possible to solve part of the problem with high accuracy/reliability.

3.3.1 Traditional Approach for Instance Creation

As previously mentioned, a mention-pair model is composed of a classifier (to identify co-referent pairs) and a clustering approach (to form chains from the identified pairs). For creating training instances for the classifier, arguably the most popular choice is the approach proposed by Soon et al. (1999) and Soon et al. (2001) which is the following.

Given an anaphoric noun phrase, NP_k, create a positive instance between NP_k and its closest preceding antecedent, NP_j , and a negative instance by pairing NP_k with each of the intervening NPs, NP_{j+1} , . . ., NP_{k-1}.

To improve the precision of the coreference resolver, Ng and Cardie (2002) proposed the following simple modification of the method described

earlier.

If NP_k is non-pronominal, a positive instance should be formed between NP_k and its closest preceding non-pronominal antecedent instead.

To further reduce skewness between positive and negative instances, some other studies employ a filtering mechanism by disallowing the creation of training instances from NP (i.e. mention) pairs that are unlikely to be co-referents, e.g. the NP pairs that violate gender and number agreement (e.g. Strube et al. (2002), Yang et al. (2003)) or agreement in the semantic types of the mentions.

We exploit these strategies but strengthen them further, which will be discussed next.

3.3.2 The Criteria for being Co-referent: A Proposal

The success of a ML technique often relies on the effective prior assumptions which can allow the learner to have a rational bias to classify unseen instances with higher accuracy. For coreference resolution, such assumptions include, among others, how to discard as many as possible less-informative negative (i.e. false) antecedent-anaphora candidate pairs in advance based on expert knowledge (rather than by using a ML technique exploiting another set of features).

We created a list of 13 criteria (driven by semantic and syntactic intuitions) and propose that a candidate antecedent (m_x) and the mention to be resolved (m_y) are unlikely to be co-referents if they violate any of them. We list some of those criteria with examples below (see Appendix C for the complete list):

- m_y is a determiner and part of an NP rather than constituting an NP itself:
  - e.g. if the word *"this"* is the mention m_y but it is part of a larger

mention (m_x) *"this patient"* then they are not co-referents.

- m_y is a determiner (or Wh-determiner) and it is not among the first two words of the sentence and m_x does not belong to the same sentence:
  - e.g. consider the following consecutive sentences (extracted from clinical text) and a candidate pair (which are not co-referent) where *"hematemesis"* in the first sentence is m_x and *"which"* in the second sentence is m_y

    Sen. 1: *Mr. Bruno is a 60 year old gentleman who initially presented with* ***hematemesis****, hemoptysis and on work-up was found to have a left lower lobe mass .*

    Sen. 2: *He previously underwent bronchoscopy with washings* ***which*** *showed to be negative for malignant cells and showed atypical bronchial epithelial cells , likely to be reactive .*

- both m_x and m_y are of type `Problem` but they are semantically attached to different persons:
  - e.g. "*Diabetics of the patient*" and "*Diabetics of the patient's father*" do not refer to the same entity.

Some of these 13 criteria ensure that there is no mismatch in semantic types of the mentions as well as in grammatical characteristics (such as gender and number types). Our system exploits these criteria during training and test instance creation to reduce the negative instances. For training instances, the chains are converted to sets of pairs of mentions (this will be discussed later) before applying the criteria.

It turns out that, although this strategy ensures filtering of a significant amount of negative instances (more in Section 3.6.1), there are also a small number of positive instances, i.e. positive pairs, which are discarded in the process, partly because of peculiarity of clinical text and partly because

of annotation errors in the data. However, we observed that in such cases almost all those corresponding chains could be still reconstructed from the remaining positive pairs.

3.3.3 Our approach for Training Instance Creation

For positive instance creation, we follow the same techniques used by Ng and Cardie (Ng and Cardie, 2002) and Soon et al. (2001) as mentioned in 3.3.1. But for the creation of negative training instances, we add some additional constraints. These are:

- a pair must not match any of the criteria for being unlikely to be co-referents (Section 3.3.2);
- the difference between the sentence indexes of the mentions must not be more than 5, i.e. the distance between a pair of mentions must not be more than 5 sentences.

The first constraint is motivated by syntactic and semantic properties, while the second constraint is influenced by the contextual properties. We observed that the majority of the co-referent pairs are within the 5 sentence boundary. Even when pairs are not within this boundary, they are often part of the chains where other co-referent mentions are in between them. So, the sentence boundary allows to reduce skewness between positive and negative instances significantly at a cost of the exclusion of only a limited number of positive instances.

3.3.4 Our approach for Test Instance Creation

Unlike previous studies, instead of creating test instances simply with every pair of entities of compatible semantic types, we enforce a number of restrictions.

We propose that any mention to be resolved (m_y) and any candidate antecedent (m_x) (which precedes m_y) could be used to create a test instance only if *none* of the following situations holds:

1. the pair matches any of the criteria for being unlikely to be co-referents (Section 3.3.2);

2. the difference between the sentence indexes of the mentions is more than 5, and either m_y is a determiner or the two mentions do not have exact string similarity or the first word of m_y is a pronoun;

3. m_y is a pronoun whose gender is male/female and already 3 other candidate test instances for m_y are created where these 3 other candidate antecedents appear after m_x but before m_y inside the text.

Sometimes mentions having exactly the same names, although they appear in different parts of clinical text, have a high probability of corresponding to the same entity. We did not want to exclude such pairs of mentions from consideration simply because they lie beyond the sentence boundary. This is why we added additional constraints in the 2nd constraint mentioned above.

The last constraint listed above (which limits the number of candidate antecedents up to 3 for the male/female pronouns) is motivated by our random analyses of the training data (a similar preference for local pronoun coreference is also enforced in other studies, e.g. Haghighi and Klein (2007)). We observed that for a male/female pronoun (e.g. *he*), if we list all the compatible entity mentions (i.e. those which satisfies other constraints mentioned above) according to their order of appearance (in the text) prior to the pronoun in question, at least one of the true antecedents (if any) of the pronoun could be located among the immediately preceding three compatible entity mentions.

3.3.5 ML Technique Chosen for Classification and Data Preprocessing

Ideally, any ML classifier can be accommodated in our proposed approach for training models and classifying test instances as co-referent and non-co-referent pairs. For this particular study, we used support vector machine (SVM). The features of the system are explicitly extracted and then used for training by the SVM-Light-TK toolkit (Moschitti, 2006; Joachims, 1999).

We used the Stanford parser (Klein and Manning, 2003) for parsing the texts.

One of the important characteristics of our approach is that we train only one SVM classifier for all the 4 semantic types of the i2b2/VA 2011 data. This is unlike the other approaches where different classifiers are used for `Person` and non-`Person` mention pairs and even for identifying *patient* and non-*patient* type mentions. Our objective was to use a simpler classifier and focus more on analysing the impact of the different contributions that we made in this study. However, one can easily adopt multiple classifiers instead of a single classifier in our approach without any loss of generality.

3.3.6 Traditional Approach Clustering Mentions into Chains

There exist several proposals about how to form co-referent chains from co-referent pairs (see Ng (Ng, 2010) for a review of different proposals). Among them, the most popular algorithms are closest-first clustering (Soon et al., 2001) and best-first clustering (Ng and Cardie, 2002).

If a mention m_y has multiple possible antecedents (identified during the classification stage), the closest-first clustering algorithm selects the closest preceding antecedent among them as co-referent. To improve the precision of the closest-first algorithm, the best-first clustering algorithm

selects that mention as co-referent which is the most probable preceding mention among the multiple possible antecedents.

3.3.7 Our Proposed Approach for Clustering Mentions into Chains

With an eye to improving both recall and precision, we propose a modified version of the closest-first and best-first clustering algorithms.

From the pairs identified as likely to be co-referents by the classifier, we form chains by the following steps based on two thresholds (see Table 3.1):

1. Discard all the likely co-referent pairs of mentions for which the scores predicted by the classifier are smaller than the *MINIMUM THRESHOLD* value.

2. Among the pairs obtained from the previous step, for each mention to be resolved (m_y) retain any candidate antecedent (m_x) if either of the following holds:

   (a) m_x is the closest mention among the possible antecedents;

   (b) the score predicted for the corresponding pair $\{m_x, m_y\}$ is greater than or equal to the *MAXIMUM THRESHOLD* value.

3. New pairs are added to the list (refined by previous steps), even if the classifier does not consider them as possible co-referents, if either of the following holds:

   (a) m_y is a cataphora for a certain mention m_x (this will be discussed in Section 3.3.8);

   (b) both m_y and m_x match either with the string *"the patient"* or *"patient"*.

4. Group any two pairs of the above list into a chain if they share a common mention.

Threshold	Action	Value
	accept all	
MAX	———————	4
	accept if special case	
MIN	———————	0.9
	exclude all	

Table 3.1: Two thresholds to cluster mention pairs into chains.

5. Merge two chains (obtained from the previous step) into a single chain if they are of type `Person` and the total number mentions in each of them is more than *MIN NUMBER OF MERGE THRESHOLD* value.

Some of the rules above are motivated by the fact that there is usually only one large `Person` chain (the one of the patient) in a clinical document. Readers are referred to the various analyses of the i2b2/VA corpus reported by Grouin et al. (2011) for more details.

The values of *MINIMUM THRESHOLD*, *MAXIMUM THRESHOLD* and *MIN NUMBER OF MERGE THRESHOLD* are 0.9, 4.0 and 7 respectively, and they are selected empirically from the experiments on the training data.

3.3.8 Cataphora Resolution

In linguistics, cataphora is used to describe an expression that co-refers with a later expression in the discourse. For example, consider the following sentence where *"your Primary care doctor"* is a cataphora that refers to *"Larry Bock"*.

Please follow-up with your Primary care doctor Larry Bock 2019-01-16 at 8:30 AM.

Note that, in the above example *"Larry Bock"* is in apposition to *"your Primary care doctor"*. [5]

Sometimes clinical texts contain some positive cataphora instances which we tried to deal with separately in a limited extent. We only try to resolve a cataphora if it is in apposition with another mention. For any two mentions m_x and m_y, we consider m_y as a cataphora of m_x if

- m_x appear after m_y in the same sentence,
- m_y consists of more than one word,
- the first word of m_y is a pronoun or determiner,
- the first word of m_x is not a pronoun or determiner,
- both m_x and m_y are of type `Person`, and
- the last word of m_y and the first word of m_x are consecutive words or there is a comma (,) between them.

3.4 Data

The i2b2/VA corpus of the i2b2/VA 2011 challenge contains de-identified discharge summaries from Beth Israel Deaconess Medical Center, Partners Healthcare, and University of Pittsburgh Medical Center (UPMC). Details about the corpus can be found in Uzuner et al. (2012).

It was noted by the challenge organisers in the task description paper that some of the participating teams could not obtain the UPMC data. So, the organisers provided two rankings – one for all the participating teams (without evaluation on the UPMC data) and the other for the teams who

[5] Had the sentence been written as *"Please follow-up with Larry Bock, your Primary care doctor, 2019-01-16 at 8:30 AM"*, then the two mentions would be still in apposition, but in that case *"your Primary care doctor"* would be an anaphora and *not* a cataphora.

were able to use the full corpus. To help the reader compare with any of the participating teams, we provide results with/without the UPMC data.

To provide an evaluation as similar as possible to the official evaluation of the challenge, we put the official test corpus aside and used it as unseen data for the final evaluation. Initially, we excluded the UPMC data from the official training corpus, and split the documents inside it into 66/33 % proportions which were used as the development training corpus (henceforth, *dev-train*) and development test corpus (henceforth, *dev-test*). All the experiments for feature selection and parameter tuning were conducted on these dev-train and dev-test corpora.

3.5 Feature Selection and Extraction

Tables 3.2, 3.3, 3.4 and 3.5 list all the features (with description) that we used in our system. These features are grouped into four categories – LEXICAL, SEMANTIC, GRAMMATICAL and CONTEXTUAL. All these features are selected because of their impact on improving the overall results (i.e. for all metrics – MUC, $CEAF$ and B_3) during the experiments based on dev-train and dev-test corpora. Features with (*) mark inside Tables 3.2, 3.3, 3.4 and 3.5 indicate new feature types proposed by us.

We heavily used POS tags and various linguistic properties for feature construction. A few lists were constructed by analysing text of the dev-train corpus as well as exploiting the UMLS Metathesaurus (Bodenreider, 2004) and some Wikipedia[6] articles related to the human organs and drug administration. These lists include position/size cues (e.g. *small, left*, etc), frequency/quantity cues (e.g. *daily, per week, q.o.d, t.i.d, mg, ml*, etc which must be followed by a number), physical location cues (e.g. *heart, lung, esophagus*, etc), and drug administration mode cues (e.g. *oral, nebuliza-*

[6]www.wikipedia.com

SEMANTIC AND GRAMMATICAL **features describing a candidate antecedent (m_x) and the mention to be resolved (m_y)**	
Feature	*Description*
Feature type: Semantic	
`HasCommonUmlsCUI`	If lists of UMLS CUIs for m_x and m_y have at least one common item
`NoCommonUmlsCUI`	If lists of UMLS CUIs for m_x and m_y have no common item
`BothHumanName*`	If both m_x and m_y are human names
`BothClinicalPatients`	If both m_x and m_y are patients
Feature type: Grammatical	
`HasSameGovernor*`	If both m_x and m_y are either subject or object and are syntactically dependent on the same word
`BothPronouns`	If both m_x and m_y are pronouns
`SimilarNumber`	If m_x and m_y have similar number (i.e. singular/plural)
`NumberMismatch`	If m_x and m_y have different number
`SimilarGender`	If m_x and m_y have similar gender (i.e. male/female/neutral)
`GenderMismatch`	If m_x and m_y have different gender

Table 3.2: Semantic and Grammatical features for a candidate antecedent (m_x) and the mention to be resolved (m_y). Features with (*) mark indicate new feature types proposed by us.

tion, transmucosal, etc). They are primarily used for contextual feature extraction (see Table 3.4).

We used MetaMap to return related UMLS concept names, corresponding CUIs (concept unique identifiers) and `matched strings` (between the queried mention names and the UMLS concept names) for each of the mention names[7] (and their corresponding types) in the i2b2/VA corpus. CUIs for duplicate `matched strings` are removed and the remaining CUIs are sorted by the number of words in the corresponding `matched strings`.

A list is created with the sorted CUIs and their corresponding `matched strings`. For any candidate antecedent (m_x) and the mention to be resolved (m_y), the shortest `matched strings`, $string_x$ and $string_y$, are identified from the list where $string_x$ contains m_x and $string_y$ contains m_y.[8]

[7]After removing determiner, if any.

[8]Ideally, this shortest `matched string` would be the MetaMap `matched string` obtained for the

LEXICAL **features describing a candidate antecedent (m_x) and the mention to be resolved (m_y)**	
Feature	*Description*
`ExactStringMatch`	If m_x and m_y are not pronouns and have exact string similarity
`NNPsExactStringMatch*`	Same as `ExactStringMatch` + If all of the words in m_x and m_y have POS tag NNP (i.e. noun, proper, sin- gular)
`FullMatchWithoutDet`	If m_x and m_y are not pronouns and have exact string similarity if determiner(s) is excluded
`NNPsFullMatchWithoutDet*`	Same as `FullMatchWithoutDet` + if all of the words in m_x and m_y have POS tag NNP
`AntContainsAnph*`	If m_x and m_y are not pronouns, do not have exact string similarity (w/o determiner) but the m_x contains m_y
`NNPsAntContainsAnph*`	Same as `AntContainsAnph` + if all of the words in m_x and m_y have POS tag NNP
`AnphContainsAnt*`	If m_x and m_y are not pronouns, do not have exact string similarity (w/o determiner) but the m_y contains m_x
`NNPsAnphContainsAnt*`	Same as `AnphContainsAnt` + If all of the words in m_x and m_y have POS tag NNP
`HeadWordMatches`	If m_x and m_y are not pronouns and neither have exact string similarity (w/o determiner) nor one of them contains the other but their syntactic head words are identical
`NNPsHeadWordMatches*`	Same as `HeadWordMatches` + If all of the words in m_x and m_y have POS tag NNP
`EqualTotWordNoStringSim*`	If m_x and m_y are not pronouns, have equal number of words excluding title/determiner) and have at least one common word (excluding head word)
`NNPsEqualTotWordNoStringSim*`	Same as `EqualTotWordNoStringSim` + If all of the words in m_x and m_y have POS tag NNP

Table 3.3: Lexical features for a candidate antecedent (m_x) and the mention to be resolved (m_y). Features with (*) mark indicate new feature types.

CONTEXTUAL **features describing a candidate antecedent (m_x) and the mention to be resolved (m_y)**	
Feature	*Description*
`SemanticallyClosestAnt`	If m_x is the closest semantically similar candidate antecedent of m_y
`SenDist`	Distance between the sentences containing m_x and m_y + 1
`FollowedByEqualNumbers*`	If m_x and m_y are `Treat./Test` followed by equal numbers
`FollowedByUnequalNumbers*`	If m_x and m_y are `Treat./Test` followed by different numbers
`SamePositionOrSizeCue`	If m_x and m_y are `Prob./Test` and their names contain at least one common *position/size cue*
`DiffPositionOrSizeCue`	If m_x and m_y are `Prob./Test`, and both of them contain *size cue(s)* but there is no common cue
`SameFreqQuantity`	If m_x and m_y are `Treat.`, and words around them contain at least one common *frequency/quantity cue*
`DiffFreqQuantity`	If m_x and m_y are `Treat.`, and they or words around them contain *frequency/quantity cue(s)* but there is no common cue
`SamePhysicalLocation`	If m_x and m_y are `Prob./Test/Treat.`, and they or words around them contain at least one common *physical location cue*
`DiffPhysicalLocation`	If m_x and m_y are `Prob./Test/Treat.`, and they or words around them contain *physical location cue(s)* but there is no common cue
`SameDrugAdminMode`	If m_x and m_y are `Prob./Test/Treat.`, and they or words around them contain at least one common *drug administration mode cue*
`DiffDrugAdminMode`	If m_x and m_y are `Prob./Test/Treat.`, they or words around them contain *drug admin. mode cue(s)* but there is no common cue
`SameTemporalExp*`	If m_x and m_y are `Prob./Test/Treat.`, and their corresponding sentences contain same *date/time*
`DiffTemporalExp*`	If m_x and m_y are `Prob./Test/Treat.`, and their corresponding sentences contain different *date/time*
`SameYearMonthCue*`	If m_x and m_y are `Prob./Test/Treat.`, and their corresponding sentences contain same *year/month*
`DiffYearMonthCue*`	If m_x and m_y are `Prob./Test/Treat.`, and their corresponding sentences contain different *year/month*

Table 3.4: Contextual features for a candidate antecedent (m_x) and the mention to be resolved (m_y). Features with (*) mark indicate new feature types.

If no such string can be found for any of the mentions, then the longest `matched strings`, $string_x$ and $string_y$, are located (if any) where m_x contains $string_x$ and m_y contains $string_y$. Finally, the corresponding CUIs for $string_x$ and $string_y$ (stored inside the list) are used for extraction of the features `HasCommonUmlsCUI` (i.e. at least one common CUI) and `NoCommonUmlsCUI`.

We also consider syntactic dependencies to create various grammatical features (apart from the traditional features such as *number, gender*, etc). For example, for any mention m (which can be either a candidate antecedent or the mention to be resolved), if it is a type of subject or object then the corresponding syntactic dependency that it has with its governor word is added as a feature.

Additionally, we split mentions of `Person` type in clinical text into four subcategories using regular expressions of various contextual clues and used them during feature extraction. These subcategories are: `Patient, Family, Doctor` and `Other`.

3.6 Experimental Results

As mentioned earlier, we built and tuned the system based on the experiments using *dev-train* and *dev-test* data. We kept the official test corpus as unseen during this stage and used it only during the final evaluation. All the scores are computed using the official evaluation scripts released by the i2b2/VA 2011 challenge organisers.

---

corresponding mention itself given that the `matched string` has the same words as in the corresponding mention name. However, for some mentions (e.g. *"severe airway obstruction"*) the MetaMap matched strings contain fewer words than in the original name. In such case, the shortest `matched string` would be an empty string.

Features describing a mention m (of a candidate mention pair) which can be either a candidate antecedent or the mention to be resolved	
Feature	*Description*
Feature type: Semantic	
`SemType`	Semantic type (i.e. `Person/Test/Treatment/Problem`) of m
`HumanName*`	If m is a human man
`ClinicalPersonType`	Clinical person type (i.e. Patient/Doctor/Family/Other) of m if it is of type `Person`
Feature type: Grammatical	
`HasPossessiveCase`	If m has possessive case
`HeadWord`	the word of m on which its other words are syntactically dependent
`subj-type*`	If m is a subject, then its dependency type(s) with its governor(s)
`obj-type*`	If m is a object, then its dependency type(s) with its governor(s)
`Pronoun`	If m is pronoun
`Reflexive`	If m is a reflexive pronoun
`FirstNPInCurSen`	If m is the first NP of the corresponding sentence
Feature type: Contextual	
`ContainsTimePeriod*`	If m is of type `Treatment` and if the corresponding sentence contains cue about *in which period of the day after how many hours*
`FollowedByNumber*`	If m is of type `Treatment/Test` and followed by a number
Additional features if m is the mention to be resolved	
Feature	*Description*
Feature type: Grammatical	
`DemonsPronoun`	If m is a demonstrative pronoun
`DemonsNP`	If m is a demonstrative NP

Table 3.5: Features describing a mention m (of a candidate mention pair) which can be either a candidate antecedent or the mention to be resolved. Features with (*) mark indicate new feature types.

	B^3			*MUC*			*CEAF*			*BLANC*			**Average**		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
Using the complete proposed system															
All	0.981	0.954	0.967	0.796	0.888	0.839	0.856	0.913	0.883	0.759	0.922	0.821	0.878	0.918	**0.897**
Test	0.981	0.957	0.969	0.735	0.904	0.811	0.900	0.958	0.928	0.956	0.884	0.917	0.872	0.94	**0.903**
Person	0.895	0.945	0.919	0.929	0.880	0.904	0.791	0.670	0.726	0.748	0.941	0.817	0.872	0.832	0.850
Problem	0.977	0.937	0.957	0.596	0.860	0.704	0.832	0.944	0.885	0.888	0.724	0.784	0.802	0.914	0.849
Treatment	0.986	0.942	0.963	0.693	0.937	0.797	0.843	0.952	0.895	0.941	0.784	0.845	0.841	0.944	**0.885**
If traditional approach of chain clustering is used instead of the proposed approach															
All	0.963	0.954	0.958	0.794	0.821	0.808	0.854	0.872	0.863	0.778	0.861	0.814	0.870	0.882	0.876
Test	0.963	0.959	0.961	0.756	0.799	0.777	0.898	0.926	0.912	0.915	0.888	0.902	0.872	0.895	0.883
Person	0.848	0.927	0.886	0.907	0.845	0.875	0.760	0.610	0.676	0.769	0.870	0.812	0.838	0.794	0.812
Problem	0.959	0.939	0.949	0.616	0.744	0.674	0.834	0.916	0.873	0.820	0.728	0.766	0.803	0.866	0.832
Treatment	0.968	0.943	0.955	0.709	0.842	0.770	0.847	0.928	0.886	0.880	0.784	0.825	0.841	0.904	0.870
If traditional approach of instance creation is used instead of the proposed approach															
All	0.961	0.913	0.936	0.628	0.759	0.687	0.748	0.830	0.787	0.743	0.897	0.802	0.779	0.834	0.804
Test	0.968	0.926	0.947	0.517	0.762	0.616	0.821	0.928	0.871	0.915	0.790	0.842	0.769	0.872	0.811
Person	0.892	0.914	0.903	0.891	0.856	0.874	0.708	0.625	0.664	0.740	0.933	0.809	0.830	0.798	0.814
Problem	0.943	0.896	0.919	0.314	0.491	0.383	0.727	0.895	0.803	0.686	0.599	0.630	0.661	0.761	0.702
Treatment	0.952	0.874	0.911	0.326	0.589	0.420	0.688	0.888	0.775	0.753	0.613	0.656	0.655	0.784	0.702

Table 3.6: Results on the i2b2/VA 2011 official test corpus excluding UPMC data. Boldface shows the best obtained results on this dataset.

	B^3			*MUC*			*CEAF*			*BLANC*			**Average**		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**	**P**	**R**	**F$_1$**
Using the complete system proposed EXCEPT lexical features															
All	0.984	0.94	0.961	0.735	0.885	0.803	0.818	0.907	0.861	0.751	0.915	0.813	0.846	0.911	0.875
Test	0.982	0.932	0.956	0.501	0.895	0.643	0.826	0.951	0.884	0.935	0.729	0.800	0.770	0.926	0.828
Person	0.89	0.933	0.911	0.917	0.868	0.892	0.748	0.632	0.685	0.742	0.942	0.813	0.852	0.811	0.829
Problem	0.983	0.925	0.953	0.525	0.902	0.664	0.808	0.95	0.873	0.928	0.708	0.780	0.772	0.926	0.830
Treatment	0.986	0.931	0.958	0.638	0.935	0.759	0.818	0.948	0.878	0.954	0.759	0.830	0.814	0.938	0.865
Using the complete system proposed EXCEPT grammatical features															
All	0.972	0.953	0.962	0.787	0.834	0.81	0.848	0.879	0.864	0.739	0.906	0.801	0.869	0.889	0.879
Test	0.972	0.958	0.965	0.742	0.859	0.796	0.898	0.954	0.925	0.942	0.885	0.912	0.871	0.924	0.895
Person	0.843	0.935	0.887	0.904	0.825	0.863	0.701	0.519	0.597	0.729	0.922	0.797	0.816	0.760	0.782
Problem	0.968	0.936	0.952	0.596	0.787	0.678	0.828	0.94	0.881	0.811	0.725	0.762	0.797	0.888	0.837
Treatment	0.976	0.942	0.959	0.698	0.892	0.783	0.842	0.95	0.893	0.923	0.786	0.842	0.839	0.928	0.878
Using the complete system proposed EXCEPT semantic features															
All	0.981	0.952	0.966	0.786	0.888	0.834	0.85	0.913	0.880	0.761	0.918	0.822	0.872	0.918	0.893
Test	0.975	0.956	0.965	0.724	0.857	0.785	0.894	0.948	0.92	0.935	0.874	0.903	0.864	0.920	0.890
Person	0.913	0.944	0.928	0.925	0.892	0.908	0.792	0.716	0.752	0.752	0.939	0.820	0.877	0.851	**0.863**
Problem	0.978	0.936	0.957	0.583	0.864	0.696	0.827	0.939	0.879	0.891	0.717	0.779	0.796	0.913	0.844
Treatment	0.986	0.939	0.962	0.672	0.929	0.78	0.835	0.948	0.888	0.935	0.769	0.833	0.831	0.939	0.877
Using the complete system proposed EXCEPT contextual features															
All	0.974	0.956	0.965	0.808	0.869	0.838	0.862	0.900	0.881	0.762	0.924	0.824	0.881	0.908	0.894
Test	0.97	0.959	0.964	0.749	0.841	0.792	0.899	0.937	0.918	0.934	0.887	0.909	0.873	0.912	0.891
Person	0.879	0.947	0.912	0.93	0.876	0.902	0.789	0.652	0.714	0.753	0.942	0.822	0.866	0.825	0.843
Problem	0.973	0.94	0.956	0.63	0.836	0.719	0.843	0.932	0.886	0.868	0.740	0.791	0.815	0.903	**0.854**
Treatment	0.981	0.944	0.962	0.715	0.904	0.798	0.852	0.941	0.895	0.909	0.794	0.842	0.849	0.930	**0.885**

Table 3.7: Results with feature type ablation on the i2b2/VA 2011 official test corpus (excluding UPMC data). Boldface shows the best obtained results on this dataset.

3.6.1 Evaluation on the i2b2/VA 2011 Official Test Corpus excluding UPMC data

Table 3.6 shows the results on the i2b2/VA 2011 official test corpus excluding UPMC data. The unweighted micro-averaged F_1 score is 0.897, which is on par with the results obtained by the top teams of the challenge (see Uzuner et al. (2012) for the results and ranking of the participating teams).

Our results are better (with respect to the results of the top participating teams) for `Test` and `Treatment` chain types and somewhat similar for `Problem` type. However, the result for `Person` type is lower, though it is still high (average F_1 score 0.85).

The lowest individual F_1 score obtained by our system is the MUC score 0.704 for `Problem`. This indicates that our system is robust enough to obtain good scores for any of the metrics and for any of the four chain types. In comparison, the lowest individual F_1 score obtained by the best system of the challenge was the MUC score 0.476 for `Test`.

To understand the impact of our aggressive and greedy chain clustering, we evaluated our system after replacing the proposed clustering with the traditional closest-first clustering. This reduces the micro-averaged F_1 score to 0.876 due to a sharp decrease in recall of MUC and $CEAF$ metrics for each of the chain types (see Table 3.6).

We also investigated the impact of the proposed controlled training and test instance creation. We found that:

Total instances created without using proposed constraints [9] *= 3,482,114*
Total instances created with all the proposed constraints = 117,936

[9]Except the rule proposed by Ng and Cardie (Ng and Cardie, 2002), sentence distance limits and semantic type agreement requirement.

When we used the instance creation approach proposed by Ng and Cardie (2002) along with sentence distance limits and semantic type agreement requirement[10], the micro-averaged F_1 score of the system degraded to 0.804 (Table 3.6). This follows the deterioration of MUC and $CEAF$ scores for each of the chain types along with comparatively smaller degradation of B^3 scores.

Our overall observation is that B^3 is always fairly high, while MUC and $CEAF$ vary more.

We also conducted experiments to evaluate the contribution of different feature types (Table 3.7). Empirical results show that lexical and grammatical features contribute more than the semantic and contextual features. When looking closer, we can see that the contribution of the latter is contrasted depending on chain type: semantic features improve `Test`, `Problem` and `Treatment`, but degrade `Person`, whereas contextual features improve `Test` and `Person` but degrade `Problem` and `Treatment`. This suggests that training separate classifiers for each chain type might help optimize the contribution of the different features for each type and hence improve the global results.

3.6.2 Results on the i2b2/VA 2011 Full Official Test Corpus

Empirical outcome (without further tuning the system) on the full test corpus is almost identical to the results obtained excluding UMPC data. There is a slight decrease of the micro-averaged F_1 score (from 0.897 to 0.895). This is because of the fact that we tuned various parameters of our system (as well as the parameters of the SVM classifier) on the dev-train data which do not contain any UPMC training data. We did not tune

[10]If both the mentions of a pair are pronouns, then the semantic type agreement requires that either both pronouns be personal pronouns (e.g. *he*) or be non-personal pronouns (e.g. *it*). If one of the mentions is a pronoun while the other is not, then agreement requires that the pronoun be a personal pronoun and the mention be of type `Person` or vice versa.

	B^3			MUC			$CEAF$			$BLANC$			**Average**		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Using the complete proposed system															
All	0.979	0.955	0.967	0.796	0.876	0.834	0.861	0.910	0.885	0.756	0.915	0.817	0.879	0.914	0.895
Test	0.976	0.958	0.967	0.728	0.874	0.794	0.906	0.957	0.931	0.939	0.874	0.904	0.870	0.930	0.897
Person	0.907	0.935	0.921	0.904	0.871	0.887	0.783	0.715	0.747	0.747	0.931	0.814	0.865	0.840	0.852
Problem	0.978	0.940	0.959	0.607	0.862	0.713	0.844	0.945	0.892	0.895	0.752	0.808	0.810	0.916	0.855
Treatment	0.983	0.947	0.965	0.689	0.916	0.786	0.856	0.951	0.901	0.941	0.784	0.846	0.843	0.938	0.884

Table 3.8: Results on the i2b2/VA 2011 full official test corpus.

the system again because parameter tuning was not the main focus of our study. Here again, the variation of results differs depending on chain type.

3.7 Comparison of Results with Other Studies

In this section, we compare our results with other recently published studies that also conducted experiments on i2b2/VA 2011 challenge corpus.

Dai et al. (2012) reported an unweighted F_1 score of 0.871 on the i2b2/VA 2011 official test corpus excluding UPMC data which is lower than our results (0.897). They did not mention the results of individual evaluation metrics for different chain types. It appears that for `Test`, `Treatment` and `Problem` their system obtained unweighted F_1 scores lower than 0.80.

Rink et al. (2012) obtained an unweighted F_1 score of 0.906 on the full official test corpus, which is slightly higher than our results. They were the second best team in the challenge. However, according to their results, our system obtains better unweighted F_1 scores for `Test` (theirs: 0.823, ours: 0.897) and `Treatment` (theirs: 0.828, ours: 0.884), and almost similar outcome for `Problem` (theirs: 0.858, ours: 0.855). They did not report scores of individual evaluation metrics for individual chain types. So, we are unable to compare whether their system is as robust as ours for

different evaluation metrics.

Ware et al. (2012) achieved an unweighted F_1 score of 0.848 on the full official test corpus (lower than our results). Their system performs poorly for MUC metrics, e.g. it obtains only 0.254 MUC F_1 score for `Test` chain type.

Gooch and Roudsari (2012) obtained an unweighted F_1 score of 0.875 on the full official test corpus (our result: 0.895). Like some of the above mentioned studies, they also did not mention results of individual evaluation metrics for different chain types.

We already discussed the limitations of Xu et al. (2012) (best system of the i2b2/VA 2011 challenge) in Section 3.1. Apart from that, we noticed that their results for `Test` and `Treatment` are lower than ours, but almost similar for the `Problem` chain type.

Some may find similarities between our proposed approach and the existing multi-pass sieve based approach (Raghunathan et al., 2010; Jonnalagadda et al., 2012) because both of the approaches exploit (not necessarily the same) heuristics. But there are some major differences. For example, in a multi-pass sieve based system, each sieve (or tier) builds on the output of the previously applied sieves. Each sieve proposes candidate co-referent entity mention pairs based on its own deterministic rules. However, in our approach, the (various semantic, linguistic and syntactic) heuristics are used to filter non-co-referent entity mention pairs from training data to reduce data skewness as well as sub-optimal instances to train more accurate machine learning classifier. Also, we use various heuristics on test data to avoid as much false positives as possible to reduce the possible errors to be made by the trained classifier.

The sieve based approach of Jonnalagadda et al. (2012) proposed for coreference resolution on clinical data does include two filters. Most of their heuristics (used in their filters) are contextual constraints. In comparison,

most of our heuristics are semantic and syntactic constraints.

There also exists some other major differences between our and their approach. For example, Jonnalagadda et al. (2012) used a ML based approach in a sieve only for pronominal co-reference resolution. This sieve is one of their several sieves (i.e. not the only classifier/component for identifying co-referent pairs of mentions). In contrast, our ML classifier is the only component[11] that identifies co-referent pairs in the test data and it considers both nouns and pronouns. The impact of the differences in various heuristics, classification and clustering between their and our approaches is visible in the outcome (our unweigthed average F_1 score 0.895; their unweigthed average F_1 score 0.843).

3.8 Errors and Inconsistencies in the i2b2/VA 2011 Challenge Data

The i2b2/VA 2011 corpus has a number of inconsistencies and errors. Some of these exist because they are part of clinical text (e.g. spelling mistakes, capitalization mistakes, inconsistent term and title usage such as *"mr"* vs *"Mr."*, etc), while others were introduced during data conversion (e.g. XML tag "<" instead of "<") and data annotation.

We found at least 45 sentences in the clinical text files of the Beth Israel Deaconess Medical Center and Partners Healthcare which are incorrectly split into multiple lines (i.e. sentences) and, therefore, some of the mentions in those sentences have boundary annotations that span over multiple sentences.

We also observed that, in some of the chains, personal pronouns of both male and female types are wrongly annotated as co-referents. In a few

[11]Except for the two heuristics used for cataphora resolution and string matching with the "the patient|patient" during chain expansion as described in Section 3.3.7.

other chains, "the patient" is incorrectly put in the same chain with non-patient persons such as "her father", "her ophthalmologist", etc. There also exist inconsistent annotations. For example, in some of the documents the term "attending" is annotated as co-referent with the immediate following name of the doctor but in other documents it is annotated in a different way. Wrong boundary annotations for some of the mentions have also been detected in some documents. These are indeed the inevitable imperfections of human annotated data: according to Uzuner et al. (2012, Appendix II, Table 2), the best inter-annotator agreement for human annotators on co-referring mention pairs was 0.81. These annotation errors probably contribute to a (hopefully small) part of the training and evaluation errors.

3.9 Limitations and Possible Future Extension of This Study

One of the limitations of our approach is that although it obtains fairly high results (average F_1 score: 0.852) for `Person`, they are not as good as the results (for this particular chain type) reported by Xu et al. (2012). However, it should be noted that they used a separate classifier solely trained for identifying co-referent pairs of type `Person`. Such an additional classifier can be easily included in our approach.

Another limitation is that we only use UMLS Metathesaurus and some (domain specific) Wikipedia articles for the exploitation of world knowledge. However, there exist a number of other resources which other approaches had used and might also be useful for our system. These includes Wordnet (Fellbaum, 1998), Probase (Song et al., 2011), NeedleSeek (Lee et al., 2011), Evidence (Zhang et al., 2011), RadLex[12], etc.

Our approach does not include any specific technique to cope with para-

[12]http://www.radlex.org/

phrases (e.g. *"left ankle wound"* and *"a small complication"*) when the expressions share no common word. Also, one other area of improvement could be to identify temporal expressions more accurately and then to normalize them using tools such as TIMEN (Llorens et al., 2012) which can help to extract informative features. Currently, we use regular expressions to identify temporal expressions.

Highly accurate POS tagging and parsing output can further uplift the performance of our system. It has been shown that the Stanford parser (which we used in our study) achieves around 77% bracketing F_1 score for POS tagging[13] on a randomly constructed sub-corpus from the 2010 i2b2/VA NLP challenge clinical data (Xu et al., 2011). The Stanford parser is currently not trained on clinical treebank. Hence, this introduces some limitations in our system.

One could extend our study by addressing the above limitations. Including clues about various sections in a clinical text might also help. Inclusion of a separate classifier for patient/non-patient identification (that has been reported as very effective for the improvement of `Person` type co-referent classification by Xu et al. (2012)) could be another possible extension.

3.10 Key Ideas in the Proposed Approach and Their Potential Usage on Biomedical Text

While we have performed experiments only on clinical texts, we argue that the results achieved in this study can be of broader interest for the coreference resolution community.

As we mentioned in Section 3.2, there are four major supervised coreference resolution approaches. Among them, the mention-pair model is by far the most widely studied and most popular, and has been applied on all

[13] Note that correct parsing depends a lot on correct POS tagging.

of the following genres (newswire, biomedical and clinical). So, we wanted to investigate whether the general architecture of the mention-pair model can be further improved.

We mentioned in Section 3.1 the issues regarding the usage of different evaluation metrics and the disparity in scores obtained. So, we also wanted to investigate whether it is possible to minimize the differences among the scores obtained for different metrics, at the same time maintaining their unweighted average high.

In conclusion, we would like to summarize some of the key ideas of the proposed variant of the popular mention-pair model. Firstly, we argue to exploit the combination of a series of linguistically and semantically motivated constraints that can control the generation of less-informative/sub-optimal training and test instances. This strategy could be equally beneficial for other genres of text. Secondly, the greedy clustering of mention pairs proposed in this study can be used on biomedical text with only some minor modifications (e.g. the cluster merging heuristic rule for *Patient* mentions would not be applicable for biomedical text). Finally, various rules such as restricting the number of candidate antecedents in case of male/female pronouns (see Section 3.3.4) or rejecting unlikely candidate antecedents in the previous sentence for a Wh-determiner (see Section 3.3.2), etc are based on general writing style in English and should be equally effective for other genres (in English).

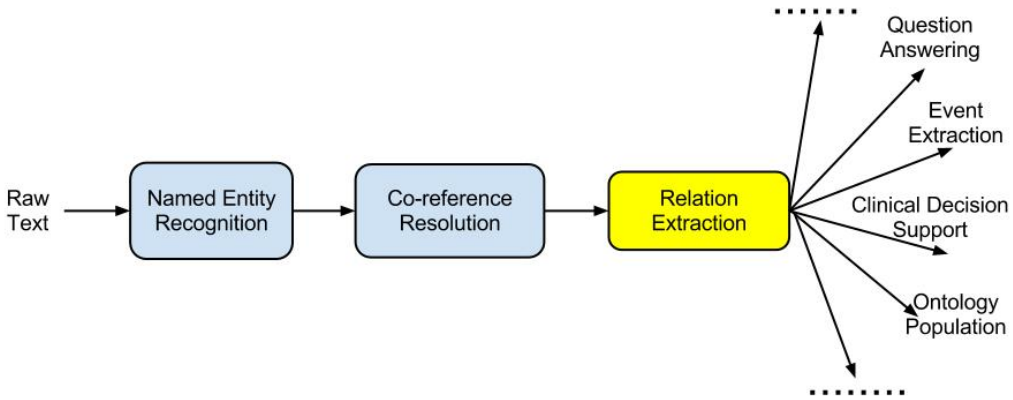
Chapter 4

Relation Extraction

"Relating two entity words in a sentence requires a skillful combination of local and nonlocal noisy clues from diverse syntactic and semantic structures in a sentence."

– Sunita Sarawagi

"Information Extraction", Foundations and Trends in Databases, 1(3):261–377. (2007)



Relation extraction (RE) is the task that aims at identifying instances of pre-defined semantic relation types between pairs of entity mentions in a given text. For example, given the following sentence

Hillary Rodham moved to Arkansas in 1974 and married Bill Clinton in 1975.

an RE system trained for identifying instances of the "wife_of" relation is expected to find that the above sentence contains an instance of such a relation (that holds between the mentions "Hillary Rodham" and "Bill Clinton"). RE often serves as an important preliminary step for various advanced text mining tasks.

As mentioned in Chapter 1, RE is the main focus of this PhD research. In this chapter, we will lay out our proposed approach for RE. Although it has been originally developed for and tested on biomedical data, the approach has been later applied to other genres of text performing equally well (see Appendix B).

4.1 Basic Terminology

Every (binary) relation instance has two arguments that have to be filled by two entity mentions. The compatible entity types for the arguments (and in some cases the order of the arguments, too) are pre-defined. For example, for the "wife_of" relation, both of the arguments are expected to be of type "person". In this case, the type "person" would be called **target entity** for the "wife_of" relation. Any entity other than the target entities (w.r.t. the particular relation type) would be called **non-target entities**.

4.2 Current state of RE research

In the following subsections we will survey the current status of the RE research area (mainly focussing on the biomedical domain), describing the prevalent approaches, the main limitations and the open issues.

We first discuss two of the main characteristics of the current RE approaches, i.e. the fact that usually they focus on individual sentences (Section 4.2.1) and approach RE as a classification problem (Section 4.2.2).

Then we briefly describe the main existing RE methodologies (Section 4.2.3). Following that, we discuss three critical issues in RE, i.e. imbalance in data distribution (Section 4.2.4), domain adaptation (Section 4.2.5) and supervision (Section 4.2.6). Finally, we describe existing biomedical RE approaches, focussing on the work done in the context of protein-protein interaction (PPI, Section 4.2.7) and drug-drug interaction (DDI, Section 4.2.8) extraction since these are the RE tasks that have been receiving more attention in the community interested in IE from biomedical texts.

4.2.1 Predominantly intra-sentential

Current RE research is mostly focused on *intra-sentential* relations, i.e. relations holding between entity mentions in the same sentence. The motivation behind such choice is that the vast majority of the relations involves entities appearing both in the same sentence. This is confirmed by the few work discussing cross-sentential relations (i.e. relations involving entity mentions beyond sentence boundaries) (Stevenson, 2006; Swampillai and Stevenson, 2010). For example, Swampillai and Stevenson (2010) report that 90.6% of the total number of relations in the ACE03[1] corpus (a benchmark news domain RE corpus) are intra-sentential. Like most of the previous RE work, in this thesis we concentrate only on intra-sentential relations. We leave the extension of our proposed approach for cross-sentential RE as future work.

4.2.2 A classification problem

State-of-the-art RE approaches are based on various Machine Learning (ML) techniques and they usually approach RE as a classification task. All possible entity mention pairs (compatible with the given relation) inside individual sentences are enumerated, and then each of the candidate pairs

[1] http://www.itl.nist.gov/iad/mig/tests/ace/2003/

is classified as one of the target class labels (in case of binary classification, positive or negative instance[2]). This is the standard method in RE. The main differences among systems concern the choice of trainable classifier and the representation of instances (McDonald et al., 2005). One exception to the standard approach of RE is Miller et al. (2000), who approached relation extraction as just a form of probabilistic parsing where parse trees are augmented to identify all relations. In a different work, Roth and Yih (2004) approached RE jointly with entity type classification (instead of approaching RE separately). They used a set of global constraints over locally trained classifiers.

4.2.3 RE Methodologies

RE approaches generally fall into three main categories: *(i)* exploitation of statistics about co-occurrences of entity mention pairs, *(ii)* usage of patterns and rules, and *(iii)* usage of machine learning (ML) classifiers. These approaches have been studied for a long period and each has its own pros and cons. Exploitation of co-occurrence statistics results in high recall but low precision, while rule or pattern based approaches can increase precision but suffer from low recall.

ML based approaches[3] can be broadly categorized into two groups: *(a)* structural similarity based ML approaches and *(b)* flat feature based ML approaches. Structural similarity based ML approaches usually employ kernel methods to automatically exploit a large amount of features (without an explicit feature representation) to measure the similarity between structures of the same type (e.g. sub-sequences, trees, graphs, etc). The

[2]**Positive examples/instances/pairs** are those candidate entity mention pairs between which a given relation holds in the corresponding context. **Negative examples** are those between which the relation does not hold.

[3]Regarding the choice of ML algorithms, support vector machines (SVMs) and maximum entropy models (MaxEnt) are the most popular choices for RE. Other choices include decision trees, integer linear programming, etc.

main characteristic of kernel methods is that they map the data into higher dimensional spaces so that, if the data is not linearly separable in a lower dimension, then it is mapped into a higher-dimensional space where the data could become more easily separated or better structured. It is beyond the scope of this thesis to discuss the theory of kernel methods. Readers are referred to Shawe-Taylor and Cristianini (2004) for an introduction.

Flat feature based ML approaches, on the other hand, depend on feature engineering to explicitly select various feature types to be used. Note that kernel based classifiers can be also trained using explicit feature sets (often dubbed as feature based kernels). Hence, both of the ML approaches can be combined into hybrid/composite kernels.

4.2.4 Imbalance in data distribution

Like in other NLP tasks, it has been claimed that advances of ML based approaches in RE are hampered by the imbalanced distribution of positive and negative instances in the annotated training data. For example, Sun et al. (2011) hypothesized (without providing empirical evidence) that the unbalanced distribution of instances is an obstacle for further improving the performance of RE. Usually, the number of negative instances is much larger than that of the positive ones and such skewness exists both in the training and in the test data.

RE approaches use different strategies to reduce the number of elements to be considered for feature or pattern extraction. These include considering only part of the phrase structure parse tree (Moschitti, 2004; Zhang et al., 2005; Zhou et al., 2007) or part of the dependency graph (Culotta and Sorensen, 2004; Bunescu and Mooney, 2005a; Chowdhury et al., 2011), limiting the window of words on the left and right of the entities (Giuliano et al., 2006a; Bunescu and Mooney, 2006), etc. However, such strategies do not reduce the number of candidate entity mention pairs.

Ideally, to reduce the skewness of instances, the informativeness of both positive and negative instances should be taken into account. In their seminal work regarding selection of features and instances, Blum and Langley (1997) pointed out that, as learning progresses and the learner's knowledge about certain parts of the training data increases, the remaining data which are similar to the already "well-understood" portion become less useful.

One of our goals is to get rid of such instances from the annotated data before training the ML classifier to reduce the imbalance in instance distribution and to obtain a more accurately learned model/classifier. Ideally, a well trained classifier is expected to successfully identify the true positive instances in the test data, distinguishing them from the negative instances; in other words, it is expected to avoid labelling any negative instance as a *(false) positive* instance. But, in practice, a classifier does mistakenly label (false) positives. To reduce the probability of such incorrect labeling, we aim to automatically get rid of as many negative instances as possible from the test data (before applying the learned classifier) using the same knowledge used to reduce skewness in the training data. The goal is to curb the number of false positives produced by the classifier.

Different techniques are employed in open domain IE[4] for filtering irrelevant data to construct datasets. For example, whether the semantic type of the retrieved entity mentions and that of the target mentions are the same[5], or the number of words between the candidate mentions is greater than a certain limit, etc (Banko et al., 2007; Wu and Weld, 2010; Wang et al., 2011). However, such filtering is applied in a setting substantially different from ours.

[4]Open domain IE has substantial differences with traditional RE some of which are discussed in Wang et al. (2011).

[5]In traditional RE, any pair of mentions to be considered as an instance must satisfy the already known argument types of the target relation. Hence, this technique does not qualify as a criterion for negative instance filtering in traditional RE.

4.2.5 Domain adaptation

One of the most pressing issues in RE is domain adaptation. Widely studied RE approaches on the news domain usually perform badly on specialized domains (such as biomedical data) and vice versa. As part of our experiments in this PhD research, we have implemented and used an existing state-of-the-art feature based RE approach (Zhou et al., 2005), originally proposed for and tested on the news domain, for biomedical RE. The results reveal startling variation in performance (discussed in detail in Section 4.7.1).

One of the main reasons behind this problem is the choice of the features. Features explicitly chosen to tune performance for a specific relation in a particular domain might prove not equally effective for other types of relations/domains. Moreover, even the same set of features for the same RE task could produce substantially different results in different corpora, due to the variation in corpora characteristics (e.g. number of target entity mentions per sentence, average length of the sentences, ...).[6] Naturally, the situation becomes even more complicated when two RE corpora are composed of completely different genres of text. In this case, the differences in linguistic aspects (e.g. change in valency of certain domain specific verbs) and terminology may require new features which are sensitive to these variations.

4.2.6 Supervision and external resources

There is a growing trend in the general (i.e. news) domain to move from fully supervised to semi-supervised, distantly supervised and unsupervised approaches (Wang et al., 2007; Mintz et al., 2009; Nguyen and Moschitti, 2011a; Wang et al., 2012). This is due to the advent of linked open data

[6]We will show such variation during the discussion of experimental results later in Section 4.7.

and the availability of resources such as Wikipedia[7], Yago (Suchanek et al., 2007), Freebase (Bollacker et al., 2008), OpenCyc (Lenat, 1995), etc.[8] Some of the semi-supervised approaches have been shown to obtain competitive results to their supervised counterparts (e.g. Sun et al. (2011)) on news data. But it is far more difficult to replicate such results on biomedical data due to the nature of the texts. Instances of many relation types in the news domain can be obtained from resources such as Wikipedia infoboxes and exploited for distant (and semi-) supervision. It is true that somewhat similar resources exist in the biomedical domain (e.g. UMLS Metathesaurus (Bodenreider, 2004)). However, things are not so straightforward for biomedical RE. For example, consider the treatment/medication of a particular disease (i.e. the disease-treatment relation). The exact treatment depends not only on the disease but also on many other parameters such as patient background (i.e., gender, age, previous medical history, etc), demographics, other drugs that the patient might be taking, etc.

Existing semi-supervised biomedical RE approaches use external resources (e.g. HUGO[9], OMIM[10], etc) to discover new, potentially meaningful (specific types of) relations between biomedical entity mentions (e.g. see Hristovski et al. (2003)). However, the problem is that the results reported in such studies might not be completely reliable. These studies often exclusively rely on the co-occurrence information of target entity mentions. Hence, they are prone to fetching a lot of false positives. In most of the cases, these studies reported only a preliminary analysis of precision (Chun et al., 2006).

It is taken for granted that supervised approaches achieve (often con-

[7]http://www.wikipedia.org/

[8]It is interesting to note that some of these semi-supervised RE approaches that rely on Yago, Freebase and other resources claim to be domain independent, although hardly any empirical evidence is reported to back up such claim.

[9]http://www.genenames.org/

[10]http://www.ncbi.nlm.nih.gov/omim

siderably) better results than semi-supervised or unsupervised approaches. But the reality is that in the biomedical domain even the supervised RE performance has not yet reached a sufficient level of maturity. So, in this thesis we focus on developing high performance supervised RE approach. At the same time, while exploiting various linguistic characteristics (e.g. the linguistic scope of negation cues), we avoid the usage of annotated corpora for those characteristics and try to collect such information through self-supervision (more details later in Section 4.4).

4.2.7 Protein-protein interaction extraction

Arguably, protein-protein interaction (PPI) extraction has garnered far more attention than any other RE tasks in biomedical domain to date. For this reason, we evaluated our proposed RE approach on this task. Below we include a brief review of the existing approaches to PPI extraction.

PPI[11] information is very critical in understanding biological processes. The following sentence contains examples of PPIs that exist between $\{HFE_1, TfR_2\}$ and $\{HFE_3, TfR_4\}$.

> *The 2.8 A crystal structure of a complex between the extracellular portions of HFE_1 and TfR_2 shows two HFE_3 molecules which grasp each side of a twofold symmetric TfR_4 dimer.*

Considerable progress has been made for this task. Nevertheless, the empirical results of previous studies show that none of the approaches already known in the literature is consistently better than other approaches when evaluated on different benchmark PPI corpora. Pyysalo et al. (2008) analysed this situation and opined that there are definite limits on the ability to compare different RE approaches evaluated on different PPI corpora. They concluded that the differences stemming from the choice of

[11] PPIs occur when two or more proteins bind together, and are integral to virtually all cellular processes, such as metabolism, signalling, regulation, and proliferation (Tikk et al., 2010).

PPI corpus for evaluation can be substantially larger than the differences between the performance of different PPI extraction methods. We believe this issue requires further study and the design of new approaches that are sensitive to the variations of complex linguistic constructions in all these PPI corpora.

Several RE approaches have been reported to date for the PPI task, most of which are kernel based methods. Tikk et al. (2010) reported a benchmark evaluation of various kernels on PPI extraction. An interesting finding is that the Shallow Linguistic (SL) kernel (Giuliano et al., 2006a) (to be discussed in Section 4.3.3), despite its simplicity, is on a par with the best kernels in most of the evaluation settings.

Kim et al. (2010) proposed walk-weighted subsequence kernel using e-walks, partial matches, non-contiguous paths, and different weights for different sub-structures (which are used to capture structural similarities during kernel computation). Miwa et al. (2009a) proposed a hybrid kernel, which combines the all-paths graph (APG) kernel (Airola et al., 2008), the bag-of-words kernel, and the subset tree kernel (Moschitti, 2006) (applied on the shortest dependency paths between target protein pairs). They used multiple parser inputs.

As an extension of their work, they boosted system performance by training on multiple PPI corpora instead of on a single corpus and adopting a corpus weighting concept with support vector machine (SVM) which they call SVM-CW (Miwa et al., 2009b). Since most of their results are reported by training on the combination of multiple corpora, it is not possible to compare them directly with the results published in the other related work (that usually adopt 10-fold cross validation on a single PPI corpus). To be comparable with the vast majority of the existing work, we also report results using 10-fold cross validation on single corpora.

Apart from the approaches described above, there also exist other stud-

ies that used kernels for PPI extraction (e.g. subsequence kernel (Bunescu and Mooney, 2006)).

A notable exception from these kernel based state-of-the-art RE approaches is the work published by Bui et al. (2011). They proposed an approach that consists of two phases. In the first phase, their system categorizes the data into different groups (i.e. subsets) based on various properties and patterns. Later they classify candidate PPI pairs inside each of the groups using SVM trained with features specific for the corresponding group.

4.2.8 Drug-drug interaction extraction

While we tested our approach for PPI extraction due to its relevance in the biomedical field, we also wanted to verify whether our approach performs well for a different biomedical RE task. The other task that we chose is drug-drug interaction (DDI). DDI is a condition when one drug influences the level or activity of another. The extraction of DDIs has significant importance for public health safety. It was reported that about 2.2 million people in USA, age 57 to 85, were taking potentially dangerous combinations of drugs (Landau, 2009). An earlier report mentioned that deaths from accidental drug interactions rose by 68 percent between 1999 and 2004 (Payne, 2007). An example of DDI between *Acamprosate* and *antidepressants* is shown below.

> *Patients taking Acamprosate$_1$ concomitantly with antidepressants$_2$ more commonly reported both weight gain and weight loss, compared with patients taking either medication alone.*

Automatic DDI extraction is a relatively new RE task. One of the earliest work reported for DDI extraction is by Segura-Bedmar et al. (2011b) where they used the SL kernel. A number of other approaches were applied

during the DDIExtraction-2011 challenge[12]. Such approaches are based either on kernel methods, or on ML classifiers trained on explicit features and patterns, or ensemble based methods, where the output of different classifiers is combined to produce the final output (Segura-Bedmar et al., 2011a; Chowdhury et al., 2011c; Björne et al., 2011; Thomas et al., 2011a).

4.3 Proposed Approach

The central component of the proposed approach is a novel hybrid kernel, whose aim is to take advantage of different types of information (e.g. syntactic, contextual, semantic, etc) and their different representations (i.e. flat features, tree structures and graphs).

One of the important characteristics of our approach is that it uses various linguistically motivated techniques to get rid of what we call *less informative sentences*[13] and *less informative instances*[14]. To this aim, we exploit information coming from other NLP areas (such as scope of negations and elementary discourse units). We will describe these aspects in detail in Section 4.4. We also indirectly take advantage of the semantic roles of entity mentions. To the best of our knowledge, these topics were not explored before for RE. Importantly, our proposed approach does not need specific annotations for exploiting the above mentioned information. It only requires the annotation of the target entity mentions and of the instances of the target relations that hold between them.

To be precise, our contributions are the following:

1. We propose an approach that, differently from what is known in the literature, obtains very good results on all the 5 widely used PPI bench-

[12] http://labda.inf.uc3m.es/DDIExtraction2011/

[13] A sentence is less informative if it is unlikely to contain any instance of the relation of interest and if its exclusion does not degrade the performance of the RE system.

[14] Less informative instances are instances that share some common characteristics and whose exclusion results in better performance.

mark corpora. As a matter of fact, none of the previous biomedical RE approaches that have been tested on 5 widely used PPI benchmark corpora consistently outperform other approaches. Besides, their performance is generally not high in most of these corpora. We propose a novel hybrid kernel based RE approach that outperforms these previous approaches in 4 out of these 5 corpora. Furthermore, our result is very close to the state of the art on the other remaining corpus.

2. We propose new structures (namely, mildly extended dependency tree and reduced graph) to identify the important parts of a sentence (with respect to a pair of candidate entity mentions) to extract the target relation.

3. We propose a number of linguistically motivated rules for extracting a variety of features as well as for preprocessing the input data.

4. Our analysis suggests that tree kernels can slightly improve the F-score (by boosting recall), when combined with an already high performance hybrid kernel, but it comes at a price of a much slower runtime.

5. We propose a self-supervised technique to exploit the scope of negations for RE without using any corpus annotated with the scope of negations. This technique can be exploited to automatically filter out less informative sentences and, as a consequence, to reduce the imbalance in data distribution.

6. We propose to exploit data driven knowledge with already known common knowledge to reduce the imbalance in data distribution. Our proposed data driven knowledge was collected by indirectly exploiting the idea of semantic role labelling.

7. We propose an approach to exploit elementary discourse units to filter negative test instances.

8. Our proposed approach also outperforms previous best results on a benchmark DDI corpus, i.e. a separate biomedical RE task.

9. Our proposed approach achieves results on a par with state-of-the-art RE approaches on news domain, too. In other words, our approach shows a certain degree of domain independence.

The following sections will describe our approach in detail.

4.3.1 Proposed Kernel Combinations

We propose two kernel combinations for RE. One of them is a new hybrid (polynomial) kernel, K_{COMP}, that combines two feature vector based kernels. It is defined as follows:

$$K_{COMP}\ (R_1,\ R_2) = K_{HF}\ (R_1,\ R_2) + K_{SL}\ (R_1,\ R_2)$$

where K_{HF} is a new feature based kernel (proposed in this PhD research; we will describe it later) that uses a heterogeneous set of features, and K_{SL} is the Shallow Linguistic (SL) kernel proposed by Giuliano et al. (2006a).

Tree kernel based approaches have been shown to be effective for RE from the news domain. So, we propose another hybrid kernel, K_{Hybrid}, that combines K_{COMP} with the Path-enclosed Tree (PET) kernel (Moschitti, 2004):

$$K_{Hybrid}\ (R_1,\ R_2) = K_{COMP}\ (R_1,\ R_2) + w\ *\ K_{PET}\ (R_1,\ R_2)$$

where w is a multiplicative constant used for the PET kernel. It allows the hybrid kernel to assign more (or less) weight to the information obtained using tree structures depending on the corpus. The proposed K_{COMP} and K_{Hybrid} kernels are valid according to the closure properties of kernels.

4.3.2 Proposed K_{HF} kernel

As mentioned above, this proposed kernel uses heterogeneous features extracted from three different sources. The first one is Zhou et al. (2005) which uses 51 different features. We select the following 27 of those features for our feature set:

> *WBNULL, WBFL, WBF, WBL, WBO, BM1F, BM1L, AM2F, AM2L, #MB, #WB, CPHBNULL, CPHBFL, CPHBF, CPHBL, CPHBO, CPHBM1F, CPHBM1L, CPHAM2F, CPHAM2L, CPP, CPPH, ET12SameNP, ET12SamePP, ET12SameVP, PTP, PTPH*

A description of these features can be found in Appendix A. The other two sources are a sub-graph, called *reduced graph*, and a sub-tree structure, called *mildly extended dependency trees (MEDTs)*, both proposed in our recent studies (Chowdhury et al., 2011; Chowdhury and Lavelli, 2012b).

Features from Reduced graph

For each of the candidate entity mention pairs, we construct a type of subgraph from the dependency graph formed by the syntactic dependencies among the words of a sentence. We call it `reduced graph` and define it in the following way:

> A **reduced graph** is a subgraph of the dependency graph of a sentence which includes:
>
> - the two candidate entity mentions and their governor nodes up to their least common governor (if exists).
> - dependent nodes (if exist) of all the nodes added in the previous step.
> - the immediate governor(s) (if exists) of the least common governor.

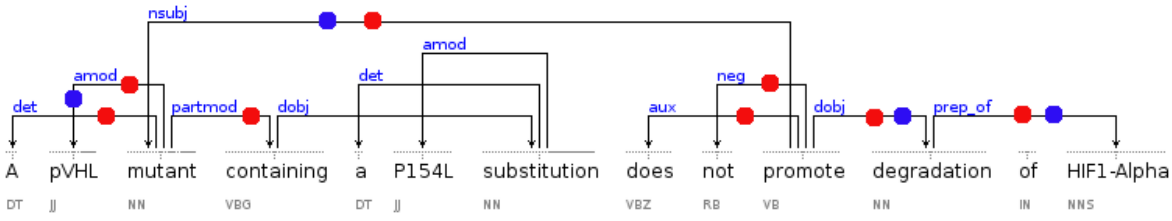


Figure 4.1: Dependency graph for the sentence "A pVHL mutant containing a P154L substitution does not promote degradation of HIF-Alpha" generated by the Stanford parser. The edges with blue dots form the smallest common subgraph for the candidate entity mention pair **pVHL** and **HIF-Alpha**, while the edges with red dots form the *reduced graph* for the pair.

Figure 1 shows an example. A reduced graph is an extension of the smallest common subgraph of the dependency graph that aims at overcoming its limitations. It is a known issue that the smallest common subgraph (or subtree) sometimes does not contain important cue words. Our objective in constructing the reduced graph is *to include any potential modifier(s) or cue word(s)* that describes the relation between the given pair of entities. Sometimes such modifiers or cue words are not directly dependent (syntactically) on any of the entity mentions (of the candidate pair). Rather they are dependent on some other word(s) which is dependent on one (or both) of the entity mentions. The word "*not*" in Figure 4.1 is one such example. The reduced graph aims to preserve these cue words.

The following types of features are collected from the reduced graph of a candidate pair:

1. *HasTriggerWord*: whether any trigger word[15] matches with one of the words inside the reduced graph.

2. *Trigger-X*: whether trigger word 'X' matches with one of the words

[15] `Trigger words` of a certain semantic relation are the words/phrases which, if present inside a text, could provide strong indication that at least one instance of that semantic relation holds between entity mentions residing inside the text.

inside the reduced graph.

3. *DepPattern-i*: whether the reduced graph contains all the syntactic dependencies of the i-th pattern of dependency pattern list.

The **dependency pattern** list is automatically constructed from the training data during the learning phase. Each pattern is a set of syntactic dependencies of the corresponding reduced graph of a (positive or negative) candidate entity mention pair in the training data. For example, the dependency pattern for the reduced graph in Figure 4.1 is {*det, amod, partmod, nsubj, aux, neg, dobj, prep_of*}. The same dependency pattern might be constructed for multiple (positive or negative) mention pairs. However, if it can be constructed for both positive and negative pairs, it is discarded from the pattern list.

The dependency patterns allow some kind of underspecification as they do not contain the lexical items (i.e. the words). Instead, they contain the likely combination of syntactic dependencies that a given related pair of candidate mentions would pose inside their reduced graph.

Features from Mildly Extended Dependency Trees (MEDTs)

An MEDT is basically a linguistically motivated extension of a minimal subtree that connects two target mentions. The goal is to include important cue words or predicates that are missing in the minimal dependency subtree, without including non-relevant words. In other words, an MEDT is more constrained than a reduced graph. We propose three expansion rules for obtaining an MEDT from a minimal dependency subtree:

- Expansion rule 1: *If the root of the minimal subtree is not a modifier (e.g. adjective) or a verb, then look for such node among its children or in its parent (in the original DT tree) to extend the subtree.*

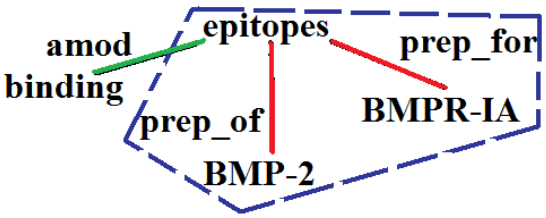


Figure 4.2: Part of the DT for the sentence "The binding epitopes of *BMP-2* for *BMPR-IA* was characterized using BMP-2 mutant proteins". The dotted area indicates the minimal subtree.

The following example shows a sentence where this rule would be applicable:

> *The binding epitopes of* ***BMP-2*** *for* ***BMPR-IA*** *was characterized using BMP-2 mutant proteins.*

Here, the cue word is "binding", the root of the minimal subtree is "epitopes" and the target entities are *BMP-2* and *BMPR-IA*. However, as shown in Figure 4.2, the minimal subtree does not contain the cue word.

- Expansion rule 2: *If the root of the minimal subtree is a verb and its subject in the original DT tree is not included in the subtree, then include it.*

Consider the following sentence:

> *Interaction was identified between* ***BMP-2*** *and* ***BMPR-IA****.*

Here, the cue word is "Interaction", the root is "identified" and the entities are *BMP-2* and *BMPR-IA*. The passive subject "Interaction" does not belong to the minimal subtree (see Figure 4.3).

- Expansion rule 3: *If the root of the minimal subtree is the head word of one of the interacting entities, then add the parent node (in the original DT tree) of the root node as the new root of the subtree.*

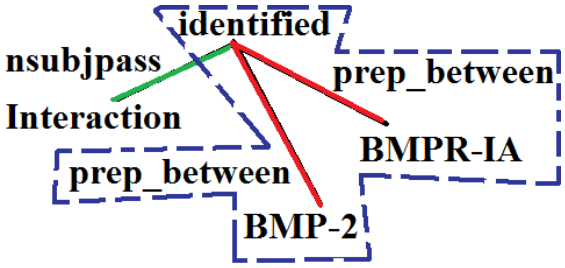


Figure 4.3: Part of the DT for the sentence "Interaction was identified between *BMP-2* and *BMPR-IA*". The dotted area indicates the minimal subtree.

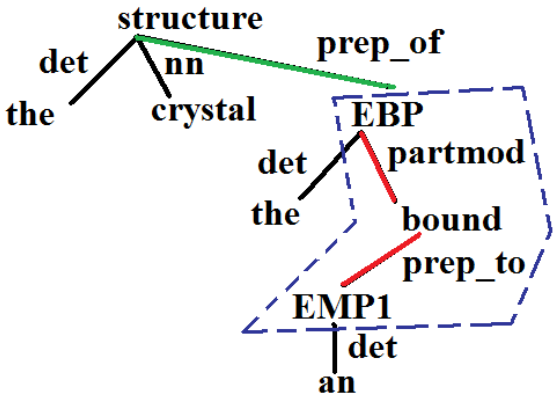


Figure 4.4: Part of the DT for the sentence "Phe93 forms extensive contacts with a peptide ligand in the crystal structure of the *EBP* bound to an *EMP1*". The dotted area indicates the minimal subtree.

This is an example sentence where this rule is applicable (see Figure 4.4):

> *Phe93 forms extensive contacts with a peptide ligand in the crystal structure of the **EBP** bound to an **EMP1**.*

We extract *e-walk* and *v-walk* features from the MEDT (expanded using all of the above expansion rules) of each candidate pair. A *v-walk* feature consists of $(word_i - dependency\_type_{i,i+1} - word_{i+1})$, and an *e-walk* feature is composed of $(dependency\_type_{i-1,i} - word_i - dependency\_type_{i,i+1})$.[16] To extract such features, we choose MEDT instead of reduced graph as we

[16]Note that, in a dependency graph, the words are nodes while the dependency types are edges.

observed that the latter often includes some uninformative words which produce uninformative walk features. The walk features extracted from MEDTs have the following properties:

- The directionality of the edges (or nodes) in an e-walk (or v-walk) is not considered. In other words, e.g., $pos(stimulatory) - amod - pos(effects)$ and $pos(effects) - amod - pos(stimulatory)$ are treated as the same feature.
- The v-walk features are of the form $(pos_i - dependency\_type_{i,i+1} - pos_{i+1})$. Here, pos_i is the POS tag of $word_i$, i is the governor node and $i+1$ is the dependent node.
- The e-walk features are of the form $(dep.\_type_{i-1,i} - pos_i - dep.\_type_{i,i+1})$ and $(dep.\_type_{i-1,i} - lemma_i - dep.\_type_{i,i+1})$. Here, $lemma_i$ is the lemmatized form of $word_i$.
- Usually, the e-walk features are constructed using dependency types between $\{governor\_of\_X,\ node\_X\}$ and $\{node\_X,\ dependent\_of\_X\}$. However, we also extract e-walk features from the dependency types between any two dependents and their common governor (i.e. $\{node\_X,\ dependent\_1\_of\_X\}$ and $\{node\_X,\ dependent\_2\_of\_X\}$).

Other Features:

In addition to the above mentioned features, surrounding tokens within the window of {-2,+2} for each candidate mention are also included as features. We extend the heterogeneous feature set by adding **features related to the scope of negations**. We use a list of 13 negation cues[17] to search inside the reduced graph of a candidate pair. If the reduced graph contains any of the negation cues or their morphological variants then we add the following features:

[17]No, not, neither, without, lack, fail, unable, abrogate, absence, prevent, unlikely, unchanged, rarely.

- *negCue:* the corresponding negation cue.
- *immediateNegatedWord:* if the word following the negation cue is neither a preposition nor a "*be* verb", then that word, otherwise the word after the next word.[18]

Furthermore, if the corresponding matched negation cue is either "no", "n't" or "not", then we add additional negation scope related features:

- *bothEntDependOnImmediateGovernor:* whether the immediate governor (if any) of the negation cue is also governor of a dependency sub-tree (of the dependency graph of the corresponding sentence) that includes both of the candidate mentions.
- *immediateGovernorIsVerbGovernor:* whether the immediate governor of the negation cue is a verb.
- *nearestVerbGovernor:* the closest verb governor (i.e. parent or grandparent inside the dependency graph), if there any, of the negation cue.

We further extend the heterogeneous feature set by adding **features related to important non-target entities** (with respect to the relation of interest). For example, for the purpose of DDI extraction, we deem the presence of *DISEASE* mentions (which might result as a consequence of DDI) can provide some clues. So, we used our proposed state-of-the-art NER system, BioEnEx (described in Chapter 2), to annotate the corpus used for RE experiments. For each candidate target entity mention pair, we add the following features in our feature set:

- *NTEMinsideSentence:* whether the corresponding sentence contains important non-target entity mention(s) (e.g. disease for DDI).

[18]For example, "interested" from "... not interested ...", and "confused" from "... not to be confused ...".

- *immediateGovernorIsVerbGovernorOfNTEM:* the immediate governor (if any) of the non-target entity mention, only if such governor is also governing a dependency sub-tree that includes both of the target candidate entity mentions.
- *nearestVerbGovernorOfNTEM:* the closest verb governor (if any) of the non-target entity mention, only if it also governs the candidate entity mentions.
- *immediateGovernorIsVerbGovernorOfNTEM:* whether the immediate governor is a verb.

4.3.3 Other component kernels

As mentioned before, we combine our proposed K_{HF} kernel with two previously proposed kernels, the Shallow Linguistic (SL) kernel and the path-enclosed tree (PET) Kernel, to design the new kernel compositions K_{COMP} and K_{Hybrid}.

Shallow Linguistic (SL) Kernel

The Shallow Linguistic (SL) kernel was proposed by Giuliano et al. (2006a). It is one of the best performing kernels applied on different biomedical RE tasks such as PPI and DDI (drug-drug interaction) extraction (Tikk et al., 2010; Segura-Bedmar et al., 2011b; Chowdhury and Lavelli, 2011b; Chowdhury et al., 2011c). It is defined as follows:

$$K_{SL}\ (R_1,\ R_2) = K_{LC}\ (R_1,\ R_2) + K_{GC}\ (R_1,\ R_2)$$

where K_{SL}, K_{GC} and K_{LC} correspond to SL, global context (GC) and local context (LC) kernels respectively. The GC kernel exploits contextual information of the words occurring before, between and after the pair of entities (to be investigated for RE) in the corresponding sentence; while the LC kernel exploits contextual information surrounding individual entities.

Path-enclosed tree (PET) Kernel

The path-enclosed tree (PET) kernel[19] was first proposed by Moschitti (2004) for semantic role labelling. It was later successfully adapted by Zhang et al. (2005) and other studies for relation extraction on general texts (such as newspaper domain). A PET is the smallest common subtree of a phrase structure tree that includes the two entities involved in a relation.

A tree kernel calculates the similarity between two input trees by counting the number of common sub-structures. Different techniques have been proposed to measure such similarity. We use the Unlexicalized Partial Tree (uPT) kernel (Severyn and Moschitti, 2010) for the computation of the PET kernel since a comparative evaluation in one of our studies (Chowdhury et al., 2011) shows that uPT kernels achieve better results for RE than the other techniques used for tree kernel computation.

4.4 Less Informative Sentence and Instance Filtering

In this section we describe in detail the idea of filtering less informative sentences and less informative instances, exploiting different sources of linguistic information.

4.4.1 Exploiting the scope of negations for sentence filtering

Negation is a linguistic phenomenon where a *negation cue* (e.g. *not*) can alter the meaning of a particular text segment or of a fact. This text segment (or fact) is said to be inside the *scope of such negation (cue)*. In the context of RE, there is little work that aims to exploit the scope of negations.[20] The only work on RE that we are aware of is Sanchez-Graillet

[19] Also known as shortest path-enclosed tree (SPT) kernel.

[20] In the context of event extraction (a closely related task of RE), there have been efforts in BioNLP shared tasks of 2009 and 2011 for (non-mandatory sub-task of) event negation detection (3 participants in 2009; 2 in 2011) (Kim et al., 2009; Kim et al., 2011). The participants approached the sub-task using

and Poesio (2007) where they used various heuristics to extract negative protein interaction.

Despite the recent interest on automatically detecting the scope of negation[21], till now there seems to be no empirical evidence supporting its exploitation for the purpose of RE. Even if we could manage to obtain highly accurate automatically detected negation scopes, it is not clear how to feed this information inside the RE approach. Simply considering whether a pair of candidate mentions falls under the scope of a negation cue might not be helpful.[22]

One of the ways in which we tried take advantage of this linguistic phenomenon is the use of the negation scope related features included in the proposed K_{HF} kernel. These features are meant to help training more accurate classifiers, which will be later applied to each test instance individually. In addition, we hypothesize that a classifier trained solely on features related to the scope of negations can be used to pro-actively filter groups of instances which are less informative and mostly negative.

To be more precise, we propose to train a classifier (which will be applied before the RE classifier proposed in Section 4.3.1) that would check whether all the target entity mentions inside a sentence along with possible relation clues (or trigger words), if any, fall (directly or indirectly) under the scope of a negation cue. If such a sentence is found, then it would be identified as less informative and discarded (i.e. the candidate mention pairs inside such sentence would not be considered). During training (and testing), we group the instances by sentences. This is inspired by the multiple instance learning (MIL) technique, a ML framework that allows weak

either pre-defined patterns or some heuristics.

[21]This task is popularized by various recently held shared tasks (Farkas et al., 2010; Morante and Blanco, 2012).

[22]There exists unpublished work where unsuccessful attempts were made to exploit the scope of negation cues for RE by considering whether a pair of candidate mentions falls under a negation scope or not (Walter Daelemans, personal communication).

supervision[23]. MIL was originally introduced to solve a problem in biochemistry (Dietterich et al., 1997), but it was later adopted for some NLP problems including RE (Bunescu, 2007). In MIL, the classifier is trained on sets of positive and negative *bags* instead of sets of positive and negative *instances*. A positive bag is a set of instances which is guaranteed to contain at least one positive example. On the contrary, a negative bag is a set of instances which are all negative. *Any sentence that contains at least one relation of interest is considered by the less informative sentence (LIS) classifier as a positive (training/test) instance.* The remaining sentences are considered as negative instances.

We propose the following features to train a binary classifier that filters out less informative sentences:

- *has2TM:* The sentence has exactly 2 target entity mentions.
- *has3OrMoreTM:* The sentence has more than 2 target entity mentions.
- *allTMonRight:* All target entity mentions inside the sentence appear after the negation cue.
- *neitherAllTMonLeftOrRight:* Some but not all target entity mentions appear after the negation cue.
- *negCue:* The negation cue itself.
- *immediateGovernor:* The word on which the cue is directly syntactically dependent.
- *nearestVerbGovernor:* The nearest verb in the dependency graph on which the cue is syntactically dependent.

[23]Readers are referred to Bunescu (2007) to know how MIL can be used for weak supervision.

- *isVerbGovernorRoot:* The *nearestVerbGovernor* is root of the dependency graph of the sentence.
- *allTMdependentOnNVG:* All target entity mentions are syntactically dependent (directly/indirectly) on the *nearestVerbGovernor*.
- *allButOneTMdependentOnNVG:* All but one target entity mentions are syntactically dependent on the *nearestVerbGovernor*.
- *although*PrecedeCue:* The syntactic clause containing the negation cue begins with "although / though / despite / in spite".
- *commaBeforeNextTM:* There is a comma in the text between the negation cue and the next target entity mention after the cue.
- *commaAfterPrevTM:* There is a comma in the text between the previous target entity mention before the negation cue and the cue itself.
- *sentHasBut:* The sentence contains the word "but".

The objective of the classifier is to decide whether all target entity mentions as well as any possible evidence[24] inside the corresponding sentence fall under the scope of a negation cue in such a way that the sentence is unlikely to contain the relation of interest (e.g. DDI). If the classifier finds such a sentence, then it assigns the negative class label to it.

At present, we limit our focus only on the first occurrence of the negation cues "no", "n't" or "not". These cues usually occur more frequently and generally have larger negation scope than other negation cues.

The LIS classifier is trained using a linear SVM classifier. Its hyperparameters are tuned during training for obtaining maximum recall. In this way we minimize the number of false negatives (i.e. sentences that

[24]For which we assume the immediate and the nearest verb governors of the negation cue would be good candidates.

contain relations but are wrongly filtered out). Once the classifier is trained using the training data, we apply it on both the training and test data. However, if the recall of the LIS classifier is found to be below a *threshold value* (we set it to 70.0) during cross fold validation on the training data of a corpus, it is not used for sentence filtering of that corpus.

Any (training/test) sentence that is classified as negative is considered as a less informative sentence and is filtered. In other words, such a sentence is not considered for RE. However, it should be noted that, if such a sentence is a test sentence and it contains positive RE instances, then **all these filtered positive RE instances are automatically considered as false negatives during calculation of RE evaluation results.**

We rule out any sentence (i.e. we consider them neither positive nor negative instances) during both training and testing if any of the following conditions holds:

- The sentence contains less than two target entity mentions (such sentence would not contain the relation of interest anyway).
- It has any of the following phrases – "not recommended", "should not be" or "must not be".[25]
- There is no "no", "n't" or "not" in the sentence.
- No target entity mention appears in the sentence after "no", "n't" or "not".

To assess the effectiveness of the proposed classifier, we defined a ***baseline*** classifier that filters any sentence that contains "no", "n't" or "not".

[25]These expressions often provide clues that one of the bio-entity mentions negatively influences the level of activity of the other.

4.4.2 Discarding instances using semantic roles and contextual evidence

For identifying less informative negative instances, we exploit static (i.e. already known, heuristically motivated) and dynamic (i.e. automatically collected from the data) knowledge as described by the following criteria:

- **C1:** If each of the two entity mentions (of a candidate pair) has *anti-positive governors* (to be defined later in this section) with respect to the type of the relation, then they are not likely to be in a given relation.
- **C2:** If two entity mentions in a sentence refer to the same entity, then it is unlikely that they would have a relation between themselves.
- **C3:** If a mention is the abbreviation of another mention (i.e. they refer to the same entity), then they are unlikely to be in a relation.

Criteria C2 and C3 (static knowledge) are quite intuitive. Criterion C1 is motivated by our analyses of some randomly selected sentences from the PPI corpora (and also by some other assumptions that we will describe later in this section). For criterion C1, we construct on the fly a list of *anti-positive governors* (dynamic knowledge) taken from the training data and use them for detecting pairs that are unlikely to be in relation. As for criterion C2, we simply check whether two mentions have the same name and there is more than one character between them[26]. For criterion C3, we look for any expression of the form "Entity1 (Entity2)" and consider "Entity2" as an abbreviation or alias of "Entity1".

The above criteria are used to filter instances from both training and test data. **Any positive test instance filtered out by these criteria**

[26]In biomedical literature sometimes expressions such as "Protein1-Protein1" refer to a PPI relation. We wanted to keep mention pairs of such expressions even if the mentions have the same name.

is automatically considered as a false negative during calculation of RE evaluation results.

Anti-positive governors

The semantic roles of the entity mentions may indirectly contribute either to relate or not to relate them in a particular relation type (e.g. PPI) in the corresponding context. To put it differently, the semantic roles of two mentions in the same context could provide an indication whether the relation of interest does *not* hold between them. Interestingly, the word on which a certain entity mention is (syntactically) dependent (along with the dependency type) could often provide a clue of the semantic role of such mention in the corresponding sentence.

Our goal is to automatically identify the words (if any) that tend to prevent mentions, which are directly dependent on those words, from participating in a certain relation of interest with any other mention in the same sentence. We call such words as ***anti-positive governors*** and assume that they could be exploited to identify negative instances (i.e. negative entity mention pairs) in advance. Below we describe our approach for the automatic identification of such words.

Let $\texttt{EN}$ be the set of entity mentions such that if ${\texttt{e}^i}_s \in \texttt{EN}$ (where s indicates the corresponding training sentence and i indicates the corresponding entity mention index inside such sentence), then ${\texttt{e}^i}_s$ does not have any relation of interest (i.e. PPI) with any other mention inside the same sentence.

Let $\texttt{EP}$ be the set of entity mentions such that if ${\texttt{e}^k}_s \in \texttt{EP}$ (where s indicates the corresponding training sentence and k indicates the corresponding entity mention index inside such sentence), then ${\texttt{e}^k}_s$ has at least one relation of interest with one of the mentions inside the same sentence.

For example, consider the following sentence (taken from the IEPA

corpus) where there are three entity mention annotations – *oxytocin*[1], *oxytocin*[2] and *IP3*[3].

> *These results indicate that oTP-1 may prevent luteolysis by inhibiting development of endometrial responsiveness to* ***oxytocin***[1] *and, therefore, reduce* ***oxytocin***[2]*-induced synthesis of* ***IP3***[3] *and PGF2 alpha.*

Here, the mention *oxytocin*[1] does not participate in any PPI relation in this sentence. So, it would be included in `EN`. The other two mentions would be added to `EP`, because they are in PPI relation with each other. Note that the two mentions of the entity *oxytocin* are treated separately.

Now, let `GV` be the set of governor words where for each `w` $\in$ `GV`, *(i)* there is at least one mention $\mathtt{e}^i{}_s \in \mathtt{EN}$ which is syntactically dependent on `w` in the corresponding training sentence s, and *(ii)* there is *no* mention $\mathtt{e}^k{}_s \in \mathtt{EP}$ which is syntactically dependent on `w` in the corresponding training sentence s. We call this set `GV` as the list of *anti-positive governors*.

4.4.3 Further test instance filtering by exploiting discourse units

In this section we investigate the possibility of finding an adequate definition of discourse units such that it can be automatically exploited for filtering less informative test instances.

Elementary sentence units

Single-sentence relation instances can be split into two groups:

- **Explicit relations:** where the relation is supported by direct textual evidence inside the corresponding sentence.
- **Implicit relations:** where the relation is supported by textual evidence in other sentence(s) of the same document or by background knowledge.

We hypothesize that:

- The majority of the single-sentence relation instances might be **explicit relations**.
- Explicitly related entity mention pairs are often connected through a cue word (e.g. a verb, modifier or noun) inside a simple sentence or a simpler unit of a long complex sentence. We call these units **Elementary Sentence Units (ESUs)**. For the time being, we do not provide a definition of ESUs but we appeal to an informal concept. Hence, single sentence relations can be further categorized into **single-unit relations**[27] and **cross-unit relations**[28].
- Even if some explicitly related entity mention pairs appear in different ESUs, most of them might be treated as single-unit relations by exploiting coreference resolution.

For example, consider the following sentence where there exists an instance of PPI relation between the mentions "prion protein[4]" and "kinase[5]" and the trigger word is "blocking". All of these three appear inside a small portion of the sentence.

> *Once the abnormally phosphorylated abnormal prion protein[1] isoform agent is initiated, any stress event ensuing in adult life induces a nerve growth factor-mediated synthesis of normal cellular prion protein[2] isoform that aggregates to abnormally phosphorylated abnormal prion protein[3] isoform, thereby becoming 'infected'/transformed into the same; due to the vicious circle of positive feedback invoked by the **blocking** of a prion protein[4]-specific kinase[5].*

Another example is the following sentence where "hTAFII18[1]" interacts with "TBP[2]", "hTAFII28[4]" and "hTAFII30[5]", i.e. they are positive PPI instances. The latter two mentions appear in a different syntactic clause

[27]Relation instances where the entities participating in the relation belong to a single ESU.
[28]Relation instances where the entities participating in the relation belong to different ESUs.

(with respect to "hTAFII18[1]"). However, the pronoun "it[3]", which refers to "hTAFII18[1]", resides in the same clause as the last two mentions.

> *hTAFII18*[1] also **interacts** with *TBP*[2], but **it**[3] **interacts** more strongly with *hTAFII28*[4] and *hTAFII30*[5].

If we manage to concretely define and automatically identify *ESUs* which would comply with the above mentioned hypotheses, then they would contain a smaller number of entities than their corresponding original sentences. As a consequence, the total number of candidate entity pairs inside individual *ESUs* would be smaller than inside the original sentences. So, if only those pairs where both the candidate mentions belong to the same unit are considered as test instances, then many of the (true) negative test instances would be filtered automatically. This will reduce the number of false positives that could be mistakenly identified by the RE classifier.

However, the automatic identification of such *ESUs* from complex sentences can be in practice quite difficult.[29] There exist approaches on splitting sentences into syntactic clauses[30], but syntactic clauses (as defined in such studies) are not adequate for our purposes.

Elementary discourse units seem more appropriate for our purposes. **Elementary discourse units (EDUs)** are simple sentences or clauses within complex sentences from which discourse trees can be constructed (Marcu, 1997; Soricut and Marcu, 2003). They are the smallest identifiable structures within a discourse. Relations among EDUs are used for

[29]The closest previous work that we could identify is reported by Ding et al. (2002) where they compare abstracts, sentences and phrases as the units of text from which to extract facts (in their case, protein-protein interaction (PPI) pairs). They considered the text between any two successive punctuation marks {. : , ;} as a phrase. Interestingly, F-score (which they referred as "effectiveness") for the interaction relation of the 10 protein pairs (that they considered) in the phrasal level is not far from that in the sentence level. Another partially related work is by Thomas et al. (2011b) where they discarded patterns if two protein entities have a common ancestor node connected by the same dependency type, assuming that those proteins do not interact with each other.

[30]For example, the CoNLL 2001 shared task on clause identification – http://www.clips.ua.ac.be/conll2001/clauses/

Rhetorical structure analysis (RSA), and have applications in various text processing tasks such as text understanding, summarization, and question-answering.

The general definition of EDUs is that they are simple sentences or clauses in complex sentences from which discourse trees can be constructed (Marcu, 1997; Soricut and Marcu, 2003). Tofiloski et al. (2009) proposed the following criteria for EDUs:

- All EDUs must contain a verb.
- Adjuncts, but not complement clauses, are EDUs.
- Coordinated clauses (but not coordinated VPs), adjunct clauses with either finite or non-finite verbs, and non-restrictive relative clauses (marked by commas) are EDUs.

For defining *ESUs*, we adopt their proposal of EDUs and add additional constraints. We define that all EDUs are *ESUs* except the following:

- Non-restrictive relative clauses are not *ESUs*.
- If two EDUs are syntactically connected with each other by a preposition, then they have to be merged into a single *ESU*.

Once we split test sentences into *ESUs*, we only consider single-unit target mention pairs as candidate test instances. **Any positive test instance which is not single-unit (and is therefore filtered) is automatically considered as a false negative during calculation of RE evaluation results.**

4.5 Data

We use seven corpora for two RE tasks. All of these corpora are collections of sentences which are obtained from abstracts of biomedical liter-

ature. They contain annotation of relevant target entity mentions and corresponding single-sentence relations.

We have performed experiments for two different types of biomedical RE tasks: *(a)* protein-protein interaction (PPI), and *(b)* drug-drug interaction (DDI) extraction. The motivation is to examine *whether the impact on performance is consistent across different RE tasks.* As a matter of fact, the linguistic expressions and constructions used in biomedical literature for the description of PPIs differ from those of DDIs.

4.5.1 Data for PPI extraction

There are 5 benchmark corpora for the PPI task that are frequently used: HPRD50 (Fundel et al., 2007), IEPA (Ding et al., 2002), LLL (Nédellec, 2005), BioInfer (Pyysalo et al., 2007) and AIMed (Bunescu and Mooney, 2005b). These corpora adopt different PPI annotation formats. For a comparative evaluation Pyysalo et al. (2008) put all of them in a common format which has become the standard evaluation format for the PPI task. In our experiments, we use the versions of the corpora converted to such format[31]. The objective of using multiple corpora for the same task is to evaluate *whether the impact on performance is consistent across different corpora.* Table 4.1 shows various statistics regarding the 5 (converted) corpora.

Although all these corpora are annotated for PPI extraction, the differences in performance of the same RE system on these corpora reported by previous studies are quite dramatic. This is due to several reasons. For example, there is no general consensus regarding the guidelines to be followed when annotating PPIs. Furthermore, there are differences in the entity types considered (i.e., the PPI annotations are not just restricted to proteins). Pyysalo et al. (2008) reported their findings of quantitative and

[31] Available from http://mars.cs.utu.fi/PPICorpora/.

Corpus	Sentences	Positive pairs	Negative pairs
BioInfer	1,100	2,534	7,132
AIMed	1,955	1,000	4,834
IEPA	486	335	482
HPRD50	145	163	270
LLL	77	164	166

Table 4.1: Basic statistics of the 5 benchmark PPI corpora.

qualitative analyses of the annotations and their differences. In a different recent study, we found that the statistics of various characteristics of these five corpora indicate that there are substantial differences between the datasets (Chowdhury and Lavelli, 2012a).

4.5.2 Data for DDI extraction

For DDI, we primarily used the DDIExtraction-2011 challenge corpus[32] (Segura-Bedmar et al., 2011a). The official training and test data of the corpus contain 4,267 and 1,539 sentences, and 2,402 and 755 DDI annotations respectively.

To evaluate on a 2nd corpus, we also participated in the DDI detection and classification task of SemEval-2013[33]. The official results of the task show that our approach yields an F-score of 0.80 for DDI detection and an F-score of 0.65 for DDI detection and classification. Our system obtained significantly higher results than all the other participating teams in this shared task and has been ranked 1st. Details of our participation and results are described in Appendix E.

[32] http://labda.inf.uc3m.es/DDIExtraction2011/
[33] http://www.cs.york.ac.uk/semeval-2013/task9/

4.6 Data Pre-processing and Experimental Settings

The Charniak-Johnson reranking parser (Charniak and Johnson, 2005), along with a self-trained biomedical parsing model (McClosky, 2010), has been used for tokenization, POS-tagging and parsing of the sentences. Before parsing the sentences, all the entities are blinded by assigning names as *EntityX* where *X* is the entity index.[34] In each example, the POS tags of the two candidate entities are changed to *EntityX*. The parse trees produced by the Charniak-Johnson reranking parser are then processed by the Stanford parser[35] (Klein and Manning, 2003) to obtain syntactic dependencies according to the (collapsed) Stanford Typed Dependency format.

The Stanford parser often skips some syntactic dependencies in output. We use the following two rules to add some of such dependencies:

- If there is a "conj_and" or "conj_or" dependency between two words *X* and *Y*, then *X* should be dependent on any word *Z* on which *Y* is dependent and vice versa.
- If there are two verbs *X* and *Y* such that inside the corresponding sentence they have only the word "and" or "or" between them, then any word *Z* dependent on *X* should be also dependent on *Y* and vice versa.

Our system uses the SVM-Light-TK toolkit[36] (Moschitti, 2006; Joachims, 1999) for computation of the proposed hybrid kernels. We made minor changes in the toolkit to compute the proposed hybrid kernel. The ratio of negative and positive examples has been used as the value of the cost-ratio-factor parameter[37]. The SL kernel is computed using the jSRE

[34] This has been done only for the PPI extraction.

[35] http://nlp.stanford.edu/software/lex-parser.shtml

[36] http://disi.unitn.it/moschitti/Tree-Kernel.htm

[37] This parameter value is the one by which training errors on positive examples would outweight errors on negative examples.

tool[38].

The SVM-Light-TK toolkit is based on the SVM^{light} tool. The multi-class classification implementation of SVM^{light} (known as SVM^{multiclass}) does not support kernel combination. Hence, for multi-class classification using our system, one has to split the task into multiple binary classification tasks for multi-class or multi-label classification. For example, one can do one-class-against-the-rest classification, and then can select the highest predicted confidence score and the corresponding label for a particular instance as the final answer.

Currently, we do not perform any post-processing on output provided by the proposed RE classifier. But since the classifier assigns confidence scores for the predicted labels of the test instances, one might use these scores to select only those instances where the classifier is more confident.

Whenever required, we did statistical significance testing using *Approximate Randomization Procedure* (Noreen, 1989). We set the number of iterations to 1,000 and the confidence level to 0.01.

4.7 Experiments for PPI Extraction

We have conducted many experiments. In this section, we will report the most important ones. Our system uses a list of PPI trigger words that contains 144 words and was previously used by Bui et al. (2011) and Fundel et al. (2007). We have followed the same experimental setting commonly used for the PPI extraction task, i.e. abstract-wise 10-fold cross validation on individual corpus and one-answer-per-occurrence criterion. In fact, we have used exactly the same (abstract-wise) fold splitting of the 5 benchmark (converted) corpora that Tikk et al. (2010) used for benchmarking various kernel methods[39].

[38] http://hlt.fbk.eu/en/technology/jSRE

[39] Downloaded from http://informatik.hu-berlin.de/forschung /gebiete/wbi/ppi-benchmark .

4.7.1 Results using individual kernels

Table 4.2 compares the results obtained using Zhou et al. (2005)[40], K_{SL} and K_{PET} kernels with those obtained using our proposed K_{HF} feature vector based kernel. We implemented Zhou et al. (2005) (following the description of features given in the corresponding paper) and used it for the comparison because it is one of the state-of-the-art feature based RE approaches on the news domain and is still being used in recent studies (Sun et al., 2011; Min and Grishman, 2012). As mentioned before, we use some of the features of Zhou et al. (2005) in our proposed K_{HF}. K_{PET} has been also shown to be among the state-of-the-art RE approaches in the news domain (Nguyen et al., 2009).

Results of K_{SL} and K_{PET} are reported partly because they are shown by other studies (Tikk et al., 2010) as very competitive biomedical RE approaches, and partly because these two kernels are later used as components of our proposed hybrid kernels. So, we wanted to investigate whether the proposed kernel combinations perform better than their individual components.

As the results of Table 4.2 show, the approach proposed by Zhou et al. (2005) performs poorly in comparison to its counterparts. None of the other three approaches clearly stands out on these 5 corpora. What is noticeable though is that the proposed K_{HF} is comparatively *more robust* than the other approaches in Table 4.2 when F-scores on these corpora are compared. Furthermore, the proposed K_{HF} seems to have considerably higher precision in almost all the corpora while K_{PET} seems to obtain higher recall in most cases. The results of the proposed K_{HF} look even more attractive when the runtime (not reported here) is taken on consideration. K_{HF} is much faster than both K_{SL} and K_{PET}.

[40]We used the full feature of Zhou et al. (2005) except the WordNet features used by them.

	BioInfer	**AIMed**	**IEPA**	**HPRD50**	**LLL**
Pos. / Neg.	2,534 / 7,132	1,000 / 4,834	335 / 482	163 / 270	164 / 166
	P / R / F	**P / R / F**	**P / R / F**	**P / R / F**	**P / R / F**
K_{SL} kernel	58.7 / 69.7 / 63.7	56.8 / 63.8 / **60.1**	71.2 / 76.1 / 73.6	60.9 / 66.9 / 63.7	74.5 / 89.0 / 81.1
Using Zhou et al. (2005)	64.2 / 72.3 / 68.0	50.1 / 51.8 / 50.9	69.9 / 74.3 / 72.1	**61.1** / 49.1 / 54.4	84.2 / **93.9** / 88.8
K_{PET} kernel	61.5 / **87.2** / 72.2	46.0 / **66.8** / 54.5	71.8 / **76.7** / **74.2**	57.6 / **74.9** / **65.1**	82.2 / 92.7 / 87.1
Proposed K_{HF}	**75.5** / 79.3 / **77.4**	**57.4** / 53.4 / 55.3	**71.8** / 75.8 / 73.7	58.8 / 65.6 / 62.0	**88.4** / 92.7 / **90.5**

Table 4.2: Comparison of the results on the 5 benchmark PPI corpora using individual kernels. *Pos.* and *Neg.* refer to the total number of positive and negative instances for each of the corpora.

4.7.2 Results using proposed kernel combinations

The combination of K_{HF} and K_{SL} (i.e. K_{COMP}) results in a significant performance improvement on both BioInfer and AIMed (see Table 4.3). For the 3 small corpora, there is a considerable improvement on HPRD50 when K_{COMP} is used, while there is a decrement in results on IEPA and LLL (with respect to the results of its components, K_{HF} and K_{SL}). To be precise, K_{COMP} obtains higher precision than its components in all the PPI corpora.

When K_{COMP} is combined with K_{PET}, as the results of K_{Hybrid} show, recall improved in each of the PPI corpora. Apart from the smallest corpus LLL, the F-scores of K_{Hybrid} are higher than its three components.

As we can see from the overall results, it is difficult to choose which one between the proposed K_{COMP} and K_{Hybrid} is the better RE approach. Apart from IEPA, none of their F-score differences is statistically significant. The numbers in Table 4.3 hint that K_{Hybrid} probably has a thin edge over K_{COMP}. We wanted to understand how much this minor gain by K_{Hybrid} is worth if we take runtime in consideration. It appears K_{Hybrid}

	BioInfer	**AIMed**	**IEPA**	**HPRD50**	**LLL**
Pos. / Neg.	2,534 / 7,132	1,000 / 4,834	335 / 482	163 / 270	164 / 166
	P / R / F	**P / R / F**	**P / R / F**	**P / R / F**	**P / R / F**
Individual components					
K_{SL} kernel	58.7 / 69.7 / 63.7	56.8 / 63.8 / 60.1	71.2 / 76.1 / 73.6	60.9 / 66.9 / 63.7	74.5 / 89.0 / 81.1
K_{PET} kernel	61.5 / 87.2 / 72.2	46.0 / 66.8 / 54.5	71.8 / 76.7 / 74.2	57.6 / 74.9 / 65.1	82.2 / 92.7 / 87.1
Proposed K_{HF}	75.5 / 79.3 / 77.4	57.4 / 53.4 / 55.3	71.8 / 75.8 / 73.7	58.8 / 65.6 / 62.0	88.4 / 92.7 / 90.5
Proposed combinations					
Proposed K_{COMP} (i.e. $K_{HF} + K_{SL}$)	**84.6** / 79.4 / 81.9	**64.4** / 62.1 / **63.2**	71.9 / 72.5 / 72.2	**62.4** / 74.2 / **67.8**	89.4 / 82.3 / 85.7
Proposed K_{Hybrid} (i.e. $K_{HF} + K_{SL} + K_{PET}$)	83.8 / **81.1** / **82.4**	63.1 / **62.9** / 63.0	**74.3** / **77.6** / **75.9**	58.8 / **76.1** / 66.3	**91.3** / **83.5** / **87.3**

Table 4.3: Comparison of the results on the 5 benchmark PPI corpora using proposed K_{COMP} and K_{Hybrid} kernels. *Pos.* and *Neg.* refer to the total number of positive and negative instances for each of the corpora.

requires at least twice more runtime in each of the corpora in comparison to K_{COMP} (see Figure 4.5). To put in another way, the gain in recall along with the gain in F-score (in some of the corpora) due to the addition of the tree kernel (i.e. K_{PST}) comes at a cost of much slower runtime.

4.7.3 Results using sentence and instance filtering

In this section, we report the outcome when each of our three proposed techniques for less informative sentence and instance filtering are used separately. The experiments are done using both of the proposed kernel compositions, i.e. K_{COMP} and K_{Hybrid}.

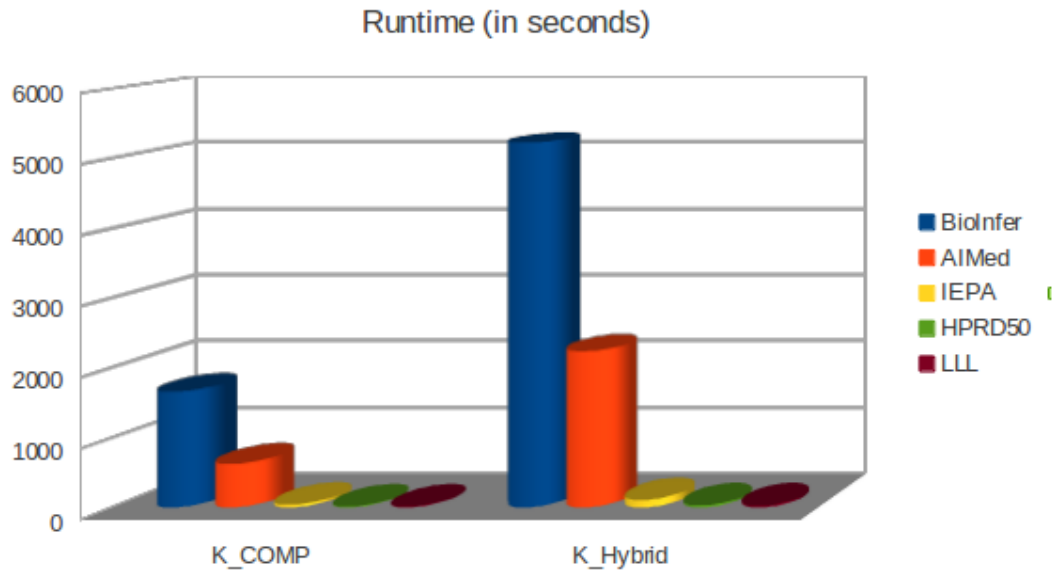


Figure 4.5: Comparison of the runtime on the 5 benchmark PPI corpora using proposed K_{COMP} and K_{Hybrid} kernels.

Less informative sentence filtering

Table 4.4 shows the results when less informative sentence (LIS) filtering (by exploiting the scope of negations) is employed. For AIMed, IEPA, HPRD50 and LLL, the recall of the classifier was below the threshold value (70.0), so no sentences were filtered from them.

On BioInfer, the usage of the baseline LIS classifier (i.e. filtering sentences that have any of the words – "no", "not" or "n't") deteriorated the results due to a considerable decrement of the recall. Our proposed LIS classifier does not improve the results either. There are minor drops in F-scores which are statistically insignificant. But what is interesting is that the usage of the proposed LIS classifier enables to obtain almost the same performance using fewer training instances and at a faster runtime. Table 4.8 shows the reduction of the total number of instances (training and test combined) due to the usage of the LIS classifier. This reduction cuts back the imbalance in data distribution.

	BioInfer	**AIMed**	**IEPA**	**HPRD50**	**LLL**
Pos. / Neg.	2,534 / 7,132	1,000 / 4,834	335 / 482	163 / 270	164 / 166
	P / R / F	**P / R / F**	**P / R / F**	**P / R / F**	**P / R / F**
K_{COMP}	**84.6** / 79.4 / **81.9**	64.4 / 62.1 / 63.2	71.9 / 72.5 / 72.2	62.4 / 74.2 / 67.8	89.4 / 82.3 / 85.7
Baseline LIS clas. + K_{COMP}	83.7 / 70.2 / 76.4	NA	NA	NA	NA
Proposed LIS clas. + K_{COMP}	82.9 / **79.7** / 81.3	NA	NA	NA	NA
K_{Hybrid}	**83.8 / 81.1 / 82.4**	63.1 / 62.9 / 63.0	74.3 / 77.6 / 75.9	58.8 / 76.1 / 66.3	91.3 / 83.5 / 87.3
Baseline LIS clas. + K_{Hybrid}	85.4 / 69.9 / 76.9	NA	NA	NA	NA
Proposed LIS clas. + K_{Hybrid}	83.0 / 80.9 / 81.9	NA	NA	NA	NA

Table 4.4: Comparison of results on the 5 PPI corpora after using proposed techniques for filtering less informative sentences by exploiting scopes of negations.

To understand why our proposed approach is not effective on these 5 corpora, we collected some statistics from these corpora which are shown in Table 4.5. As we can see, the total number of sentences that satisfy our proposed criteria to be eligible as training and test instances (for sentence filtering using negation scopes) is very low. So, it is not surprising that the recall of the LIS classifier in 4 out of the 5 corpora is not high enough to be considered for filtering sentences.

Although the recall did cross the threshold (70.0) on BioInfer, the learning of the LIS classifier was highly biased. Among the 90 eligible sentences of BioInfer, only 30 do not contain any PPI relations. That means there are twice more positive instances (for the LIS classifier) with respect to the negative instances, which is a highly unusual NLP data distribution. Hence, the less informative sentence filtering was not helpful for BioInfer either. Later in this chapter (while discussing DDI extraction), we would

	BioInfer	**AIMed**	**IEPA**	**HPRD50**	**LLL**
Total sentences eligible as training and test instances for the sentence filtering experiments using the proposed LIS classifier	90	85	40	13	1

Table 4.5: Total number of sentences in each of the PPI corpora that satisfy our proposed criteria to be eligible as training and test instances for sentence filtering using negation scopes.

show that when the number of eligible sentences is not so small and the ratio of the distribution of positive instances of the LIS classifier with respect to that of its negative instances is not unusually high, our proposed approach for less informative sentence filtering does significantly improve the results.

Less informative instance filtering

Table 4.6 shows the results when the proposed dynamic knowledge (collected by exploiting intuition of semantic roles) and static knowledge are used for filtering less informative instances (LII). As we can see, the differences in F-score vary from corpus to corpus. Our proposed technique evidently appears helpful on IEPA and HPRD50. On AIMed, it slightly improves the F-score in case of K_{Hybrid} and has no effect on F-score in case of K_{COMP}. In all of these three corpora, recall is boosted thanks to the usage of our technique.

On LLL, the smallest PPI corpus, F-score drops. We believe this is due to the creation of imbalance in distribution. Unlike other corpora, LLL is a balanced corpus with an almost equal number of positive and negative instances. So, when the number of negative instances decreased due to our filtering technique, it hurts the performance.

The increments of F-scores on IEPA and HPRD50 are statistically sig-

	BioInfer	**AIMed**	**IEPA**	**HPRD50**	**LLL**
Pos. / Neg.	2,534 / 7,132	1,000 / 4,834	335 / 482	163 / 270	164 / 166
	P / R / F	**P / R / F**	**P / R / F**	**P / R / F**	**P / R / F**
K_{COMP}	84.6 / **79.4** / **81.9**	**64.4** / 62.1 / **63.2**	71.9 / **72.5** / 72.2	62.4 / 74.2 / 67.8	89.4 / **82.3** / **85.7**
Baseline for LII filt. + K_{COMP}	**84.8** / 68.7 / 75.9	**64.4** / 61.1 / 62.7	71.9 / **72.5** / 72.2	**63.2** / 73.6 / 68.0	**90.5** / 81.1 / 85.5
Proposed LII filt. + K_{COMP}	83.1 / 79.1 / 81.1	60.3 / **66.5** / **63.2**	**73.6** / **72.5** / **73.1**	61.1 / **81.6** / **69.8**	89.7 / 79.9 / 84.5
K_{Hybrid}	**83.8 / 81.1 / 82.4**	**63.1** / 62.9 / 63.0	74.3 / 77.6 / 75.9	58.8 / 76.1 / 66.3	91.3 / **83.5** / **87.3**
Baseline for LII filt. + K_{Hybrid}	83.0 / 75.2 / 78.9	63.0 / 42.5 / 50.7	74.3 / 77.6 / 75.9	60.0 / 75.5 / 66.9	**92.5** / 82.3 / 87.1
Proposed LII filt. + K_{Hybrid}	83.6 / 80.1 / 81.8	58.5 / **68.9** / **63.3**	**76.2** / **78.5** / **77.4**	**62.8** / **85.9** / **72.5**	92.3 / 79.9 / 85.6

Table 4.6: Comparison of results on the 5 PPI corpora after using proposed techniques for filtering less informative instances by using dynamic and static knowledge.

	BioInfer	**AIMed**	**IEPA**	**HPRD50**	**LLL**
Pos. / Neg.	2,534 / 7,132	1,000 / 4,834	335 / 482	163 / 270	164 / 166
	P / R / F	**P / R / F**	**P / R / F**	**P / R / F**	**P / R / F**
K_{COMP}	84.6 / **79.4** / **81.9**	64.4 / **62.1** / 63.2	71.9 / **72.5** / 72.2	62.4 / **74.2** / 67.8	89.4 / **82.3** / 85.7
Proposed ESUs + K_{COMP}	**84.7** / 78.4 / 81.4	**64.8** / 62.0 / **63.4**	**77.9** / **72.5** / **75.1**	**63.0** / **74.2** / **68.2**	**90.6** / **82.3** / **86.3**
K_{Hybrid}	**83.8** / **81.1** / **82.4**	63.1 / **62.9** / **63.0**	74.3 / **77.6** / **75.9**	58.8 / **76.1** / 66.3	91.3 / **83.5** / 87.3
Proposed ESUs + K_{Hybrid}	83.7 / 80.2 / 81.9	**63.8** / 61.8 / 62.8	**74.8** / 77.0 / **75.9**	**59.1** / **76.1** / **66.5**	**92.5** / 82.9 / **87.5**

Table 4.7: Comparison of results on the 5 PPI corpora after filtering test sentences by exploiting proposed elementary sentence units.

	BioInfer	**AIMed**	**IEPA**	**HPRD50**	**LLL**
Pos. / Neg.	2,534 / 7,132	1,000 / 4,834	335 / 482	163 / 270	164 / 166
Only after using scope of negations for sentence filtering, i.e. using proposed LIS classifier					
Reduction of positive instances	3.95%	–	–	–	–
Reduction of negative instances	12.31%	–	–	–	–
Only after using proposed LII filtering, i.e. using dynamic and static knowledge					
Reduction of positive instances	2.46%	0.60%	0.60%	0.61%	1.83%
Reduction of negative instances	9.22%	20.18%	24.07%	26.30%	19.88%
Only after using proposed ESUs for test instance filtering					
Reduction of positive instances	2.17%	0.90%	1.49%	0.00%	1.83%
Reduction of negative instances	5.58%	10.28%	7.05%	2.59%	4.82%

Table 4.8: Percentage of the decrease in the number of instances for the proposed techniques.

nificant, while the differences (increment or decrement) on the other three PPI corpora are not statistically significant.

Regarding the impact of the baseline classifier for identifying less informative instances (i.e. the technique proposed by (Sun et al., 2011)), apart from a small change in HPRD50, in none of the other PPI corpora there was any improvement in F-score. In fact, the performance dropped sharply in the two biggest PPI corpora (i.e. BioInfer and AIMed).

What we would like to underline is that the proposed strategy for less informative instance filtering enables to obtain almost the same (in some cases) or better performance (in other cases) at a cost of much shorter runtime and a smaller number of training instances (see Table 4.8).

Table 4.7 shows the results when we exploited elementary sentence units (which are derived from discourse units) to filter less informative instances from the test data. Since no instance is filtered from the training data, the data imbalance remains the same. According to the results, it appears that the usage of our proposed filtering improves precision in almost all

	BioInfer	**AIMed**	**IEPA**	**HPRD50**	**LLL**
Pos. / Neg.	2,534 / 7,132	1,000 / 4,834	335 / 482	163 / 270	164 / 166
	P / R / F	**P / R / F**	**P / R / F**	**P / R / F**	**P / R / F**
APG kernel Airola et al. (2008)	56.7 / 67.2 / 61.3	52.9 / 61.8 / 56.4	69.6 / 82.7 / 75.1	64.3 / 65.8 / 63.4	72.5 / 87.2 / 76.8
Multiple kernels and multiple parser input Miwa et al. (2009a)	65.7 / 71.1 / 68.1	55.0 / 68.8 / 60.8	67.5 / 78.6 / 71.7	**68.5** / 76.1 / 70.9	77.6 / 86.0 / 80.1
SVM-CW, multiple parser input and graph, walk and BOW features Miwa et al. (2009b)	– / – / 67.6	– / – / **64.2**	– / – / 74.4	– / – / 69.7	– / – / 80.5
kBSPS kernel Tikk et al. (2010)	49.9 / 61.8 / 55.1	50.1 / 41.4 / 44.6	58.8 / **89.7** / 70.5	62.2 / **87.1** / 71.0	69.3 / **93.2** / 78.1
Walk weighted subsequence kernel Kim et al. (2010)	61.8 / 54.2 / 57.6	**61.4** / 53.3 / 56.6	73.8 / 71.8 / 72.9	66.7 / 69.2 / 67.8	76.9 / 91.2 / 82.4
2 phase extraction Bui et al. (2011)	61.7 / 57.5 / 60.0	55.3 / 68.5 / 61.2	– / – / –	– / – / –	– / – / –
Proposed K_{Hybrid} with instance filtering using dynamic and static knowledge	**83.6** / **80.1** / **81.8**	58.5 / **68.9** / 63.3	**76.2** / 78.5 / **77.4**	62.8 / 85.9 / **72.5**	**92.3** / 79.9 / **85.6**

Table 4.9: Comparison of the results on the 5 benchmark PPI corpora. *Pos.* and *Neg.* refer to the number of positive and negative relations respectively. The results of Bui et al. (2010) on LLL, HPRD50, and IEPA are not reported since thy did not use all the positive and negative examples during cross validation. As for Miwa et al. (2009b), we consider only those results of their experiments where they used a single training corpus, as it is the standard evaluation approach adopted by all the other studies on PPI extraction for comparing results. All the results of the previous approaches reported in this table are directly quoted from their respective original papers. We use exactly the same folds that are used by Tikk et al. (2010).

cases. There are minor and statistically insignificant decrements of F-score in BioInfer and AIMed. For K_{COMP}, F-score improved in the three smaller corpora. But there is virtually no improvement when K_{Hybrid} is used. It appears that the structural similarity features (of the K_{PET}) are quite effective to accurately identify cross-sentence unit relations which have been filtered. As a result, such filtering negatively affected the performance because of the filtering of some positive cross-sentence unit relation instances which, otherwise, could have been correctly extracted using structural similarity features.

4.7.4 Why is there so much discrepancy in performance?

There are two things to notice from all these results: firstly, the variation of performance on the different corpora for both K_{Hybrid} and K_{COMP}, and secondly, the variation of the impact of reducing skewness in distribution on these corpora. Previous studies stated that there are definite limits on the ability to compare different RE approaches evaluated on these different PPI corpora (Pyysalo et al., 2008). Nevertheless, we make an attempt below to understand the possible causes of such discrepancies on these PPI corpora.

	Corpora				
Description	**LLL**	**IEPA**	**HPRD50**	**AIMed**	**BioInfer**
Avg. no. of words between each target entity pair	10.46	8.89	7.11	6.92	8.44
Avg. no. of words per target entity name	1.05	1.22	1.21	1.29	1.24
Avg. no. of target entities per sentence	3.10	2.30	2.79	3.25	4.05
Avg. no. of words in (all) target entity names per sent.	3.26	2.80	3.38	4.19	5.03
Excluding target entities avg. no. of words per sent.	22.57	26.07	20.93	19.71	21.93
Avg. no. of words per sentence	25.83	28.87	24.31	23.90	26.96

Table 4.10: Statistics of different characteristics of the 5 benchmark PPI corpora. All sentences (in each corpus) are considered during analyses.

We collected statistics of different characteristics of the 5 benchmark PPI corpora which are shown in Table 4.10. As we can see, the target entity names in LLL are smaller and there is a reasonably higher average number of words between each candidate entity pair. Furthermore, the instance distribution in LLL is already balanced (positive = 164, negative = 166). So, it is probably relatively easier to extract PPIs from LLL with simpler techniques. The reduction of negative training instances on LLL is only going to hurt the performance as our results indicate. In fact, almost all the previous state-of-the-art RE approaches for PPI obtained quite high results on LLL but often failed to perform equally well on the other corpora.

Both IEPA and BioInfer also have a relatively higher average number of words between each candidate entity pair. The average length of sentences (i.e. number of total words) in these two corpora is also higher than that of the other corpora. This means that they contain probably sufficient context to include cue words/phrases relevant for PPIs, which led to a high performance on them. Interestingly, phrase structural syntactic features seem to be very useful in these corpora (see results obtained using K_{HF} and K_{PET}). While our proposed instance filtering techniques were useful for IEPA, they totally failed to have any impact (on performance) on BioInfer. Our investigation (summarized in the next paragraph) suggests that this stems from the peculiar annotation guidelines of BioInfer.

The first peculiarity that we observed in the BioInfer corpus is that 2.19% of its PPIs (i.e. positive instances) are between entity mentions having the same name. The only other corpus which has such annotations is AIMed, but only 0.20% of its PPIs . So, although the criterion C1 of the proposed static knowledge discarded 6.69% negative instances in BioInfer, perhaps it was not enough to counter the loss of information due to the discarded positive instances. Another peculiarity is that the usage of anti-

positive governors (criterion C2; proposed dynamic knowledge) actually discarded positive instances in BioInfer and failed to filter any negative instance. To check why it is so, we extracted the list of anti-positive governors from the whole BioInfer corpus (total 1,100 sentences) and found there are only 10 such words. By comparison, the number of anti-positive governors in AIMed (total 1,955 sentences) and IEPA (total 486 sentences) are 300 and 161 respectively. Further investigation revealed that there are startling differences for the concentration of PPIs/sentence between BioInfer and the other corpora. For BioInfer, it is 2.30 PPIs/sentence. If we compare this with AIMed and IEPA then the respective numbers are 0.51 PPIs/sentence and 0.70 PPIs/sentence. As a result, it is quite difficult to spot a word which is not governing any mention that participates in PPI and which is only governing those mentions that are not in any PPI in the corresponding sentence.

The F-scores on AIMed and HPRD50 are somewhat similar for K_{Hybrid} and K_{COMP}. The average number of words between each candidate entity pairs and the average length of sentences in these two corpora are quite close. But the imbalance in data distribution in AIMed is much larger. Interestingly, recall on AIMed is significantly lower than that on HPRD50. We suspect that the data imbalance might have a correlation with the recall in AIMed. Because after the reduction of skewness in the training data (by using dynamic and static knowledge), a considerable improvement of recall in AIMed has been observed.

4.7.5 Comparisons with the state-of-the-art results

Based on the analysis described in the earlier sections, we argue that, among the proposed filtering techniques, only the usage of static and dynamic knowledge is effective for PPI extraction. So, we compare the results of our proposed approach (i.e. K_{Hybrid} with instance filtering using dynamic

and static knowledge) with that of the existing state-of-the-art approaches.

As we can see in Table 4.9, on 4 out of the 5 benchmark PPI corpora (i.e. except only AIMed), our proposed approach outperforms the previous best results. In the case of AIMed, our RE approach obtains slightly lower result than the best result, obtained by Miwa et al. (2009b). But the results of their system in the other corpora are not as good as ours.

4.8 Experimental Results for DDI extraction

We have also conducted a number of experiments for the other RE task, DDI extraction. In this section, we will report the most important ones among these. We are unaware of any available list of trigger words for DDI. So, we created such a list (which will be made publicly available along with the RE system developed for this PhD thesis).

Some sentences in the DDIExtraction-2011 corpus contain unknown symbols (can be identified by the presence of question (*?*) marks), perhaps due to encoding problems. This fact could produce negative effects on the processing performed by our system. Given that we have no access to the original documents, we replace such symbols with "@". If an entity name does not contain space characters immediately before and after its boundaries, space characters are inserted automatically in such positions to avoid tokenization errors. These pre-processing were done before parsing the data.

4.8.1 Results using individual kernels and kernel combinations

As shown in Table 4.11, the proposed K_{HF} obtains higher precision than the other approaches – Zhou et al. (2005), K_{SL} and K_{PET}. However, K_{SL} obtains considerably higher recall, due to which it outperforms K_{HF} in F-score. The combination of K_{HF} and K_{SL} (i.e. K_{COMP}) results in a sig-

	P	R	F-score
Individual components			
K_{SL}	55.1	**73.1**	**62.9**
Using Zhou et al. (2005)	57.1	40.0	47.0
K_{PET}	51.7	64.1	57.2
Proposed K_{HF}	**58.5**	56.2	57.3
Proposed combinations			
Proposed K_{COMP} (i.e. $K_{HF} + K_{SL}$)	57.8	**75.9**	65.6
Proposed K_{Hybrid} (i.e. $K_{HF} + K_{SL} + K_{PET}$)	**60.0**	74.3	**66.4**

Table 4.11: Comparison of results on the official test set of the 2011 DDI Extraction challenge using the proposed K_{COMP} and K_{Hybrid} kernels as well as their individual components.

nificantly better F-score. These improvements are statistically significant. If K_{PET} is combined with K_{COMP}, there is a further increase in precision. Consequently, there is more improvement in F-score.

The non-target entity specific features that we proposed as part of the feature set of K_{HF} are only used during DDI extraction.[41] We found that these features improve recall (by 0.9 points) and F-score (by 0.3 points). However, these improvements are not statistically significant. The usage of negation scope related features improves recall (by 1.1 points) and F-score (by 0.9 points) as well.

4.8.2 Results using sentence and instance filtering

According to the experimental results (see Rows 2-7 Table 4.12), less informative sentence filtering is found to be very effective for DDI extraction. It

[41]We knew that appearance of disease/symptoms inside text, where co-administration of two drugs are mentioned, could provide clue for potential DDI. In case of PPI, it was not clear to us whether there is any such specific non-target entity for PPI that could provide hints.

improves the F-scores of K_{COMP} and K_{Hybrid} by 1.4 and 1.0 points respectively, which are found to be statistically significant. These improvements are due to the encouraging increase in precision (more than 2 points in each case).

K_{Hybrid} seems to benefit at most from the usage of the proposed static and dynamic knowledge for less informative instance filtering (see Rows 8-13 Table 4.12). The improvement of F-score for K_{Hybrid} is statistically significant, but not for K_{COMP}.

The usage of elementary sentence units to filter less informative test instances (see Rows 14-17 Table 4.12) boosted precision (by 3.7 points) and F-score (by 1.1 points) of K_{COMP}. These increments are statistically significant. But this technique failed to make an impact on the outcome of K_{Hybrid}, similar to the findings for PPI extraction.

As it was mentioned during the discussion on PPI extraction results, sentence filtering using negation scope was not successful in any of the PPI corpora because the total number of instances (for 10-fold cross validation experiments) for the LIS classifier was quite small. Furthermore, even in this small amount of data in some of these corpora, the number of positive instances was unusually high.

By contrast, sentence filtering using negation scope has been very effective for DDI extraction. As we can see from Table 4.14, the total number of instances for the LIS classifer is 607; almost 3 times higher than that of all the 5 PPI benchmark corpora combined. We found that the ratio of positive and negative instances for the LIS classifier in training and test data are 1:4.6 and 1:5.6 respectively, which is quite typical since negative instances are usually found more often in NLP data.

	P	**R**	**F-score**
K_{COMP}	57.8	**75.9**	65.6
Baseline LIS classifier + K_{COMP}	59.8	72.1	65.3
Proposed LIS classifier + K_{COMP}	**60.0**	**75.9**	**67.0**
Proposed K_{Hybrid}	60.0	**74.3**	66.4
Baseline LIS classifier + K_{Hybrid}	61.8	68.9	65.1
Proposed LIS classifier + K_{Hybrid}	**62.1**	73.8	**67.4**
K_{COMP}	57.8	75.9	65.6
Baseline LII filtering + K_{COMP}	**58.4**	67.3	62.5
Proposed LII filtering + K_{COMP}	54.9	**82.0**	**65.8**
Proposed K_{Hybrid}	60.0	74.3	66.4
Baseline LII filtering + K_{Hybrid}	58.8	66.8	62.5
Proposed LII filtering + K_{Hybrid}	**61.1**	**75.1**	**67.4**
K_{COMP}	57.8	**75.9**	65.6
Proposed ESUs + K_{COMP}	**61.5**	72.9	**66.7**
Proposed K_{Hybrid}	**60.0**	**74.3**	**66.4**
Proposed ESUs + K_{Hybrid}	59.9	74.0	66.2

Table 4.12: Comparison of results on the official test set of the 2011 DDI Extraction challenge after using each of the proposed techniques for filtering less informative sentences and instances. The LIS classifier and its baseline are described in Section 4.4.1. The proposed approaches for LII filtering and its baselines are described in Section 4.4.2. Section 4.4.3 includes details regarding how proposed ESUs are exploited.

	Training data	**Test data**
Pos. / Neg.	2,402 / 21,377	755 / 6,271
Only after using scope of negations for sentence filtering, i.e. using proposed LIS classifier		
Reduction of positive instances	0.54%	0.66%
Reduction of negative instances	0.28%	14.77%
Only after using proposed LII filtering, i.e. using dynamic and static knowledge		
Reduction of positive instances	0.92%	0.66%
Reduction of negative instances	6.87%	6.36%
Only after using proposed ESUs for test instance filtering		
Reduction of positive instances	–	0.93%
Reduction of negative instances	–	1.85%

Table 4.13: Percentage of the decrease in the number of instances for the proposed techniques on the 2011 DDI Extraction challenge data.

	Training data	**Test data**
Total sentences eligible for the sentence filtering experiments using the proposed LIS classifier	455	152

Table 4.14: Total number of sentences in the 2011 DDI Extraction challenge corpus eligible as training and test instances for sentence filtering using negation scopes.

4.8.3 Comparisons with the state-of-the-art results

Table 4.15 compares the performance of our proposed approach with the previously reported best results. Both precision and recall of our approach are significantly higher than the reported best result. Consequently, our approach outperforms previous approaches in F-score (3.2 points higher than the next best).

The joint exploitation of less informative sentence and instance filtering improves the F-score by 2.5 points, which is quite encouraging, with respect to using the proposed K_{Hybrid} kernel alone without these filterings. This

	P	R	F-score
Ensemble of multiple ML based methods (Thomas et al., 2011a)	60.5	71.9	65.7
Ensemble of two ML based methods (Chowdhury et al., 2011c)	58.6	70.5	64.0
Combination of multiple kernels (Chowdhury and Lavelli, 2011b)	58.4	70.1	63.7
Regularized least-squares classifiers (Björne et al., 2011)	58.0	68.9	63.0
Proposed K_{Hybrid} with (i) sentence filtering using negation scopes, and (ii) instance filtering using dynamic and static knowledge	**63.5**	**75.2**	**68.9**

Table 4.15: Comparison of the results of our proposed approach with the previous state-of-the-art approaches, obtained on the official test set of the 2011 DDI Extraction challenge.

improvement is statistically significant.

4.9 Additional Experiments

We performed many other experiments, whose results are not reported here, which include experiments after instance filtering in training data by exploiting sentence units, less informative sentence filtering only from training data (and also only from test data), less informative instance filtering (using dynamic and static knowledge) only from training data (and also only from test data), etc. The general trend is that filtering only test instances has more impact on results improvement than filtering training instances only.

4.10 Limitations and Future Work

While our proposed hybrid kernels proved to be very robust (on the basis of the achieved state-of-the-art results) for different tasks on multiple corpora, there is further room of improvement. The strength of our proposed approach is that it obtains much higher recall than the other approaches.

But in case of precision, for most of the corpora on which we tested our proposed approach, it could not (in general) outperform other approaches in terms of precision. So, any future extension should focus on this issue.

Our assumption was that the exploitation of the proposed elementary sentence units would strengthen the precision by reducing false positives and actually it did for K_{COMP} (see Tables 4.7 and 4.12). But unfortunately, it was not helpful for K_{Hybrid}. More investigation is required to re-adjust the definition of elementary sentence units for RE.

Our proposed techniques for reducing data imbalance brought mixed results. They were very effective on DDI and some of the PPI corpora. But they were found ineffective on BioInfer, one of the biggest PPI corpora. We already mentioned several reasons for such disparity. Further investigation may be required to understand the influence of the annotation guidelines and their impact on RE.

The proposed hybrid kernels are very dependent on the proposed feature based kernel, K_{HF}. We assume the usage of automatically collected paraphrases to generate new useful features could contribute to improve precision. Weakly supervised collection of paraphrases for RE has already been investigated (Romano et al., 2006).

Chapter 5

Conclusion

> *"There may be more text data in electronic form than ever before, but much of it is ignored. No human can read, understand, and synthesize megabytes of text on an everyday basis. Missed information – and lost opportunities – has spurred researchers to explore various information management strategies to establish order in the text wilderness."*
>
> – Jim Cowie and Wendy Lehnert
>
> "Information Extraction", Communications of the ACM, 39(1):80–91. (1996)

5.1 Summary

With escalating health care costs in most countries, it is important to develop precision healthcare technologies that will be evidence-based, patient-centred, pro-active and preventive. To be used in real time, these technologies also need to be scalable, sustainable and fast. The success of these new technologies will greatly depend on the advancements of the core IE tasks.

Keeping this in mind, throughout this PhD research, we examined a number of different techniques for the improvement of the current state of NER, coreference resolution and RE research in biomedical domain. For

each of these tasks, we approached some novel research questions which can push the barrier ahead for the ultimate goal of the accurate extraction of complex biological knowledge hidden inside biomedical literature and other form of biomedical text. Some of our proposed techniques proved to be very effective and, consequently, we were able to obtain state-of-the-art results. However, there are a few of our proposed techniques which either did not have any impact or hurt the performance, defying our hypothesis. In the corresponding chapters we make an attempt to discuss and to reason on the success and failure of all these techniques.

Although our solutions are (machine learning based) supervised approaches, we did exploit the outcomes produced in other computational linguistics fields (the scope of negation cues, the discourse units of sentences and the semantic roles of entity mentions) without the need of annotated data. Almost all of our proposed techniques for the three IE tasks are largely domain independent. On the long run, we hope that the research described in this thesis would be useful towards building an integrated information extraction approach that is robust and accurate not only on the biomedical domain but also on other genres of text.

5.2 Possible Future Extensions

5.2.1 Named entity recognition

As described in Chapter 2, there are at least two potential topics which need further investigation as they could further improve the performance of biomedical named entity recognition. One of them concerns the selection of appropriate syntactic dependency types to identify links between cue words and probable entity names (see Section 2.3.4). Another topic is the attempt of exploiting dictionaries automatically created from training data during post-processing since the usage of global perspective features

showed that such dictionaries could considerably increase the number of true positives (see Section 2.3.4).

During the study of the exploitation of silver standard corpus annotation for NER, we found that if only annotated sentences (which we call *condensed* corpus) are considered, then the number of annotations as well as the performance increases significantly (see Section 2.6.4). However, it appears that correctly unannotated sentences influence the achievement of high precision. One extension of our study concerns the investigation of a more sophisticated approach for discarding only a part of the unannotated sentences instead of discarding them completely. Measuring lexical similarity between annotated and unannotated sentences might help in this case.

The outcome of our preliminary study of the practical usability of silver standard corpus annotation for NER is encouraging. An interesting perspective concerns a similar investigation with respect to other NLP tasks.

5.2.2 Coreference resolution

The extension of the research described in Chapter 3 would imply to exploit the combination of the proposed linguistically and semantically motivated constraints for coreference resolution on biomedical text. Some of these constraints (used for clinical text) might not be effective on biomedical text but the general idea behind controlling the generation of less-informative/sub-optimal training and test instances could be useful. The greedy clustering of mention pairs proposed in our study can be used on biomedical text with only some minor modifications.

5.2.3 Relation extraction

While the hybrid kernels proposed in Chapter 4 proved to be very robust for different RE tasks on multiple corpora, there is further room for

improvement, especially for obtaining much higher precision. We assume the usage of automatically collected paraphrases to generate new useful features could contribute to improve precision. Weakly supervised collection of paraphrases for RE has already been investigated (Romano et al., 2006). So, inclusion of such techniques in our proposed approach could be a possible extension.

Our assumption was that the exploitation of the proposed elementary sentence units would strengthen the precision by reducing false positives. However, it was not so helpful when a tree kernel was incorporated inside our proposed hybrid kernel (see Tables 4.7 and 4.12). More investigation is required to re-adjust the definition of elementary sentence units for RE.

Finally, although we partially exploited our proposed NER approach for RE (see Section 4.3.2), we did not perform end-to-end information extraction (i.e. evaluating relation extraction on test data when gold-standard named entities and coreferences are not provided). This could be another potential extension.

Appendix A

Features Selected from Zhou et al. (2005)

The 27 features that we selected from Zhou et al. (2005), as part of the feature set of our system, are described below. Here, $M1$ and and $M2$ refer to any two target entity mentions which form a candidate pair/instance.

- *WBNULL:* when no word in between M1 and M2
- *WBFL:* the only word in between when only one word in between M1 and M2
- *WBF:* first word in between when at least two words in between M1 and M2
- *WBL:* last word in between when at least two words in between M1 and M2
- *WBO:* other words in between except first and last words when at least three words in between M1 and M2
- *BM1F:* first word before M1
- *BM1L:* second word before M1
- *AM2F:* first word after M2
- *AM2L:* second word after M2

- *#MB:* number of other mentions in between M1 and M2
- *#WB:* number of words in between M1 and M2
- *CPHBNULL:* when no phrase in between M1 and M2
- *CPHBFL:* the only phrase head when only one phrase in between M1 and M2
- *CPHBF:* first phrase head in between when at least two phrases in between M1 and M2
- *CPHBL:* last phrase head in between when at least two phrase heads in between M1 and M2
- *CPHBO:* other phrase heads in between except first and last phrase heads when at least three phrases in between M1 and M2
- *CPHBM1F:* first phrase head before M1
- *CPHBM1L:* second phrase head before M1
- *CPHAM2F:* first phrase head after M2
- *CPHAM2L:* second phrase head after M2
- *CPP:* path of phrase labels connecting M1 and M2 in the chunking
- *CPPH:* path of phrase labels connecting M1 and M2 in the chunking augmented with head words, if at most two phrases in between
- *ET12SameNP:* combination of the types of M1 and M2 with whether M1 and M2 included in the same NP
- *ET12SamePP:* combination of the types of M1 and M2 with whether M1 and M2 exist in the same PP

- *ET12SameVP:* combination of the types of M1 and M2 with whether M1 and M2 included in the same VP
- *PTP:* path of phrase labels (removing duplicates) connecting M1 and M2 in the parse tree
- *PTPH:* path of phrase labels (removing duplicates) connecting M1 and M2 in the parse tree augmented with the head word of the top phrase in the path.

Appendix B

Preliminary Results of Our Proposed RE Approach on News Domain

	Zhou et al. (2005) (as reported in Sun et al. (2011))	**Sun et al. (2011)**	**Nguyen and Moschitti (2011b)**	$\mathbf{K}_{Hybird}$ (our proposed kernel for RE)
EMP-ORG	77.6	79.3	82.8	**84.1**
OTHER-AFF	52.2	54.6	–	**80.0**
GPE-AFF	63.3	65.9	**76.9**	72.3
PHYS	66.9	65.3	69.5	**74.4**
PER-SOC	70.3	70.7	–	**82.4**
ART	73.4	**79.3**	–	74.4
DISC	55.7	58.6	–	**72.9**

Table B.1: Comparison of F-scores, obtained using 5-fold cross validation, for the 7 coarse-grained relation types on the ACE 2004 benchmark corpora for new domain.

Table B.1 shows RE results of our proposed K_{Hybrid} kernel on the ACE 2004 corpus, a benchmark news domain corpus. The ACE 2004 corpus is constructed from newswire and broadcast news text, and contains 7 coarse-grained relation types as listed below with examples[1]

- EMP-ORG (e.g. "US president")

[1]In these examples, Candidate entity mentions are shown in red and blue colors.

- PHYS (e.g. “a military base in Germany”)
- GPE-AFF (e.g. “U.S. businessman”)
- PER-SOC (e.g. “a spokesman for the senator”)
- DISC (e.g. “each of whom”)
- ART (e.g. “US helicopters”)
- OTHER-AFF (e.g. “Cuban-American people”)

Appendix C

List of Criteria for Being Unlikely to be Co-referents in Clinical Text

We created the following list of 13 criteria (as mentioned in Section 3.3.2) for clinical text, and propose that a candidate antecedent (m_x) and the mention to be resolved (m_y) are unlikely to be co-referents if they violate any of them.

1. m_y is a determiner and part of an NP rather than constituting an NP itself:

   - e.g. if the word *"this"* is the mention m_y but it is part of a larger mention (m_x) *"this patient"* then they are not co-referents.

2. m_y is a determiner (or Wh-determiner) and it is not among the first two words of the sentence and m_x does not belong to the same sentence:

   - e.g. consider the following consecutive sentences (extracted from clinical text) and a candidate pair (which are not co-referent) where *"hematemesis"* in the first sentence is m_x and *"which"* in the second sentence is m_y:

     Sen. 1: *Mr. Bruno is a 60 year old gentleman who initially presented with* ***hematemesis****, hemoptysis and on work-up was found to have a left lower lobe mass .*

Sen. 2: *He previously underwent bronchoscopy with washings* ***which*** *showed to be negative for malignant cells and showed atypical bronchial epithelial cells , likely to be reactive .*

3. both m_x and m_y are of type `Problem` but they are semantically attached to different persons:
   - e.g. "*Diabetics of the patient*" and "*Diabetics of the patient's father*" do not refer to the same entity.
4. m_y is non-pronominal but m_x is pronominal
5. the gender types (male/female/neutral) of the mentions are known and they are not same
6. the number types (singular/plural) of the mentions are known and they are not same
7. m_x is of `Person` type or a personal pronoun but m_y is not, and the vice-versa
8. neither m_x nor m_y is pronoun, and their semantic types are not same
9. both m_x and m_y are one of the following clinical person types `Patient, Doctor, Family` and `Other`, and their clinical person types are not same
10. either m_x or m_y is tagged as a preposition (i.e. IN) by the parser
    - e.g. consider the following sentence and a candidate pair (which are not co-referent) where *"Mr. Anders"* (annotated as `Person`) is m_x and *"that"* (annotated as `Pronoun`) is m_y. The Stanford parser assigns the POS tag "IN" to *"that"*.

      Mr. Anders states that 2 weeks prior to his presentation he hurt his left knee .

11. the syntactic head word of either m_x or m_y is tagged as a conjunction (i.e. CC) by the parser

12. either m_x or m_y is a section heading

13. the 1st word of m_y is neither a determiner nor a pronoun but it is for m_x and these two mentions do not have any common words

Appendix D

Results of Our Proposed NER Approach on A Automatically Transcribed Broadcast News Corpus

We participated in the "NER on Transcribed Broadcast News" (closed) task of the EVALITA 2011 evaluation campaign to verify how well our proposed NER approach, primarily developed for bio-entity mention identification, adapts on a completely different genre with few adjustments. We used only the training data distributed by the organizers and no additional resource. Participants were expected to identify four named entity (NE) types: *(i) Person (PER), (ii) Organization (ORG), (iii) Location (LOC),* and *(iv) Geo-Political Entities (GPE).* Our system was the 2nd best system among the participating teams with an F1-score of *57.02* on the automatically transcribed broadcast news. On the manual transcriptions of the

NE type	**Total NEs identified**	**Correct NEs**	**Precision**	**Recall**	**F1-score**
GPE	660	-	75.45	73.78	74.61
LOC	55	-	69.09	40.43	51.01
ORG	371	-	52.02	35.41	42.14
PER	364	-	51.92	41.45	46.10
ALL	1450	918	63.31	51.86	57.02

Table D.1: Official results of our NER approach on the test data of the EVALITA 2011 NER (closed) task.

NE type	Total NEs identified	Correct NEs	Precision	Recall	F1-score
GPE	633	-	75.99	71.68	73.77
LOC	51	-	70.59	36.00	47.68
ORG	241	-	55.60	24.68	34.18
PER	350	-	52.00	40.00	45.22
ALL	1275	833	65.33	47.09	54.73

Table D.2: Results of our NER approach, *excluding global perspective features*, on the EVALITA 2011 NER (closed) task test data.

same test data, where transcription errors are fixed but sentence boundaries and punctuation symbols are still missing, the system achieves an F1-score of *73.54*. Table D.1 shows our official results.

Table D.2 shows the impact on the results when the proposed global perspective features (see Section 2.2.2) are excluded. As we can see, leaving out these features causes a considerable decrease (approximately 2.3%) of overall F_1 score (see results in Table D.1 for comparison). Particularly, there is almost 8% drop off in the identification of *ORG* NEs. More details are available in Chowdhury (2013).

Appendix E

Drug-Drug Interaction Detection and Classification in SemEval-2013

The task #9[2] of SemEval-2013 concerns the recognition of drugs and the extraction and classification of drug-drug interactions from biomedical literature. The dataset of the shared task is composed by texts from the DrugBank database as well as MedLine abstracts in order to deal with different type of texts and language styles. Participants were asked to not only extract DDIs but also classify them into one of four pre-defined classes: advise, effect, mechanism and int. A detailed description of the task settings and data can be found in Segura-Bedmar et al. (2013).

We participated in this shared task (Chowdhury and Lavelli, 2013b) using our RE system described in Section 4. The official results of the task show that our approach yields an F-score of 0.80 for DDI detection and an F-score of 0.65 for DDI detection and classification. Our system obtained significantly higher results than all the other participating teams in this shared task and has been ranked 1st.

For DDI detection, we automatically discarded less informative sentences and instances, and then trained the system (a single model regardless of DDI types) on the remaining training instances to identify possible DDIs from the remaining test instances.

[2]http://www.cs.york.ac.uk/semeval-2013/task9/

The next step is to classify the extracted DDIs into different categories. For this, we train 4 separate models for each of the DDI types (one Vs all) to predict the class label of the extracted DDIs. During this training, all the negative instances from the training data are removed. The filtering techniques described in Section 4 are not used in this stage.

The extracted DDIs are assigned a default DDI class label. Once the above models are trained, they are applied on the extracted DDIs from the test data. The class label of the model which has the highest confidence score for an extracted DDI instance is assigned to such instance.

The DDIExtraction 2013 shared task data include two types of texts: texts taken from the DrugBank database and texts taken from MedLine abstracts. During training we used both types together. The parameters are tuned by doing 5-fold cross validation on the training data.

Experimental Results

Table E.1 shows the results of 5-fold cross validation for DDI detection on the training data. As we can see, the usage of the LIS and LII filtering techniques improves both precision and recall.

We submitted three runs for the DDIExtraction 2013 shared task. The only difference between the three runs concerns the default class label (i.e. the class chosen when none of the separate models assigns a class label to a predicted DDI). Such default class label is "int", "effect" and "mechanism" for run 1, 2 and 3 respectively. According to the official results provided by the task organisers, our best result was obtained by run 2 (shown in Table E.2).

According to the official results, the performance for "advise" is very low (F_1 0.29) in MedLine texts, while the performance for "int" is comparatively much higher (F_1 0.57) with respect to the one of the other DDI types. In comparison, the performance for "int" is much lower (F_1 0.55) in

	P	**R**	F_1
K_{Hybrid}	0.66	0.80	0.72
LIS filtering + K_{Hybrid}	0.67	0.80	0.73
LIS filtering + LII filtering + K_{Hybrid}	**0.68**	**0.82**	**0.74**

Table E.1: Comparison of results for DDI detection on the training data using 5-fold cross validation. Parameter tuning is not done during these experiments.

	P	**R**	F_1
All text			
DDI detection only	0.79	0.81	0.80
Detection and Classification	0.65	0.66	0.65
DrugBank text			
DDI detection only	0.82	0.84	0.83
Detection and Classification	0.67	0.69	0.68
MedLine text			
DDI detection only	0.56	0.51	0.53
Detection and Classification	0.42	0.38	0.40

Table E.2: Official results of the best run (run 2) of our system in the DDIExtraction 2013 shared task.

DrugBank texts with respect to the one of the other DDI types.

In MedLine test data, the number of "effect" (62) and "mechanism" (24) DDIs is much higher than that of "advise" (7) and "int" (2). On the other hand, in DrugBank test data, the different DDIs are more evenly distributed – "effect" (298), "mechanism" (278), "advise" (214) and "int" (94).

Initially, it was not clear to us why our system (as well as other participants) achieves so much higher results on the DrugBank sentences in comparison to MedLine sentences. Statistics of the average number of words show that the length of the two types of training sentences are substantially

similar (DrugBank : 21.2, MedLine : 22.3). It is true that the number of the training sentences for the former is almost 5.3 times higher than the latter. But it could not be the main reason for such high discrepancies.

So, we turned our attention to the presence of the cue words. In the 4,683 sentences of the DrugBank training set (which have at least one drug mention), we found that the words "increase" and "decrease" are present in 721 and 319 sentences respectively. While in the 877 sentences of the MedLine training set (which have at least one drug mention), we found that the same words are present in only 67 and 40 sentences respectively. In other words, the presence of these two important cue words in the DrugBank sentences is twice more likely than that in the MedLine sentences. We assume similar observations might be also possible for other cue words. Hence, this is probably the main reason why the results are so much better on the DrugBank sentences.

Appendix F

Tools Released for The Community

As part of this PhD research, we have developed the following tools which are made publicly available for research purpose.

- **BioEnEx**: Biomedical Entity Extractor
  - Download url: `https://sites.google.com/site/fmchowdhury2/bioenex`
- **CoRefLinker**: Coreference Linker
  - Download url: `https://github.com/fmchowdhury/CoRefLinker`
- **HyREX**: Hybrid Relation Extractor
  - Download url: `https://github.com/fmchowdhury/HyREX`

References

P Agarwal and DB Searls. 2008. Literature mining in support of drug discovery. *Brief Bioinform*, 9(6):479–492.

A Airola, S Pyysalo, J Bj̈orne, T Pahikkala, F Ginter, and T Salakoski. 2008. A graph kernel for protein-protein interaction extraction. In *Proceedings of BioNLP 2008*, pages 1–9, Columbus, USA.

C Aone and SW Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL 1995)*, pages 122–129.

AR Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings AMIA Symposium*, pages 17–21.

M Banko, MJ Cafarella, S Soderland, M Broadhead, and O Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI 2007)*, pages 2670–2676, Hyderabad, India.

J Björne, A Airola, T Pahikkala, and T Salakoski. 2011. Drug-drug interaction extraction with RLS and SVM classiffers. In *Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*, pages 35–42, Huelva, Spain, September.

AL Blum and P Langley. 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, December.

AL Blum and T Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational learning theory (COLT'98)*, pages 92–100.

O Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270.

A Bodnari, P Szolovits, and O Uzuner. 2012. MCORES: a system for noun phrase coreference resolution for clinical records. *Journal of the American Medical Informatics Association*, 19(5):906–912.

K Bollacker, C Evans, P Paritosh, T Sturge, and J Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data (SIGMOD 2008)*, pages 1247–1250, New York, NY, USA. ACM.

Q Bui, S Katrenko, and PMA Sloot. 2011. A hybrid approach to extract protein-protein interactions. *Bioinformatics*, 27(2):259–265.

M Bundschus, M Dejori, M Stetter, V Tresp, and HP Kriegel. 2008. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, 9:207.

R Bunescu and RJ Mooney. 2005a. A shortest path dependency kernel for relation extraction. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 724–731, Morristown, NJ, USA. Association for Computational Linguistics.

R Bunescu and RJ Mooney. 2005b. A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 724–731, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

R Bunescu and RJ Mooney. 2006. Subsequence kernels for relation extraction. In *Proceedings of the 19th Conference on Neural Information Processing Systems*, pages 171–178.

R Bunescu. 2007. *Learning for Information Extraction: From Named Entity Recognition and Disambiguation To Relation Extraction.* Ph.D. thesis, The University of Texas at Austin.

J Castano, J Zhang, and J Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *Proceedings of the International Symposium on Reference Resolution for NLP.*

E Charniak and M Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005).*

NV Chawla, N Japkowicz, and A Kotcz. 2004. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6(1):1–6, June.

MFM Chowdhury and A Lavelli. 2010a. Disease mention recognition with specific features. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 83–90, Uppsala, Sweden, July. Association for Computational Linguistics.

MFM Chowdhury and A Lavelli. 2010b. Disease mention recognition with specific features. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP 2010), 48th Annual Meeting of the Association for Computational Linguistics*, pages 83–90, Uppsala, Sweden, July.

MFM Chowdhury and A Lavelli. 2011a. Assessing the practical usability of an automatically annotated corpus. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 101–109, Portland, Oregon, USA, June. Association for Computational Linguistics.

MFM Chowdhury and A Lavelli. 2011b. Drug-drug interaction extraction using composite kernels. In *Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*, pages 27–33, Huelva, Spain, September.

MFM Chowdhury and A Lavelli. 2012a. An Evaluation of the Effect of Automatic Preprocessing on Syntactic Parsing for Biomedical Relation Extraction. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 544–551, May.

MFM Chowdhury and A Lavelli. 2012b. Combining tree structures, flat features and patterns for biomedical relation extraction. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 420–429, Avignon, France, April. Association for Computational Linguistics.

MFM Chowdhury and A Lavelli. 2012c. Impact of Less Skewed Distributions on Efficiency and Effectiveness of Biomedical Relation Extraction. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India, December.

MFM Chowdhury and A Lavelli. 2013a. Exploiting the Scope of Negations and Heterogeneous Features for Relation Extraction: Case Study Drug-Drug Interaction Extraction. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2013)*, Atlanta, USA, June. Association for Computational Linguistics.

MFM Chowdhury and A Lavelli. 2013b. Fbk-irst : A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013).*

MFM Chowdhury and P Zweigenbaum. 2013. A Controlled Greedy Supervised Approach for Coreference Resolution on Clinical Text. *Journal of biomedical informatics*.

MFM Chowdhury, A Lavelli, and A Moschitti. 2011. A study on dependency tree kernels for automatic extraction of protein-protein interaction. In *Proceedings of BioNLP 2011*, pages 124–133, Portland, Oregon, USA, June.

MFM Chowdhury, A Ben Abacha, A Lavelli, and P Zweigenbaum. 2011c. Two different machine learning techniques for drug-drug interaction extraction. In *Proceedings of DDIExtraction2011: First Challenge Task: Drug-Drug Interaction Extraction*, pages 19–26, Huelva, Spain, September.

MFM Chowdhury. 2013. A Simple Yet Effective Approach for Named Entity Recognition from Transcribed Broadcast News. In B Magnini, F Cutugno, M Falcone, and E Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian, LNCS 7689*, pages 98–106. Springer-Verlag Berlin Heidelberg.

HW Chun, Y Tsuruoka, JD Kim, R Shiba, N Nagata, T Hishiki, and J Tsujii. 2006. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In *Proceedings of the Pacific Symposium on Biocomputing (PSB) 11*, pages 4–15, Maui, Hawaii, USA, January.

S Clark, JR Curran, and M Osborne. 2003. Bootstrapping POS taggers using unlabelled data. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL 2003)*, pages 49–55.

M Collins. 2002. Ranking algorithms for named-entity extraction: boosting and the voted perceptron. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 489–496, Stroudsburg, PA, USA. Association for Computational Linguistics.

A Culotta and J Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423, Morristown, NJ, USA. Association for Computational Linguistics.

HJ Dai, YC Chang, RT Tsai, and WL Hsu. 2010. New challenges for biological text-mining in the next decade. *Journal of Computer Science and Technology*, 25(1):169–179.

HJ Dai, CY Chen, CY Wu, PT Lai, RTH Tsai, and WL Hsu. 2012. Coreference resolution of medical concepts in discharge summaries by exploiting contextual information. *Journal of the American Medical Informatics Association*, 19:888–896.

P Denis and J Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 660–669.

TG Dietterich, RH Lathrop, and T Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, January.

J Ding, D Berleant, D Nettleton, and E Wurtele. 2002. Mining medline: Abstracts, sentences or phrases? *Pacific Symposium in Biocomputing: 2002; Kauau, Hawaii*, pages 326–337.

R Farkas, V Vincze, G Móra, J Csirik, and G Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL 2010)*, pages 1–12, Uppsala, Sweden, July. Association for Computational Linguistics.

C Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

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

C Gasperin and T Briscoe. 2008. Statistical anaphora resolution in biomedical texts. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 257–264, Stroudsburg, PA, USA. Association for Computational Linguistics.

C Giuliano, A Lavelli, and L Romano. 2006a. 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 (EACL 2006)*, pages 401–408.

C Giuliano, A Lavelli, and L Romano. 2006b. Simple Information Extraction (SIE): A portable and effective IE system. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*, pages 9–16.

AM Gliozzo, C Giuliano, and R Rinaldi. 2005. Instance filtering for entity recognition. *SIGKDD Explor. Newsl.*, 7(1):11–18, June.

P Gooch and A Roudsari. 2012. Lexical patterns, features and knowledge resources for coreference resolution in clinical notes. *Journal of Biomedical Informatics*, 45(5):901–912.

R Grishman. 2003. Information extraction. In R Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, pages 545–559. Oxford University Press.

C Grouin, M Dinarelli, S Rosset, G Wisniewski, and P Zweigenbaum. 2011. Coreference resolution in clinical reports. The LIMSI participation in the i2b2/VA 2011 challenge. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data*, Washington D.C., USA.

A Haghighi and D Klein. 2007. Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 848–855, Prague, Czech Republic, June. Association for Computational Linguistics.

U Hahn, K Tomanek, E Beisswanger, and E Faessler. 2010. A proposal for a configurable silver standard. In *Proceedings of the 4th Linguistic Annotation Workshop, 48th Annual Meeting of the Association for Computational Linguistics*, pages 235–242, Uppsala, Sweden, July.

S He and D Gildea. 2006. Self-training and co-training for semantic role labeling: Primary report. Technical report, University of Rochester.

T Y He. 2007. Coreference resolution on entities and events for hospital discharge summaries. Master's thesis, MIT.

D Hristovski, B Peterlin, JA Mitchell, and SM Humphrey. 2003. Improving literature based discovery support by genetic knowledge integration. *Stud. Health Technol. Inform.*, 95:68–73.

A Jimeno, E Jiménez-Ruiz, V Lee, S Gaudan, R Berlanga, and D Rebholz-Schuhmann. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(S-3).

T Joachims. 1999. Making large-scale support vector machine learning practical. In *Advances in kernel methods: support vector learning*, pages 169–184. MIT Press, Cambridge, MA, USA.

SR Jonnalagadda, D Li, S Sohn, ST Wu, K Wagholikar, M Torii, and H Liu. 2012. Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules. *Journal of the American Medical Informatics Association*, 19:867–874.

JD Kim, T Ohta, Y Tateisi, and J Tsujii. 2003. Genia corpus - semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl 1):i180–182.

JD Kim, T Ohta, S Pyysalo, Y Kano, and J Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.

S Kim, J Yoon, J Yang, and S Park. 2010. Walk-weighted subsequence kernels for protein-protein interaction extraction. *BMC Bioinformatics*, 11(1).

JD Kim, Y Wang, T Takagi, and A Yonezawa. 2011. Overview of Genia event task in BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 7–15, Portland, Oregon, USA, June. Association for Computational Linguistics.

D Klein and C Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 2003)*, pages 423–430, Sapporo, Japan.

SB Kotsiantis and PE Pintelas. 2003. Mixture of Expert Agents for Handling Imbalanced Data Sets. *Annals of Mathematics, Computing and Teleinformatics*, 1(1):46–55.

S Kulick, A Bies, M Liberman, M Mandel, R McDonald, M Palmer, A Schein, and L Ungar. 2004. Integrated annotation for biomedical information extraction. In *Proceedings of HLT/NAACL 2004 BioLink Workshop*, pages 61–68.

JD Lafferty, AK McCallum, and FCN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

E Landau. 2009. Jackson's death raises questions about drug interactions [Published in CNN; June 26, 2009]. `http://edition.cnn.com/2009/HEALTH/06/26/jackson.drug.interaction.caution/index.html`.

R Leaman and G Gonzalez. 2008. Banner: An executable survey of advances in biomedical named entity recognition. In *Proceedings of Pacific Symposium on Biocomputing*, volume 13, pages 652–663.

R Leaman, C Miller, and G Gonzalez. 2009. Enabling recognition of diseases in biomedical text with machine learning: Corpus and benchmark. In *Proceedings of the 3rd International Symposium on Languages in Biology and Medicine*, pages 82–89.

T Lee, Z Wang, H Wang, and S Hwang. 2011. Web scale taxonomy cleansing. In *Proceedings of the VLDB Endowment*.

DB Lenat. 1995. Cyc: a large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38, November.

T Liang and YH Lin. 2005. Anaphora resolution for biomedical literature by exploiting multiple resources. In R Dale, KF Wong, J Su, and OY Kwong, editors, *Natural Language Processing – IJCNLP 2005*, volume 3651 of *Lecture Notes in Computer Science*, pages 742–753. Springer Berlin / Heidelberg.

YH Lin and T Liang. 2004. Pronominal and sortal anaphora resolution for biomedical literature. In *Proceedings of the 16th Conference on Computational Linguistics and Speech Processing*.

H Liu, YA Lussier, and C Friedman. 2001. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *Journal of Biomedical Informatics*, pages 249–261.

H Liu, AR Aronson, and C Friedman. 2002. A study of abbreviations in MEDLINE abstracts. In *Proceedings of the AMIA Annual Symposium*, pages 464–468, November.

H Llorens, L Derczynski, R Gaizauskas, and E Saquete. 2012. TIMEN: An Open Temporal Expression Normalisation Resource. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*.

X Luo, A Ittycheriah, H Jing, N Kambhatla, and S Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the Bell tree. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 135–142.

D Marcu. 1997. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph.D. thesis, Department of Computer Science, University of Toronto.

AK McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu,.

J McCarthy and W Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 1995)*, pages 1050–1055.

D McClosky, E Charniak, and M Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics*, pages 337–344, Sydney, Australia.

D McClosky. 2010. *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*. Ph.D. thesis, Department of Computer Science, Brown University.

R McDonald, F Pereira, S Kulick, S Winters, Y Jin, and P White. 2005. Simple algorithms for complex relation extraction with applications to biomedical IE. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 491–498, Ann Arbor, Michigan, June.

SM Meystre, GK Savova, KC Kipper-Schuler, and JF Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *IMIA Yearbook of Medical Informatics*, pages 128–144.

R Mihalcea. 2004. Co-training and self-training for word sense disambiguation. In *Proceedings of the 8th Conference on Computational Natural Language Learning (CoNLL 2004)*, pages 33–40.

S Miller, H Fox, L Ramshaw, and R Weischedel. 2000. A novel use of statistical parsing to extract information from text. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (NAACL 2000)*, pages 226–233, Stroudsburg, PA, USA. Association for Computational Linguistics.

B Min and R Grishman. 2012. Compensating for annotation errors in training a relation extractor. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 194–203.

M Mintz, S Bills, R Snow, and D Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the*

AFNLP (ACL 2009), pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

M Miwa, R Saetre, Y Miyao, T Ohta, and J Tsujii. 2009a. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *International Journal of Medical Informatics*, 78.

M Miwa, R Sætre, Y Miyao, and J Tsujii. 2009b. A rich feature vector for protein-protein interaction extraction from multiple corpora. In *Proceedings of EMNLP 2009*, pages 121–130, Singapore.

R Morante and E Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

A Moschitti. 2004. A study on convolution kernels for shallow semantic parsing. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, ACL '04, Barcelona, Spain.

A Moschitti. 2006. Making tree kernels practical for natural language learning. In *Proceedings of 11th Conference of the European Chapter of the Association for computational Linguistics (EACL 2006)*, pages 113–120, Trento, Italy.

A Névéol, W Kim, WJ Wilbur, and Z Lu. 2009. Exploring two biomedical text genres for disease recognition. In *Proceedings of the BioNLP 2009 Workshop*, pages 144–152, June.

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

C Nédellec. 2005. Learning language in logic - genic interaction extraction challenge. In *Proceedings of the ICML 2005 workshop: Learning Language in Logic (LLL05)*, pages 31–37.

V Ng and C Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 104–111.

V Ng and C Cardie. 2003. Weakly supervised natural language learning without redundant views. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-2003)*, pages 173–180.

V Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1396–1411, Uppsala, Sweden.

TT Nguyen and A Moschitti. 2011a. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 277–282, Stroudsburg, PA, USA. Association for Computational Linguistics.

TT Nguyen and A Moschitti. 2011b. Joint distant and direct supervision for relation extraction. In *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 732–740, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

TT Nguyen, A Moschitti, and G Riccardi. 2009. Convolution kernels on constituent, dependency and sequential structures for relation extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 1378–1387, Singapore, August.

N Nguyen, JD Kim, and J Tsujii. 2011. Overview of bionlp 2011 protein coreference shared task. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 74–82, Portland, Oregon, USA, June. Association for Computational Linguistics.

EW Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses : An Introduction.* Wiley-Interscience, April.

JW Payne. 2007. A Dangerous Mix [Published in The Washington Post; February 27, 2007]. `http://www.washingtonpost.com/wp-dyn/content/article/2007/02/23/AR2007022301780.html`.

D Pierce and C Cardie. 2001. Limitations of co-training for natural language learning from large datasets. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-2001)*, pages 1–9.

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

S Pyysalo, A Airola, J Heimonen, J Björne, F Ginter, and T Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl 3):S6.

K Raghunathan, H Lee, S Rangarajan, N Chambers, M Surdeanu, D Jurafsky, and C Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 492–501.

A Rahman and V Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 968–977.

D Rebholz-Schuhmann and U Hahn. 2010c. Silver standard corpus vs. gold standard corpus. In *Proceedings of the 1st CALBC Workshop*, Cambridge, U.K., June.

D Rebholz-Schuhmann, AJ Jimeno-Yepes, E van Mulligen, N Kang, J Kors, D Milward, P Corbett, E Buyko, E Beisswanger, and U Hahn. 2010a. CALBC silver standard corpus. *Journal of Bioinformatics and Computational Biology*, 8:163–179.

D Rebholz-Schuhmann, AJ Jimeno-Yepes, E van Mulligen, N Kang, J Kors, D Milward, P Corbett, E Buyko, K Tomanek, E Beisswanger, and U Hahn. 2010b. The CALBC silver standard corpus for biomedical named entities – a study in harmonizing the contributions from four independent named entity taggers. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May.

D Rebholz-Schuhmann, A Yepes, C Li, S Kafkas, I Lewin, N Kang, P Corbett, D Milward, E Buyko, E Beisswanger, K Hornbostel, A Kouznetsov, R Witte, J Laurila, C Baker, CJ Kuo, S Clematide, F Rinaldi, R Farkas, G Móra, K Hara, LI Furlong, M Rautschka, M Neves, A Pascual-Montano, Q Wei, N Collier, MFM Chowdhury, A Lavelli, R Berlanga, R Morante, V Van Asch, W Daelemans, J Marina, E van Mulligen, J Kors, and U Hahn. 2011. Assessment of ner solutions against the first and second calbc silver standard corpus. *Journal of Biomedical Semantics*, 2(Suppl 5):S11.

B Rink, K Roberts, and SM Harabagiu. 2012. A supervised framework for resolving coreference in clinical records. *Journal of the American Medical Informatics Association*, 19:875–882.

L Romano, M Kouylekov, I Szpektor, I Dagan, and A Lavelli. 2006. Investigating a generic paraphrase–based approach for relation extraction. In *Proceedings of EACL 2006*, pages 409–416.

B Rosario and M Hearst. 2004. Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, pages 430–437, Barcelona, Spain.

D Roth and W Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the 8th Conference on Natural Language Learning (CoNLL 2004)*.

O Sanchez-Graillet and M Poesio. 2007. Negation of protein–protein interactions: analysis and extraction. *Bioinformatics*, 23(13):i424–i432.

AS Schwartz and MA Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of Pacific Symposium on Biocomputing*, pages 451–62.

I Segura-Bedmar, P Martínez, and Cd Pablo-Sánchez. 2011a. The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. In *Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*, pages 1–9, Huelva, Spain, September.

I Segura-Bedmar, P Martínez, and Cd Pablo-Sánchez. 2011b. Using a shallow linguistic kernel for drug-drug interaction extraction. *Journal of Biomedical Informatics*, 44(5):789–804.

I Segura-Bedmar, P Martínez, and M Herrero-Zazo. 2013. SemEval-2013 task 9: Extraction of drug-drug interactions from biomedical texts. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, USA, June.

A Severyn and A Moschitti. 2010. Fast cutting plane training for structural kernels. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2010)*.

J Shawe-Taylor and N Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA.

L Smith, L Tanabe, R Ando, CJ Kuo, and et al. 2008. Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(Suppl 2).

Y Song, H Wang, Z Wang, H Li, and W Chen. 2011. Short text conceptualization using a probabilistic knowledgebase. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence (IJCAI 2011)*, pages 2330–2336.

WM Soon, HT Ng, and D CY Lim. 1999. Corpus-based learning for noun phrase coreference resolution. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP 1999)*, pages 285–291.

WM Soon, HT Ng, and DCY Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

R Soricut and D Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003)*, pages 149–156, Edmonton, Canada.

M Stevenson. 2006. Fact distribution in Information Extraction. *Language Resources and Evaluation*, 40:183–201.

M Strube, S Rapp, and C Muller. 2002. The influence of minimum edit distance on reference resolution. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 312–319.

FM Suchanek, G Kasneci, and G Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web (WWW 2007)*, pages 697–706.

A Sun, R Grishman, and S Sekine. 2011. Semi-supervised relation extraction with large-scale word clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011)*, pages 521–529, Portland, Oregon, USA, June. Association for Computational Linguistics.

K Swampillai and M Stevenson. 2010. Inter-sentential relations in information extraction corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 19–21, May.

DR Swanson. 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18.

L Tanabe, N Xie, L Thom, W Matten, and WJ Wilbur. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S3.

P Thomas, M Neves, I Solt, D Tikk, and U Leser. 2011a. Relation extraction for drug-drug interactions using ensemble learning. In *Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*, pages 11–18, Huelva, Spain, September.

P Thomas, S Pietschmann, I Solt, D Tikk, and U Leser. 2011b. Not all links are equal: Exploiting dependency types for the extraction of protein-protein interactions from text. In *Proceedings of BioNLP 2011*, pages 1–9, Portland, Oregon, USA, June.

D Tikk, P Thomas, P Palaga, J Hakenberg, and U Leser. 2010. A Comprehensive Benchmark of Kernel Methods to Extract Protein-Protein Interactions from Literature. *PLoS Computational Biology*, 6(7), July.

M Tofiloski, J Brooke, and M Taboada. 2009. A syntactic and lexical-based discourse segmenter. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2009)*, pages 77–80, Suntec, Singapore, August.

M Torii, Z Hu, CH Wu, and H Liu. 2009. Biotagger-GM: a gene/protein name recognition system. *Journal of the American Medical Informatics Association : JAMIA*, 16:247–255.

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

A Vlachos. 2007. Tackling the BioCreative2 gene mention task with conditional random fields and syntactic parsing. In *Proceedings of the 2nd BioCreative Challenge Evaluation Workshop*, pages 85–87.

G Wang, Y Yu, and H Zhu. 2007. Pore: positive-only relation extraction from wikipedia text. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ISWC'07/ASWC'07, pages 580–594, Berlin, Heidelberg. Springer-Verlag.

W Wang, R Besançon, O Ferret, and B Grau. 2011. Filtering and clustering relations for unsupervised information extraction in open domain. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM 2011)*, pages 1405–1414, New York, NY, USA. ACM.

C Wang, A Kalyanpur, J Fan, BK Boguraev, and DC Gondek. 2012. Relation extraction and scoring in deepqa. *IBM Journal of Research and Development*, 56(3.4):9:1 –9:12, may-june.

H Ware, CJ Mullett, V Jagannathan, and O El-Rawas. 2012. Machine learning-based coreference resolution of concepts in clinical documents. *Journal of the American Medical Informatics Association*, 19:883–887.

G Weiss and F Provost. 2001. The effect of class distribution on classifier learning: An empirical study. Technical report, Rutgers University.

JO Wrenn, DM Stein, S Bakken, and PD Stetson. 2010. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association*, 17(1):49–53.

F Wu and DS Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 118–127, Uppsala, Sweden. Association for Computational Linguistics.

H Xu, S AbdelRahman, M Jiang, JW Fan, and Y Huang. 2011. An initial study of full parsing of clinical text using the Stanford Parser. In *IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW 2011)*, pages 607 –614.

Y Xu, J Liu, J Wu, Y Wang, Z Tu, JT Sun, J Tsujii, and EIC Chang. 2012. A classification approach to coreference in discharge summaries: 2011 i2b2 challenge. *Journal of the American Medical Informatics Association*.

X Yang, G Zhou, J Su, and CL Tan. 2003. Coreference resolution using competitive learning approach. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 176–183.

X Yang, J Su, G Zhou, and C L Tan. 2004. An NP-cluster based approach to coreference resolution. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 226–232.

M Zhang, J Su, D Wang, G Zhou, and CL Tan. 2005. Discovering relations between named entities from a large raw corpus using tree similarity-based clustering. In Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong, editors, *Natural Language Processing – IJCNLP 2005*, volume 3651 of *Lecture Notes in Computer Science*, pages 378–389. Springer Berlin / Heidelberg.

F Zhang, S Shi, J Liu, S Sun, and C Y Lin. 2011. Nonlinear evidence fusion and propagation for hyponymy relation mining. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 1159–1168.

J Zheng, WW Chapman, RS Crowley, and GK Savova. 2011. Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of Biomedical Informatics*, 44(6):1113–1122.

J Zheng, WW Chapman, TA Miller, C Lin, RS Crowley, and GK Savova. 2012. A system for coreference resolution for the clinical narrative. *Journal of the American Medical Informatics Association*, 19(4):660–667.

G Zhou, J Su, J Zhang, and M Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 2005)*, pages 427–434, Ann Arbor, Michigan, USA.

GD Zhou, M Zhang, DH Ji, and QM Zhu. 2007. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 728–736, Prague, Czech Republic, June. Association for Computational Linguistics.

P Zweigenbaum, D Demner-Fushman, H Yu, and KB Cohen. 2007. Frontiers of biomedical text mining: current progress. *Brief Bioinform*, 8(5):358–375.

P Zweigenbaum, G Wisniewski, MFM Chowdhury, M Dinarelli, S Rosset, and C Grouin. 2013. Résolution des coréférences dans des comptes rendus cliniques. *Revue d'Intelligence Artificielle (submitted)*.