

# Detecting Biomedical Relations using Distant Supervision



Roland Roller

Faculty of Engineering

University of Sheffield

A thesis submitted for the degree of

*Doctor of Philosophy*

December 2015

---

In Erinnerung an meinen Vater Franz Roller.

## **Acknowledgements**

First of all, I would like to thank Dr. Mark Stevenson for offering me a scholarship and for supervising my PhD. I am very happy that he introduced me to biomedical information extraction. I am grateful for all his patience, his support and his valuable advice to guide me through the PhD. I learnt so much from him. Furthermore, I highly appreciate all the time he took for me, such as our weekly meetings and the many times I just came around and knocked at his door.

I am very grateful to my PhD panel committee members Prof. Robert Gaizauskas and Dr. Dirk Sudholt for their constructive comments and their help to improve my thesis.

Next, I would like to thank Dr. Ahmet Aker and Dr. Frederic Blain for proofreading parts of my thesis. This was very helpful. I highly appreciate the support of Frederic and other colleagues in the final stage of my PhD ('You are nearly there, Roland!'). Special thanks to all colleagues for the highly valuable feedback during (and outside) the NLP group meetings. I would like to thank all current and former colleagues at the University of Sheffield for the amazing time. I very much enjoyed the common lunches, coffee breaks in the kitchen and after-work drinks. In particular, I would like to thank everybody in room G35. I also would like to thank everybody of the IXA research group in the Basque Country for the fantastic time in Donostia and for 'Pintxo Pote'. In particular, I would like to thank Dr. Eneko Agirre and Dr. Aitor Soroa. I learnt much from our collaboration.

Many thanks to my partner Eletta Giusto for supporting me in every possible way and making my life much happier. I also would like to thank my family for their unreserved support and love.

## **Abstract**

This work concerns the detection of relationships between key information in biomedical publications, such as treatments for diseases or side-effects of drugs. Given a sentence containing some medical concepts the goal is to determine their relationship to each other.

Supervised machine learning methods are a very popular way to address this problem and often provide reliable results. Those methods require manually labelled examples to extract characteristics of particular relationships in order to detect similar information in unlabelled data. However, manually labelled data is not always available and its generation is time consuming and expensive.

The main objective of this thesis is the exploration of distant supervision, a method which generates those labelled examples automatically using prior knowledge to detect relationships between key facts.

First, relation extraction using a limited amount of training data is explored to detect adverse-drug effects in natural language. Then, work focuses on automatically labelling data using a large biomedical knowledge base, the Unified Medical Language System (UMLS). The effectiveness of a popular evaluation method that does not require manually labelled data is examined in more detail. The main goal is the investigation of whether UMLS is suitable to be used to label data automatically so as to detect similar information in natural language. Finally, a method to reduce falsely labelled instances in the automatically generated data is presented and found to improve the detection of relationships.

# Contents

<b>Contents</b>	<b>4</b>
<b>List of Figures</b>	<b>7</b>
<b>List of Tables</b>	<b>9</b>
<b>Nomenclature</b>	<b>12</b>
<b>1 Introduction</b>	<b>13</b>
1.1 Contribution . . . . .	16
1.2 Published Material . . . . .	16
1.3 Thesis Overview . . . . .	17
<b>2 Related Work and Background</b>	<b>20</b>
2.1 Relation Extraction . . . . .	20
2.1.1 Biomedical Supervised Relation Extraction . . . . .	22
2.1.2 Distant Supervision . . . . .	23
2.1.3 Mixed Classification Models . . . . .	30
2.2 Evaluating Relation Extraction Systems . . . . .	31
2.2.1 Evaluation Levels . . . . .	32
2.2.2 Evaluation Metrics . . . . .	34
2.3 Resources . . . . .	35
2.3.1 Medline repository . . . . .	36
2.3.2 UMLS . . . . .	36
2.3.3 MetaMap . . . . .	40
2.4 Summary . . . . .	42

<b>3</b>	<b>Bootstrapping Limited Training Data</b>	<b>44</b>
3.1	ADE Data . . . . .	45
3.1.1	Corpus overview . . . . .	45
3.2	Automatic Generation of Additional Training Data . . . . .	47
3.3	Relation Extraction System . . . . .	51
3.4	Experiment . . . . .	52
3.4.1	Analysis of Generated Data . . . . .	52
3.4.2	Results . . . . .	54
3.5	Conclusion . . . . .	55
 <b>4</b>	 <b>Detecting relations from the UMLS Metathesaurus in Medline abstracts</b>	 <b>57</b>
4.1	Selection of UMLS Metathesaurus relations . . . . .	58
4.2	Data Generation . . . . .	62
4.3	Corpus Statistics . . . . .	64
4.4	Further filtering steps . . . . .	66
4.5	MultiR - Relation Extraction . . . . .	68
4.6	Experiment . . . . .	70
4.6.1	Setup . . . . .	71
4.6.2	Held-Out Evaluation . . . . .	73
4.6.3	Manual Evaluation . . . . .	75
4.7	Summary . . . . .	76
 <b>5</b>	 <b>Comparison of Evaluation Strategies for Distantly Supervised Relation Extraction</b>	 <b>77</b>
5.1	Data Generation . . . . .	78
5.1.1	Distant labelling . . . . .	78
5.1.2	Manually labelled Test Data . . . . .	79
5.2	Label Comparison . . . . .	81
5.2.1	Sentence Level Labels . . . . .	81
5.2.2	Entity Level Labels . . . . .	82
5.3	Experiment . . . . .	84
5.3.1	Experimental Setup and Overview . . . . .	84
5.3.2	Entity level evaluation . . . . .	85

5.3.3	Sentence level evaluation . . . . .	87
5.4	Conclusion . . . . .	89
<b>6</b>	<b>Reduction of Falsely Labelled Data</b>	<b>91</b>
6.1	Motivation . . . . .	91
6.2	PRA-Reduction . . . . .	94
6.2.1	Removing False Negative Instances . . . . .	95
6.3	Experimental Setup . . . . .	96
6.3.1	Training Data Sets . . . . .	96
6.3.2	Evaluation . . . . .	97
6.4	Results . . . . .	98
6.4.1	Held-out data . . . . .	98
6.4.2	Manually labelled . . . . .	100
6.5	Data Analysis . . . . .	101
6.5.1	Examination of PRA-reduced data . . . . .	101
6.5.2	Pattern Analysis . . . . .	102
6.6	Conclusions . . . . .	103
<b>7</b>	<b>Conclusions</b>	<b>105</b>
7.1	Summary of thesis . . . . .	105
7.2	Future directions . . . . .	106
	<b>References</b>	<b>110</b>
	<b>Appendix 1</b>	<b>129</b>
	<b>Appendix 2</b>	<b>131</b>



# List of Figures

1.1	Sentence mentions a possible treatment for depression. . . . .	14
2.1	Example of the relation <i>PREVENT</i> . . . . .	21
2.2	Correctly labelled sentence using distant supervision. . . . .	24
2.3	Falsely labelled sentence using distant supervision. . . . .	24
2.4	Held-Out Evaluation . . . . .	31
2.5	Correctly predicted adverse drug effect between target entities. . . . .	33
2.6	Incorrectly predicted adverse drug effect between target entities. . . . .	33
2.7	MetaMap-segmented sentence . . . . .	41
2.8	MetaMap annotation of UMLS-CUIs . . . . .	42
3.1	Example of a drug-related adverse effect . . . . .	46
3.2	Mapping UMLS annotations to a sentence using MetaMap . . . . .	47
3.3	Replacing UMLS-CUIs with semantic types . . . . .	47
3.4	Keep entities that match to semantic groups . . . . .	48
3.5	Automatic generation of training data for ADE relations . . . . .	48
3.6	Positive ADE sentence (top) and its MetaMap annotation (below) . . . . .	49
3.7	Example of an insufficient entity normalisation . . . . .	50
3.8	Automatically labelled positive example . . . . .	50
4.1	Group of CUIs might cause false negatives . . . . .	67
4.2	Using non-nouns for distant supervision might cause errors . . . . .	67
4.3	Adjusting bias of training data (Held-Out Setup) . . . . .	73
4.4	Precision/Recall Curve for Held-out data . . . . .	74
6.1	Using isa-relations to reduce false negatives . . . . .	92

## LIST OF FIGURES

---

6.2	Using more relations to reduce false negatives (example 2)	94
6.3	Bias adjustment	97
6.4	Precision/Recall Curve for Held-out data using PRA	100
6.5	PRA-reduced example 1	102
6.6	PRA-reduced example 2	102
7.1	Example sentence of a positively predicted, but unknown entity pair	107
7.2	False named entity recognition	108

# List of Tables

2.1	Confusion Matrix: Predicting instances . . . . .	34
2.2	Excerpt of MRCONSO . . . . .	38
2.3	Excerpt of MRREL . . . . .	38
2.4	CUI candidate list (including semantic types) provided by MetaMap .	41
3.1	Data input for classifier of a positive sentence from publication PMID <i>10048291</i> , taken from ADE data set. . . . .	51
3.2	ADE training data size (mean across five runs) . . . . .	53
3.3	Effect of varying size of training data set . . . . .	54
4.1	Relation frequencies in the UMLS Metathesaurus . . . . .	59
4.2	UMLS Metathesaurus relation examples . . . . .	60
4.3	Top-30 vocabularies of the UMLS Metathesaurus based on relation instances . . . . .	61
4.4	NDF-RT relations . . . . .	63
4.5	Amount of CUI pairs for the generation of the baseline data (all); #given: amount of CUI pairs in MRREL, #pos: amount of positive CUI pairs, #neg (g): amount of negative CUI pairs generated from the positive pairs, #u-pos: unique positive pairs, #u-neg: unique negative pairs . . . . .	65
4.6	Selection of features used for distant supervision . . . . .	70
4.7	Best results using held-out . . . . .	74
4.8	Evaluation of distantly labelled classifier using manually labelled data	76
5.1	Comparison of manual and distantly labelled annotations . . . . .	81

## LIST OF TABLES

---

5.2	Comparison of manual and distantly labelled annotations at entity level	83
5.3	Results for relation extraction system evaluated against DL and ML data sets (may-prevent) . . . . .	86
5.4	Results for relation extraction system evaluated against DL and ML data sets (may-treat) . . . . .	86
5.5	Results for relation extraction system evaluated against DL and ML data sets (Overall) . . . . .	87
5.6	Sentence-level results for relation extraction system evaluated against DL and ML data sets (may-prevent) . . . . .	88
5.7	Sentence-level results for relation extraction system evaluated against DL and ML data sets (may-treat) . . . . .	88
5.8	Sentence-level results for relation extraction system evaluated against DL and ML data sets (Overall) . . . . .	89
6.1	Example PRA-induced paths and weights for the NCI relation <i>biological-process-involves-gene-product</i> . . . . .	95
6.2	Example PRA-induced paths and weights for the NDF-RT relation <i>contraindicating-class-of</i> . . . . .	95
6.3	Evaluation using held-out data . . . . .	99
6.4	Evaluation using manually labelled data . . . . .	101
6.5	Example PRA-induced paths and weights for the NCI relation <i>biological-process-involves-gene-product</i> . . . . .	103
6.6	Example PRA-induced paths and weights for the NCI relation <i>may-prevent</i> . . . . .	104
1	Relationship definitions of MRREL . . . . .	129
2	Semantic groups according to Bodenreider and McCray [2003]. . . . .	130

# Nomenclature

ADE	Adverse Drug Effect
BB	Bacteria Biotopes
CDR	Chemical Disease Relation
CG	Cancer Genetics
CRF	Conditional Random Field
CTD	Comparative Toxicogenomics Database
CUI	Concept Unique Identifier
DDI	Drug Drug Interaction
DL	Distantly Labelled
DS	Distant Supervision
GE	Genia Event
GO	Gene Ontology
JSRE	Java Simple Relation Extraction
KB	Knowledge Base
MeSH	Medical Subject Headings

## **LIST OF TABLES**

---

MIL	Multi-Instance Learning
ML	Manually Labelled
NCI	National Cancer Institute Thesaurus
NDF-RT	National Drug File Repository Thesaurus
NE	Named Entity
NER	Named Entity Recognition
NLM	National Library of Medicine
NLP	Natural Language Processing
PDB	Protein Data Bank
PET	Path-Enclosed Tree
PID	Pathway Interaction Database
PMID	PubMed-Id
POS	Part of Speech
PPI	Protein-Protein Interaction
PRA	Path Ranking Algorithm
REL	Relation Label of MRREL
RELA	Relation Attribute Label of MRREL
SA	seed abstracts
SAB	Source Abbreviation
SL	Shallow Linguistic
STY	Semantic Type
SVM	Support Vector Machine
YPD	Yeast Protein Database

# Chapter 1

## Introduction

Every day people produce textual data containing information related to the life sciences. Data is generated for various purposes and published using different channels such as news-pages, scientific papers, clinical records or forums. PubMed<sup>1</sup> for instance, is a large repository for scientific papers from the biomedical domain. Each year hundreds of thousands of new publications are added to the repository<sup>2</sup> [Zheng and Blake, 2015]. Valuable information, experimental outcomes and new discoveries, which are relevant for pharmacological laboratories or doctors might be published in papers but disappear in the large amount of data. This information is described in natural language, and therefore, difficult to access by computer programs. It is not possible for humans to read all documents available to find informations of interest.

For instance, a search request on PubMed to find new or alternative treatments for depression might provide a large number of documents. Searching for a term such as “depression”, PubMed returns 333,955 documents<sup>3</sup>. A more fine-grained search request such as “pharmacological treatment of depression” decreases the list down to 71,900. PubMed offers numerous search filters to narrow down the list of relevant information [Ebbert et al., 2003; Lindsey and Olin, 2013], such as a restriction per year, per journal or by Medical Subject Headings (MeSH<sup>4</sup>). In PubMed MeSH terms are used by medical experts to index publications which provide a categorisation. Addi-

---

<sup>1</sup><https://www.ncbi.nlm.nih.gov/pubmed/>

<sup>2</sup>[https://www.nlm.nih.gov/bsd/medline\\_cit\\_counts\\_yr\\_pub.html](https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html), 19th of September 2015

<sup>3</sup>according to a search request on 22th of September 2015

<sup>4</sup><https://www.nlm.nih.gov/mesh/>

---

tionally, PubMed offers document ranking by year or relevance. However, it might be helpful using a program that is able to detect all different mentions of pharmacological treatments for depression. Firstly, this could improve the search by selecting only those publications that are really relevant for the treatment. Secondly, pharmacological treatments could be extracted in order to select only publications with a treatment of interest. Moreover, such a program could be used to detect new treatments that are not known by any database. The example in Figure 1.1 shows a sentence expressing a treatment for depression. The sentence is taken from a publication with the PubMed-Id=11092117 (PMID) and mentions the usage of the drug *amitriptyline* as treatment. How can those connections between medical concepts (entities) successfully be detected to improve information access?

```
She had been taking [DRUG:amitriptyline] 75 mg at night for  
[DISEASE:depression] for four months before her admission.  
(PMID=11092117)
```

Figure 1.1: Sentence mentions a possible treatment for depression.

Relation extraction addresses the task of detecting relationships between entities in natural language, such as in the previous example. In this context a relation in a sentence can express, for instance, that a drug has a particular side effect or that a drug can be used to treat a disease<sup>5</sup>.

Relations expressed in natural language can be detected in different ways (see Chapter 2). Currently, machine learning techniques such as supervised learning have been proven the most effective for detecting relations in natural language, as shown in different shared tasks and challenges (see e.g. [Kim et al., 2011; Nédellec et al., 2013; Segura-Bedmar et al., 2013]).

Classifiers often provide better results when more training data is available<sup>6</sup> (see e.g. in [Banko and Brill, 2001; Brants et al., 2007]). Nevertheless, training data is not available in all cases. The generation of manually labelled training data is, unfortu-

---

<sup>5</sup>Relations are described as binary relations between two entities in this work.

<sup>6</sup>This statement can be very controversial, because more data does not necessarily lead always to further improvements, without paying attention to the method [Curran and Osborne, 2002]. However, using a training data set with a few thousand instances might lead to better results than using only a hundred training instances, as seen in Thomas et al. [2011] for instance.



---

nately, a time consuming and expensive process. Time is usually precious for people, such as doctors, who might be suitable to carry out a biomedical annotation. This makes supervised learning methods not ideal for all situations even though the results are very efficient.

An alternative approach to address relation extraction is distant supervision (DS; also known as self-supervision or weak supervision). The technique generates training data automatically using a set of known facts for a relation (e.g. *may-treat*(‘*aspirin*’, ‘*pain*’)). A sentence is labelled as positive if it contains a positively related fact and as negative if it contains a negative fact. Negative facts are often generated following the closed-world assumption: If an information is not known to be positive it is assumed to be negative. Obviously, the assumption of distant supervision is not always correct. The words *aspirin* and *pain* can occur together in a sentence without expressing the *may-treat* relation. Contrariwise, a sentence might be labelled as negative even though the sentences expresses the relation of interest. This situation might occur if the set of known facts is not complete. Thus, training data generated by distant supervision may contain noisy (e.g. falsely labelled) examples but, despite this, classification results are often reasonable (although not as good as using manually labelled data).

This thesis focuses on distantly supervised relation extraction from biomedical abstracts. It aims to address the following research problems:

- The Unified Medical Language System (UMLS) (see Chapter 2.3.2) is a large medical knowledge base (KB) which contains millions of medical concepts and relations between them. This thesis examines whether UMLS is a useful knowledge source for distant supervision.
- Distant supervision is typically used if no manually labelled training data exists. This means that often also no evaluation data is available. Consequently, this work analyses the efficiency of existing evaluation techniques related to distant supervision. In particular, a commonly used evaluation technique *held-out* is examined to determine whether it is useful to estimate the quality of a classifier.
- Distant supervision labels training data automatically using given facts. However, distantly labelled data may contain false annotations. This work examines a method to reduce the amount of false labels.

- 
- In general supervised learning methods using manually labelled training data provide more reliable results than using much larger amounts of (noisy) distantly labelled training data. This work investigates under which circumstances distantly labelled data can support supervised learning methods to reduce the amount of required manually labelled data.

## 1.1 Contribution

This thesis makes the following research contributions:

- The thesis explores the utility of the UMLS Metathesaurus for relation extraction and shows that UMLS is a useful knowledge source for distant supervision.
- Various evaluation methods related to distant supervision are explored. In this context the following contributions and findings are made:
  - held-out evaluation provides reasonable results, in particular for entity-level evaluation
  - system optimisation against held-out data does not necessarily improve system performance against a gold standard data
- This work introduces a novel method to reduce false negatives in distantly labelled data using inference learning.
- The thesis explores the advantage of using distantly labelled data when only a limited amount of training data is available.

## 1.2 Published Material

Parts of in this thesis have been published in a range of peer reviewed conferences and workshops:

- Roller and Stevenson [2015a] explore methods to support supervised learning with a limited amount of training data using distantly labelled data. The content of the paper is included in Chapter 3.

- 
- In Roller and Stevenson [2014a], held-out evaluation results are presented for a range of UMLS relations. The work described in Chapter 4 is partially based on this paper.
  - Chapter 5 compares evaluation strategies for distantly supervised relation extraction against distantly and manually labelled data. Parts of the results of this chapter are published in Roller and Stevenson [2015b].
  - Roller et al. [2015] explore the usage of inference learning to detect potentially false negatives in distantly labelled data. The work is presented in Chapter 6.
  - Roller and Stevenson [2014b] describe the first attempt to use UMLS for distant supervision. The publication unveiled the difficulties of evaluating a distantly supervised classifier without gold standard data. The system serves as baseline for Chapter 4. Furthermore, it triggered the experiment described in Chapter 5.
  - Parts of the system architecture, including the processing of natural language (e.g. stemming, dependency parsing) have been developed in Roller and Stevenson [2013] to detecting gene events and are used, with slight modifications, throughout the thesis (in particular Chapter 3 and Chapter 4).

## 1.3 Thesis Overview

The remainder of this thesis is structured as follows:

Chapter 2 (Related Work and Background) describes related work on relation extraction. The chapter starts with a brief overview of the history of relation extraction and reports different supervised methods in a biomedical context. Then the chapter presents related work in the context of distantly supervised relation extraction from in and outside the biomedical domain, followed by an overview of evaluation methods for distant supervision. Finally, a range of relevant resources are presented, including UMLS and MetaMap.

Chapter 3 (Bootstrapping Limited Training Data) explores relation extraction using a limited amount of manually labelled training data. In the context of adverse-drug

---

effects a small set of instances is used to train a supervised classifier. The chapter investigates whether it is possible to improve classification results by using a large set of automatically labelled training instances. The new training data is generated from facts provided by the small set of manually labelled training instances.

Chapter 4 (Detecting relations from the UMLS Metathesaurus in Medline abstracts) introduces the relations from the UMLS Metathesaurus used in this thesis. It describes how distantly labelled data from sentences of PubMed abstracts is generated. It also describes a range of different filtering techniques that are applied in order to provide a good quality of distantly labelled data. Furthermore, the chapter presents distantly supervised classification results for a range of different UMLS Metathesaurus relations from two different vocabularies (NDF-RT and NCI). The results are evaluated using the held-out approach and a small gold standard of two different UMLS relations. The chapter supports the assumption that UMLS is a useful knowledge source for distantly supervised relation extraction.

Chapter 5 (Comparison of Evaluation Strategies for Distantly Supervised Relation Extraction) examines the efficiency using held-out evaluation for distantly supervised relation extraction in more detail. The question that is explored in this chapter is whether held-out is a useful method to evaluate distant supervision and what the results mean. To ensure reliable results distantly labelled data is manually re-annotated to generate a new gold standard. Then a classifier is trained and evaluated on the evaluation subset; once using the distant labels and once using the manual labels.

Chapter 6 (Reduction of Falsely Labelled Data) explores a technique to remove potential false negatives from the distantly labelled data to improve classification results. The proposed approach uses an inference learning method implemented in the path ranking algorithm (PRA) [Lao and Cohen, 2010; Lao et al., 2011]. Using the same set of relations from Chapter 4, this chapter shows that the proposed method significantly improves the classification results.

Chapter 7 (Conclusions) finishes this thesis with a summarisation and a conclusion. Moreover, the final chapter provides an overview of possible future directions for the

---

research described in this thesis.

## Chapter 2

# Related Work and Background

This chapter presents literature and resources related to this thesis. First, an overview of different relation extraction approaches is provided, in particular supervised and distantly supervised relation extraction. Then, different techniques are presented to evaluate distant supervision. Finally, an overview about some relevant resources is provided.

### 2.1 Relation Extraction

Relation extraction is an important topic in natural language processing (NLP) [Zeng et al., 2015] and focuses on the detection of relationships between entities (concepts, things) from natural language. Given a sentence with some key entities, the task is to determine their relationship to each other. In the biomedical domain a relationship can be described, for instance, between a drug and its contraindicating effect, proteins which interact with each other or a disease and a drug which prevents it. Figure 2.1 shows an example of the relation *PREVENT* (PREV) taken from the Rosario & Hearst [Rosario and Hearst, 2004] data set. The sentence describes a prevention of *swine enzootic pneumonia* by using a *vaccination*.

Relation extraction can be defined in various ways and be addressed using different methods. Relations can be extracted with information across documents [Yao et al., 2010], across sentence boarders [Roberts et al., 2008; Swampillai and Stevenson, 2011] or combined using co-reference resolution [Chowdhury and Zweigenbaum,

---

A field study was carried out to evaluate the effect of [vaccination] against [swine enzootic pneumonia] in different production systems. (PMID=11129801)

Figure 2.1: Example of the relation *PREVENT*

2013; Xu et al., 2008], a method that is able to detect pronouns of target entities. In some cases relation extraction includes the detection of entities beforehand (named entity recognition; also called NER) [Björne and Salakoski, 2013; Kang et al., 2014] and in some other cases those entities are already provided [Chowdhury and Lavelli, 2012; Liu et al., 2013; Rosario and Hearst, 2004]. This thesis focuses on the task of detecting relations between given entities in single sentences. NER can be a challenging task, in particular in the biomedical domain due to nomenclature, abbreviations and ambiguity [Kim et al., 2004; Zhu and Shen, 2012]. For this reason relation extraction is carried out by MetaMap (see Section 2.3.3), a widely used tool to annotate UMLS concepts to natural language [Simpson and Demner-Fushman, 2012].

According to Angeli et al. [2014] relation extraction can be divided into one of four approaches: supervised relation extraction [Björne and Salakoski, 2013; Gurulingappa et al., 2012a], distant supervision [Abad and Moschitti, 2014; Nunes and Schwabe, 2014; Yao et al., 2010], rule-based systems [Riloff, 1993; Soderland et al., 1995] and open information extraction [Banko et al., 2007; Wu and Weld, 2010], which acquires related information without prior knowledge and without training data. Another method that could be included in this list is minimal supervised learning<sup>1</sup> or bootstrapping (such as [Agichtein and Gravano, 2000; Batista et al., 2015; Brin, 1999; Riloff and Jones, 1999; Yangarber et al., 2000]) which uses a small set of given instances (examples of related information, called seeds), or patterns to acquire further examples from a large corpus by applying an iterative process. While this approach does not require labelled training data it often suffers from low precision or semantic drift [Mintz et al., 2009]. However, grouping systems into one to those categories can be sometimes difficult; approaches might combine techniques across different categories. For this work supervised relation extraction and, in particular, distant supervision are

---

<sup>1</sup>also sometimes referred as *semi-supervised*

---

of interest with a focus on the biomedical domain.

### 2.1.1 Biomedical Supervised Relation Extraction

In the last decade machine learning techniques became very popular for relation extraction (such as [Culotta, 2004; Giuliano et al., 2006; Zelenko et al., 2003; Zhou et al., 2005]), in particular supervised learning. Supervised relation extraction is a method that requires training data, usually consisting of positive and negative training examples to *train* a classifier. There are different methods and techniques to address supervised relation extraction. In the biomedical domain various shared tasks have been developed for different relations and use cases. These challenges often provide a good overview about efficient approaches to solve different problems such as relation extraction. Some relevant competitions include the following:

- BioNLP Shared Task [Kim et al., 2011; Nédellec et al., 2013], including many sub-challenges such as Genia Event (GE) [Kim et al., 2013], Cancer genetics (CG) [Pyysalo et al., 2013] or Bacteria Biotopes (BB) [Bossy et al., 2013]
- Drug-Drug Interaction (DDI) challenge [Segura-Bedmar et al., 2011b, 2013]
- Chemical Disease Relation (CDR) task [Wei et al., 2015]

Some other challenges address a range of further problems related to clinical data, such as:

- i2b2 challenge [Uzuner et al., 2011]
- CLEF eHealth Evaluation Lab [Goeriot et al., 2015]

Competitions such as BioNLP or DDI are more relevant for this work, since both aim at relation extraction from biomedical publications as in this work. A short overview about successful methods is now provided.

An analysis of the BioNLP GE task (the task with the most participants) shows, that the majority of the participating systems rely on machine learning methods to detect events and relations [Björne and Salakoski, 2013; Hakala et al., 2013; Li et al., 2013]. The two most successful systems EVEX [Hakala et al., 2013] and TEES [Björne and



---

Salakoski, 2013] rely on a support vector machine (SVM) [Joachims, 1999] with a large set of different features (e.g. token features, sentence features, dependency chains) [Bjorne et al., 2012]. Some other approaches in the challenge use rule-based methods [Bui et al., 2013; Tran et al., 2013]. The BioSEM system of Bui et al. [2013] for instance is ranked third and uses a rule-based approach by generating patterns for each event (e.g. binding, regulation).

The situation within the DDI 2013 challenge is similar. The best systems here (FBK-irst [Chowdhury and Lavelli, 2013a], WBI [Thomas et al., 2013] and UTurku [Björne et al., 2013]) also rely on machine learning methods using a SVM. The winning system, FBK-irst, combines a range of different kernel methods such as a feature based kernel [Chowdhury and Lavelli, 2013b], shallow linguistic (SL) kernel [Giuliano et al., 2006] and path-enclosed tree (PET) kernel [Moschitti, 2004]. Moreover, the authors introduced a technique to reduce less informative candidate sentences in the training data.

Supervised machine learning techniques are very efficient to detect relations in natural language, but require annotated data (positively and negatively labelled sentences) to train the classifier. Machine learning methods tend to provide better classification results the more training data is available (see e.g. [Banko and Brill, 2001; Brants et al., 2007]). Unfortunately data is not always available for all different tasks. Moreover, the generation of an annotated data set for training can be time consuming [Kim et al., 2008] and expensive [Angeli et al., 2014]. In particular in the biomedical domain often expert knowledge is required to annotate medical data (such as in [Kim et al., 2008; van Mulligen et al., 2012]). For this reason, distantly supervised relation extraction has become a popular alternative to supervised learning and is introduced in the following subsection.

## 2.1.2 Distant Supervision

Distant supervision (also known as self-supervision or weak supervision)<sup>2</sup> is a widely applied technique for training relation extraction systems [Krause et al., 2012; Nunes

---

<sup>2</sup>Depending on the literature there might be different names for the same method, but also various interpretations of the different terms. However, in this work the three terms (*distant supervision*, *self-supervised learning* and *weak supervision*) are considered to be equivalent according to Riedel et al. [2010].

---

and Schwabe, 2014; Ritter et al., 2013; Roth and Klakow, 2013; Vlachos and Clark, 2014] that avoids the need for annotated training data. In particular in the last 5-7 year the technique gained popularity. According to Mintz et al. [2009] distant supervision is defined as follows:

“The distant supervision assumption is that if two entities participate in a relation, any sentence that contain those two entities might express that relation.”.

Hence, training examples are annotated automatically using known facts (usually from a knowledge base, for example [Ellendorff et al., 2014; Pershina et al., 2014; Poon et al., 2015]). These facts are matched against text and used as training examples. For example, a knowledge base may assert that the entity pair (“*hair loss*”, “*paroxetine*”) is an instance of the relationship *adverse-drug effect*. Distant supervision approaches normally assume that sentences containing both entities assert the relation between them and, consequently, the sentence in Figure 2.2 would be used as a positive example of the *adverse-drug effect* relation.

*Findings on discontinuation and rechallenge supported the assumption that the **hair loss** was a side effect of the **paroxetine**. (PMID=10442258)*

Figure 2.2: Correctly labelled sentence using distant supervision.

However, this assumption does not always hold which can lead to sentences containing entity pairs being mistakenly identified as asserting a particular relation between them. For example, the sentence in Figure 2.3 contains the same entity pair but does not assert the *adverse-drug effect* relation.

*There are a few case reports on **hair loss** associated with tricyclic antidepressants and serotonin selective reuptake inhibitors (SSRIs), but none deal specifically with **paroxetine**. (PMID=10442258)*

Figure 2.3: Falsely labelled sentence using distant supervision.

---

Distantly supervised relation extraction normally requires positive and negative training examples. Negative data can be generated in different ways. Ling et al. [2013] for instance produce negative data for the relation *has-part* in a controlled way. Examples are generated by reverse pairs (if  $e_1$  has-part  $e_2$ , then  $e_2$  NOT-has-part  $e_1$ ) and transitive characteristics of existing negative instances. Nonetheless, this technique cannot be applied for all target relations. Thomas et al. [2011] focus on protein-protein interactions and use Negatome [Smialowski et al., 2010] to generate negative data, a knowledge base that contains protein pairs which are known to not interact with each other. Nevertheless, in most of the cases knowledge bases do not contain negative relations. For this reason negative data is often generated following a closed world assumption (such as Nguyen and Moschitti [2011a]; Takamatsu et al. [2012]). If an entity pair is not known as instance of a relation, the entity pair is considered to be negative. Those negative entity pairs can be generated by creating new combinations between existing entities of a relation. However knowledge bases are often incomplete, thus this process might lead to false negative training data.

Data annotated using distant supervision is noisy and unlikely to be of as high quality as manually labelled data (see e.g. analysis in [Riedel et al., 2010]). Despite this, distantly supervised relation extraction provides reasonable results compared to those based on supervised learning (see e.g. in [Thomas et al., 2011]). Furthermore, distant supervision has the advantage that large amounts of training data can be generated without the need for manual annotation.

In context of distant supervision, many approaches exist with similar techniques but use a different terminology. Contrarily, some of those terms are also used in a different context. To provide a clear definition, in this thesis distantly supervised relation extraction is defined as follows:

1. Training data is labelled according to prior knowledge (knowledge base, seeds).
2. The distantly labelled data is used as input to train a machine learning classifier.
3. Training data is generated only once; that means, no iterative process is involved to gather more data (such as usually used in bootstrapping approaches). In other words, no iterative process is used to increase training data in order to iteratively increase the amount of distantly labelled data.

---

## Introducing History of Distant Supervision

The term ‘distant supervision’ was introduced by Mintz et al. [2009] in the context of relation extraction. The authors use Freebase [Bollacker et al., 2008], a large semantic knowledge base, and label sentences of Wikipedia for training. However, Craven and Kumlien [1999], Wu and Weld [2007] and Bunescu and Mooney [2007] are considered as the first approaches applying distant supervision.

Craven and Kumlien<sup>3</sup> introduced the technique of distantly labelling data using prior knowledge. The authors used the Yeast Protein Database (YPD) [Hodges et al., 1998] to label training data from Medline abstracts (see Section 2.3.1) for subcellar-localisation relations. Their relation extraction approach uses bag-of-word features with a Naive-Bayes classifier.

Some years later Wu and Weld [2007] and Bunescu and Mooney [2007] introduced a technique using a set of positive and negative seed instances to generate training examples from a web corpus. The approaches have similarities to bootstrapping methods, without the need of an iterative knowledge acquisition, but with techniques from machine learning. Wu and Weld apply a technique where info-boxes of Wikipedia are used to label training data from Wikipedia articles. Classifiers are trained using a conditional random field (CRF).

The interesting aspect of the work of Bunescu and Mooney is the introduction of multi-instance learning (MIL) into the use case of noisy labels. MIL is a technique which deals with incomplete knowledge about labels of training data. The approach was originally introduced by Dietterich et al. [1997] in context of detecting drug molecules which bind to a target protein. While some known molecules bind very well to proteins, other known ones do not bind well. Proteins may have different shapes but it is not clear which shapes make the molecule bind well. The authors propose a technique to detect useful shapes of binding molecules, in which each molecule is modelled as ‘bag’ containing its different shapes. A bag with different shapes of a molecule that is known to bind well is labelled as positive bag. Conversely, a bag with the different shapes of a molecule known to bind not well is labelled as negative. The paper shows that the new method outperforms other approaches which do not take this

---

<sup>3</sup>Note, the authors use the terminology ‘weak supervision’, but according to different publications (e.g. [Riedel et al., 2010; Surdeanu et al., 2012]) the work is considered as the first approach using distant supervision.

---

multi-instance characteristic into account [Dietterich et al., 1997].

Distantly labelled training data has a similar characteristic. Known facts provide information about which pair of entities express a target relation. Conversely, it is not known which sentences explicitly express the relation and which not. Hence, sentences containing a positive entity pair can be considered as a bag of examples with an unknown number of sentences expressing the target relation. Bunescu and Mooney use an extended relation extraction approach using a SVM to handle MIL in context of corporate acquisitions and person-birthplace relations. Multi-instance learning became a very popular and successful technique for distant supervision (such as in [Liu et al., 2014b; Ritter et al., 2013; Surdeanu et al., 2012]).

Riedel et al. [2010] improve multi-instance learning. Authors such as Craven and Kumlien or Mintz et al. model the distantly supervised problem as single-instance single-label supervised learning [Surdeanu et al., 2012]. However, Riedel et al. introduced a novel graphical model that assumes that at least one sentence expresses the target relation, if a (multi-instance) bag is labelled as positive. The model is able to predict relations between entities, but also sentences which express the relations. This setup is different to the previous approaches using distant supervision. Hoffmann et al. [2011] extends the models of Riedel et al. to a multi-instance multi-label problem, by taking into account that entity pairs can occur in different relations at once (overlapping relations) (e.g. a person might be the CEO and the founder of a company). The system developed is called MultiR and is publicly available<sup>4</sup>, easily adjustable for different knowledge bases and used for the experiments in this thesis (see Section 4.5).

A range of different publications focussed on further improvements of the MIL method (such as [Ling et al., 2013; Min et al., 2013; Ritter et al., 2013; Surdeanu et al., 2012]). However, other machine learning methods have been used in context of distant supervision, such as multi-class logistic regression classifier [Augenstein et al., 2014; Mintz et al., 2009], SVM (Bobic et al. [2012]; Nguyen and Moschitti [2011a]) or Neural Networks ([Zeng et al., 2015]) for instance.

An overview of distantly supervised relation extraction from out and inside the biomedical domain is now provided. Finally, some mixture models (supervised and distantly supervised) are presented.

---

<sup>4</sup><http://raphaelhoffmann.com/mr/>

---

## **Non-Biomedical Distant Supervision**

Distant supervision was introduced in biomedical context [Craven and Kumlien, 1999]. Nonetheless, the technique first found popularity outside of the biomedical domain (see e.g. Bunescu and Mooney [2007]; Mintz et al. [2009]; Wu and Weld [2007]). Mintz et al. introduced the use of Freebase, which became a very popular use case for many different approaches working with distantly supervised relation extraction (such as [Abad and Moschitti, 2014; Min et al., 2013; Takamatsu et al., 2012; Yao et al., 2010; Zhang et al., 2013]). Some other approaches use Wikipedia’s info-boxes (such as [Hoffmann et al., 2010; Wu and Weld, 2007]), YAGO [Suchanek et al., 2007], a knowledge base derived from WordNet and Wikipedia (such as Nguyen and Moschitti [2011a]) or DBPedia [Auer et al., 2007] (such as Aprosio et al. [2013]).

Within the last few years many approaches focussed on further improvements on distant supervision in the non-biomedical domain. A large range of methods aim at the reduction of noise (such as Intxaurreondo et al. [2013]; Nguyen and Moschitti [2011a]; Ritter et al. [2013]; Roth et al. [2013]; Takamatsu et al. [2012] for instance) or the reduction of false negatives (such as [Min et al., 2013; Xu et al., 2013] for instance). Takamatsu et al. [2012] for example work on the reduction of incorrect annotations or Xu et al. [2013] face the problem of knowledge base gaps. Some other approaches focus directly on the noise in the data: Intxaurreondo et al. [2013] use a range of heuristics to remove noise in the distantly labelled data and Augenstein et al. [2014] apply techniques to detect highly ambiguous entity pairs and discard them from their labelled training set.

## **Biomedical Distant Supervision**

In recent years there has been also an increasing interest on distant supervision in the biomedical domain. Rather than focussing mainly on one source (such as Freebase) a wide range of different knowledge sources have been explored. This might be connected to the fact that Freebase contains a large amount of information from outside of the biomedical domain. On the other hand, there is a large range of different biomedical topics and domains that are spread across different knowledge bases.

Craven and Kumlien [1999] introduced the idea of distantly supervised relation extraction using YPD (Yeast Protein Database). For many years the technique did not

---

attract any attention in the life sciences. Thomas et al. [2011] started using distant supervision in context of protein-protein interactions (PPI) using IntAct and Negatome. Bobic et al. [2012] extended the approach of Thomas et al. by exchanging the knowledge base to Kansas Proteomics Service (KUPS) database [Chen et al., 2011], which unifies several PPI knowledge sources. Furthermore, the authors use the database DrugBank [Knox et al., 2011] to detect drug-drug interactions as well. Ravikumar et al. [2011] proposed a method to find protein-residue associations using the Protein Data Bank (PDB) [Berman et al., 2000], Ling et al. [2013] detect meronyms (such as *has-part*), Ellendorff et al. [2014] use the Comparative Toxicogenomics Database (CTD) [Davis et al., 2009] to detect interactions between genes and chemicals and Poon et al. [2015] the Pathway Interaction Database (PID) [Schaefer et al., 2009] for the extraction of cancer pathway. As the overview shows, there is a wide range of different biomedical knowledge bases and, therefore, distant supervision can be used for different biological and medical fields.

Various approaches have used UMLS (see Section 2.3.2 for more information), as a knowledge source. Nikolova and Angelova [2011] applied distant supervision to *is\_a* relations of the UMLS Metathesaurus, as well as the relation *affect* of the UMLS Semantic Network. Liu et al. [2014b] detect genes in brain regions (*location-of*) from literature using the UMLS Semantic Network. Finally, the work of Tymoshenko et al. [2012] focussed on a method to mine relations of the National Drug File Repository Thesaurus (NDF-RT or NDFRT) a subset of UMLS Metathesaurus. Their technique relies on entity-level semantics and uses hierarchical information of UMLS to extract relation from text. As comparison to their proposed technique, the authors present a distantly supervised model using some relations of NDF-RT. However, their main focus is entity-level semantics with UMLS taxonomy and (Wikipedia) link features and not distant supervision. Furthermore, their setup for the distantly supervised classifier appears to be unrealistic with a bias<sup>5</sup> 1:1. Usually the amount of positive and negative instances are highly unbalanced towards negative data which makes classification much more challenging.

Many different knowledge bases have been tested and used for distant supervision. The UMLS Metathesaurus is a large medical knowledge base which has been used for many different NLP tasks. However, the Metathesaurus has not been examined in

---

<sup>5</sup>In this context bias represents the ratio of positive and negative instances to each other.

---

detail whether it is a useful source for distantly supervised relation extraction. This issue will be explored in this thesis.

### **2.1.3 Mixed Classification Models**

Supervised relation extraction often provides reliable results but requires a sufficient amount of manually labelled training data. Unfortunately, a sufficient number of this data is not always available and its generation can be time consuming and expensive. Conversely, distantly supervised relation extraction is able to produce large amount of data, but of a varying quality, thereby reducing the classification results. Some approaches make use of both techniques using machine learning in order to achieve further improvements. Different terminologies are used for the combination of manually labelled and automatically labelled data. One frequently used term is semi-supervised learning, which addresses this task by ‘using large amount of unlabeled data, together with the labeled data, to build better classifiers’ [Zhu, 2006].

Nguyen and Moschitti [2011b], Pershina et al. [2014] and Angeli et al. [2014] focus on improving distant supervision by including manually labelled data. Nguyen and Moschitti [2011b] use a SVM and combine the supervised and the distantly supervised classifier with a linear combination. Pershina et al. [2014] and Angeli et al. [2014] integrate the manually labelled data directly within their distantly supervised multi-instance learning approach.

Kordjamshidi et al. [2015] instead focus on improving supervised learning by using distantly labelled data. The authors argue that: ‘The main issue of a fully supervised system is the difficulty to generalise towards unseen patterns. This problem is more apparent the sparser the data, and the richer the representation.’. Their technique is tested in biomedical context (Gene Regulation Network).

Chapter 3 presents the use of manually and distantly labelled data in order to improve classification results. In contrast to previous approaches, Chapter 3 explores the impact of distantly labelled data in combination with an increasing number of manually labelled training instances. The presented approach is tested in context of adverse drug effects.



---

## 2.2 Evaluating Relation Extraction Systems

This section discusses how distant supervision can be evaluated. Usually distant supervision is applied when no annotated training data is available. However, this often means that also no data is available for the evaluation, which raises the issue of how such a system can be evaluated. Some previous approaches have made use of existing labelled data sets (with annotations similar to the related information in the knowledge base) to evaluate approaches based on distant supervision, such as Thomas et al. [2011] and Bobic et al. [2012]. Other approaches such as Craven and Kumlien [1999] generate their own gold standard to annotate relevant relations of their knowledge base. But the effort required to generate manually labelled evaluation data somewhat negates the benefit of distant supervision reducing development time. Unfortunately for many relations no annotated data set is available, thereby different evaluation techniques are required, apart from gold standard evaluation.

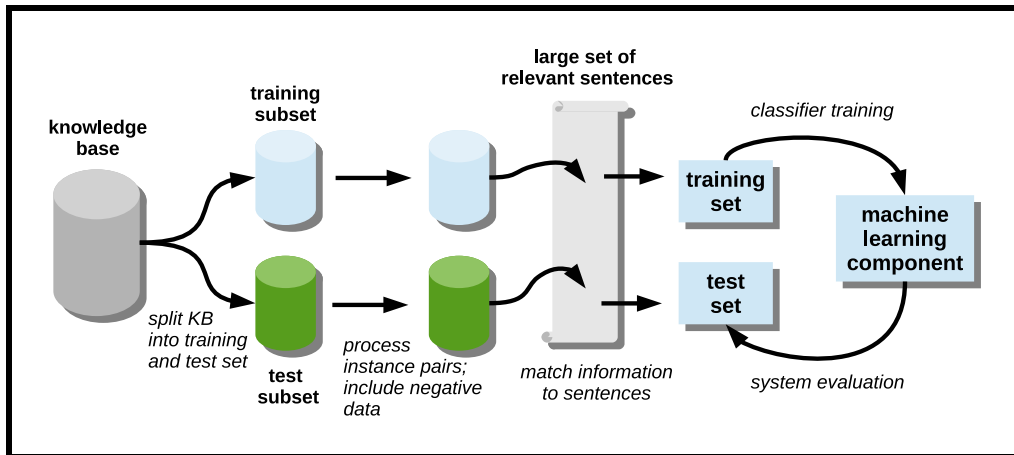


Figure 2.4: Held-Out Evaluation

An alternative approach, which does not require any labelled data, is held-out (also sometimes called hold-out) evaluation. This approach splits facts from the knowledge base into two parts: one to generate distantly supervised training data and the other to generate distantly supervised evaluation data (such as in [Hoffmann et al., 2010; Mintz et al., 2009; Riedel et al., 2010]). Consequently the system is both, trained and evaluated on/with noisy data. Held-out evaluation usually means, that the system is evaluated against the knowledge base (Is it possible to detect the known facts of held-

---

out knowledge base?). Held-out evaluation is visualised in Figure 2.4.

This approach is often combined with a manual evaluation in which a subset of the predictions is selected to be examined in more detail. For example, Riedel et al. [2010] supplemented the held-out evaluation of their distant supervision approach for Freebase by selecting the top 1000 facts and evaluating them manually. Others such as Surdeanu et al. [2012] and Intxaurreondo et al. [2013] work with the same knowledge base and are able to re-use the manually labelled data generated by Riedel et al. [2010]. However, this data is only available for some Freebase relations and evaluation data has to be generated for each new relation. Approaches such as Takamatsu et al. [2012], Zhang et al. [2013] and Augenstein et al. [2014] combine a held-out evaluation with a manual evaluation of a randomly chosen subset or the top-k predictions. This technique is a more reliable evaluation method but requires more effort including (potentially) domain knowledge and needs to be repeated for each version of the classifier.

Held-out evaluation is also used in this work. Held-out evaluation using distantly labelled data is a simple and quick technique for estimating the accuracy of distantly supervised relation extraction systems. Nonetheless, this evaluation data is noisy and it is unclear what effect this has on the accuracy of performance estimates. Chapter 5 examines evaluation on held-out data in comparison to a manually labelled version of the same data, in order to examine its advantages and drawbacks.

### 2.2.1 Evaluation Levels

In this work two evaluation approaches will be used: **sentence level evaluation** and **entity level evaluation**. Sentence level evaluation means that an entity pair will be considered as correct if the two entities express the relation of interest in the sentence they occur. Each single prediction (each sentence with the given entity pair) will be examined and influences the evaluation results. Sentence level evaluation can be useful to find documents and in particular sentences that express relevant information in large data sets.

Entity level evaluation instead, is evaluated based on the fact whether the extracted entity pair represents the target relation (evaluation against KB) or not. Predicted sentences are reduced to a set of entity pairs. A prediction is considered as correct if either a (known) related entity pair is predicted as positive in at least one of the sentences or

---

if an entity pair which is not known to be related, is never predicted as positive. In this context it is not necessarily important, how often an entity pair occurs within a set of given sentences. Correct and incorrect predictions dependent only on the overall prediction of an entity entity pair.

*Findings on discontinuation and rechallenge supported the assumption that the **hair loss** was a side effect of the **paroxetine**. (PMID=10442258)*

Figure 2.5: Correctly predicted adverse drug effect between target entities.

*There are a few case reports on **hair loss** associated with tricyclic antidepressants and serotonin selective reuptake inhibitors (SSRIs), but none deal specifically with **paroxetine**. (PMID=10442258)*

Figure 2.6: Incorrectly predicted adverse drug effect between target entities.

The given example in Figure 2.5 and Figure 2.6 exemplifies both evaluation levels, assuming a classifier predicts an adverse drug effect between *hair loss* and *paroxetine* in both sentences. On sentence level each prediction is examined and evaluated based on the information provided in the sentence. It means, that the classifier predicts the first sentences correct and the second one incorrect. The incorrect prediction reduces the quality of the classifier according the metrics presented in the following subsection.

On entity level evaluation only one prediction is considered for the evaluation of the classifier (for the given example). Assuming that *hair loss* is a known adverse effect of *paroxetine*, at least once the entity pair (*hair loss*, *paroxetine*) has to be predicted as an adverse drug effect. If the entity pair is predicted at least once as positive, the prediction is counted as a *true positive* prediction, otherwise as *false negative*. The ‘false’ prediction of the second sentence does not matter on entity level evaluation since the context where the information occurs does not matter, only whether an positive (correct) entity pair can be detected as instance of the relation.

---

## 2.2.2 Evaluation Metrics

*Precision*, *Recall* and *F1-Score* (often just called F-Score or F1) are very common measures of performance for a classification task, such as relation extraction. The measures base on the correlation between ‘true’ labels and predicted labels presented in Table 2.1. In the following the different measures are explained in detail.

		annotated (gold standard) labels	
		positive	negative
predicted labels	positive	<b>true-positives</b>	<b>false-positives</b>
	negative	<b>false-negatives</b>	<b>true-negatives</b>

Table 2.1: Confusion Matrix: Predicting instances

$pos_t$  is defined as the amount of positively predicted relations that are predicted correctly (**true-positive**) and  $pos_f$  the amount of positively predicted relations predicted incorrectly (**false positive**).  $neg_t$  is the amount of negative predictions that are correctly predicted as negatives (**true-negative**) and  $neg_f$  the amount of negative predictions that are incorrectly predicted as negatives (**false negative**). Then *Precision*, *Recall* and *F1-Score* are defined in the following way:

### **Precision:**

Precision defines the proportion of correctly predicted relations in comparison to all entity pairs predicted as positive.

$$Precision = \frac{pos_t}{(pos_t + pos_f)}$$

### **Recall:**

Recall defines the proportion of correctly predicted relations in comparison with all existing relations in the test set.

$$Recall = \frac{pos_t}{(pos_t + neg_f)}$$

---

**F1-Score:**

The F1-Score is the harmonic mean between Precision and Recall. The F1-Score is the most frequently used F-Measure. However, variations of the F1-Score exist which emphasise *Precision* over *Recall* ( $F_{0.5}$ -Score) or vice versa ( $F_2$ -Score). The F1-Score is defined as follows:

$$F1 = 2 * \frac{Precision * Recall}{(Precision + Recall)}$$

**Micro average F1-Score:**

The Micro average F1-Score is a technique to deal with the prediction and evaluation of multiple classes at the same time. It takes under consideration that some classes involve instances with a higher frequency than other ones. Assuming for example a classification task with two classes (class A with 10,000 instances and class B with 100 instances) and a classifier which is able to predict class A with an F1-Score of 20 and class B with a F1-Score of 80. In the given example the low F1-Score of the class A has a stronger impact on the overall results (micro avg. F1) due to the fact that it contains more instances.

The Micro average F1-Score is calculated by using the harmonic mean of the Micro avg. Precision (MiPrecision) and the Micro avg. Recall. (MiRecall). MiPrecision and MiRecall are generated using the true-positives and false-positives of each different class ( $c \in C$ ), whereas  $pos_{t_c}$  defines the amount of true-positives of class  $c$  and  $pos_{f_c}$  the amount of false-positives of class  $c$ :

$$MiPrecision = \frac{\sum_{c \in C} pos_{t_c}}{\sum_{c \in C} (pos_{t_c} + pos_{f_c})}$$

$$MiRecall = \frac{\sum_{c \in C} pos_{t_c}}{\sum_{c \in C} (pos_{t_c} + neg_{f_c})}$$

## 2.3 Resources

In the following some key resources that are important for this work are introduced.

---

### 2.3.1 Medline repository

Medline is a large repository containing abstracts and links to complete publications from life science and biomedical domain. The data is made freely available on the Internet by the United States National Library of Medicine (NLM) and can be accessed and searched by the search engine PubMed<sup>6</sup>. Medline contains selected publications from generally 1946 to present<sup>7</sup> (with some older material) and is a permanently growing data base. Every day approximately 2,000-4,000<sup>7</sup> completed references have been added since 2005. Currently Medline contains more than 22 million references<sup>7</sup> of biomedical publications. Sentences extracted from Medline abstracts are used to conduct the experiments in this work. Many examples given in this thesis are provided with a PMID (PubMed-ID), which can be used to find the abstract on PubMed the sentence was extracted from.

### 2.3.2 UMLS

The Unified Medical Language System<sup>8</sup> (UMLS) is a large biomedical knowledge base containing millions of medical terms and relations between them, and will be used to train distantly supervised classifiers in this work. Overall UMLS can be divided into three different parts: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The SPECIALIST Lexicon provides lexical information about medical terms, for instance it defines words as nouns or adjectives. For this work only the UMLS Metathesaurus and parts of the Semantic Network are relevant.

The Metathesaurus is the core of UMLS and unifies existing biomedical knowledge bases (vocabularies), such as the Gene Ontology (GO) or the National Drug File - Reference Terminology (NDFRT). Currently UMLS contains more than 120 different vocabularies. The vocabularies that include the largest number of medical concepts are SNOMED-CT, NCBI and MSH.

Medical concepts can be described in different ways with different spellings, different abbreviations and also in different languages. UMLS unifies those variations using the Concept Unique Identifier (CUI). The two most important parts of the UMLS

---

<sup>6</sup><https://www.ncbi.nlm.nih.gov/pubmed/>

<sup>7</sup><http://www.nlm.nih.gov/pubs/factsheets/medline.html>, date: 26th of September 2015

<sup>8</sup><https://www.nlm.nih.gov/research/umls/>

---

Metathesaurus for this work are the subsets MRCONSO and MRREL. Both subsets can be seen as two large tables.

### **MRCONSO - UMLS Metathesaurus**

MRCONSO contains concepts, concept names and their identifiers. It maps medical terms across the different vocabularies to a CUI. It contains information about different vocabularies and languages. For instance, the version used in this work<sup>9</sup> contains six different entries in MRCONSO for the string (in lower case) ‘*headache*’ involving two different CUIs (*C0018681* and *C2096315*) across five different vocabularies. Conversely, 268 MRCONSO entries are found for CUI=*C0018681*. These are defined within 51 different vocabularies across 17 different languages. For English alone it is possible to find 51 different strings such as ‘*Headache*’, ‘*headache*’, ‘*Pain head*’, ‘*Cranial Pains*’ or ‘*cephalalgia*’. An excerpt of MRCONSO for the CUI *C0018681* is given in Figure 2.2. In the example the most relevant information for this thesis are highlighted (CUI, source vocabulary and string).

### **MRREL - UMLS Metathesaurus**

MRREL defines binary relations between medical concepts. Each relation instance is defined by a pair of CUIs. MRREL contains a range of different information for each entry. However, the most important information beside the two related CUIs are the relation label (*REL*), the relation attribute label (*RELA*) and the source abbreviation (*SAB*), which refers to its source vocabulary. *REL* can be assigned to one of 12 different labels (see Table 1 in Appendix 1<sup>10</sup>) and defines a more general concept of a relation. The *RELA* labels are very useful for this thesis and provide a more detailed name for each relation, such as *has\_tradename*, *associated\_morphology\_of* or *part\_of*. If an MRREL instance does not contain any *RELA* label it is not always apparent which relation is described between the two CUIs.

For CUI *C0018681* (‘*headache*’) the MRREL table contains 2809 entries across 47 different vocabularies. The same CUI pair (relation) can be defined within different

---

<sup>9</sup>Version 2013AA.

<sup>10</sup>Data extracted from <https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CCC/relationships.html>.

```

C0018681|ENG|P|L0018681|VO|S|1640551|Y|A|1601383|||10019211|MDR|LT|10019218|Headache NOS|3|N|1024|
C0018681|ENG|S|L0272777|PF|S|1844481|Y|A|18684548|0000019413|000005820||CHV|SY|000005820|ache head|0|N|1792|
C0018681|ENG|S|L0272777|VW|S|11895648|Y|A|18628891|0000019419|000005820||CHV|SY|000005820|head ache|0|N|1536|
C0018681|ENG|S|L0290361|PF|S|0475640|N|A|2957274|41993017|25064002||SNOMEDCT|SY|25064002|Cephalalgia|9|N|3840|
C0018681|ENG|S|L0306029|PF|S|1688029|N|A|18014038||N000001418||NDFRT|SY|N000001418|Head Pain|0|N|1792|
C0018681|ENG|S|L0306029|PF|S|1688029|Y|A|1648998||M0009824|D006261|MSH|EN|D006261|Head Pain|0|N|1792|
C0018681|ENG|S|L0306029|VW|S|1916340|Y|A|1860951|||ICPC2E|PT|N01004|Pain;head|3|N||
C0018681|ENG|S|L0306029|VC|S|6653130|N|A|18573077|0000019420|000005820||CHV|SY|000005820|head pain|0|N|1536|
C0018681|ENG|S|L0306029|VO|S|0982847|Y|A|8364407|||784.0|MTHICD9|ET|784.0|Pain in head NOS|0|N|1536|
C0018681|ENG|S|L1406212|PF|S|1680378|N|A|17990267||N000001418||NDFRT|SY|N000001418|Cranial Pain|0|N|1792|
C0018681|ENG|S|L1406212|PF|S|1680378|Y|A|1641923||M0009824|D006261|MSH|EN|D006261|Cranial Pain|0|N|1792|
C0018681|ENG|S|L1406212|VW|S|1680380|Y|A|1641925||M0009824|D006261|MSH|PM|D006261|Pain, Cranial|0|N|1792|
C0018681|ENG|S|L17659919|PF|S|8880524|Y|A|18037960||N000001418||NDFRT|FN|N000001418|Headache [Disease/Finding]|0|N|1536|

```

Table 2.2: Excerpt of MRCONSO

```

1) C0000052|A0016535|SDUI|SIB|C0017915|A0064131|SDUI||R71296977||MSH|MSH||N||
2) C0000163|A0016691|SDUI|CHD|C0020268|A0070119|SDUI||R71355474||MSH|MSH||N||
3) C0000039|A0016515|SDUI|AQ|C0005768|A3879709|SDUI||R120502291||MSH|MSH||N||
4) C0000618|A17996759|AUI|RO|C1177390|A18008188|AUI|has active met abolites|R115950896||NDFRT|NDFRT||N||
5) C0000039|A0016514|AUI|SY|C0000039|A1317707|AUI|permutated term of|R28482431||MSH|MSH||N||
6) C0032951|A20635134|SCUI|RN|C1382748|A10666793|SCUI|isa|R1301865261|RXNORM|RXNORM||N||
7) C1382748|A10666793|SCUI|RE|C0032951|A20635134|SCUI|inverse isa|R130162179||RXNORM|RXNORM||N||

```

Table 2.3: Excerpt of MRREL



---

vocabularies, which means that only 1046 of those CUI pairs are unique. Moreover, a large amount of CUI pairs do not assign a RELA label. However, CUI *C0018681* occurs within 28 different relations (with RELA labels), including *isa*, *may-treat*, *has-expanded-form*, *induces* or *disease-may-have-finding*.

Figure 2.3 presents some example instances of MRREL. Slot 1 and Slot 5 contain the related CUI pair, Slot 4 the REL label, Slot 8 RELA label and Slot 11 SAB (mentioned positions are highlighted). Example 1) contains a sibling (SIB) relation between the the CUIs (C0000052, C0017915) (*'1,4-alpha-Glucan Branching Enzyme'*, *'Transfer-Glucosidase'*). The instance does not specify a value in the field RELA and it is defined within the source vocabulary MSH (Medical Subject Headings; MeSH). Example 2) presents a parent-child relation between two CUIs (C0000163, C0017915) (*'17-Hydroxycorticosteroids'*, *'Hydrocortisone'*). Example 3) defines a vaguely defined relation (C0000039, C0005768) (*'1,2-Dipalmitoyl-Glycerophosphocholine'*, *'blood'*). The instance does not contain any RELA label which specifies the relationship in more detail. Moreover, the label REL is defined as AQ (allowed qualifier), which is very general and possibly not useful for distant supervision. Example 4) instead contains a relation of NDFRT and is defined as *has\_active\_metabolites* (C0000618, C1177390) (*'6-Thiohypoxanthine'*, *'Azathioprine 75 MG Oral Tablet'*).

Usually relations are directed and defined in both directions (see Table 2.3, Example 6) and Example 7)): For example MRREL contains the relation *disease-may-have-finding* for the CUI-pair (C0018681,C0025149) (*'Headache'*, *'Glioblastoma'*). Thus, MRREL contains also the reverse CUI-pair (C0025149, C0018681) for the relation *may-be-finding-of-disease*. To reduce redundancy reverse relations of UMLS are ignored for this work.

## Semantic Network

The Semantic Network assigns semantic types (STY) to each medical concept and defines semantic relations between the different semantic types. CUI *C0018681* for instance, refers to the semantic type *'Sign or Symptom'*, whereas *'Sign or Symptom'* can be involved in 68 different relations of the Semantic Network, such as *isa*(*'Sign or Symptom'*, *'Finding'*), *diagnoses*(*'Sign or Symptom'*, *'Injury or Poisoning'*) and *treats*(*'Drug Delivery Device'*, *'Sign or Symptom'*). Relations in the Semantic Net-

---

work are more general than those in the UMLS Metathesaurus. Obviously, not every CUI that is assigned to ‘Drug Delivery Device’ can treat a headache.

### 2.3.3 MetaMap

MetaMap [Aronson and Lang, 2010] is a system to map UMLS concepts to text documents and can be used to carry out a named entity recognition. In order to generate distantly labelled data in this work, sentences containing medical concepts related in UMLS have to be selected for further processing. However, UMLS defines relations as CUI pairs. For this reason it is essential to ‘annotate’ (map) UMLS-CUIs in text, such as medical abstracts from the Medline repository. These annotations are provided by MetaMap.

In order to map CUIs to sentences of the medical domain, MetaMap applies a range of different lexical and syntactic processing steps [Aronson and Lang, 2010]. First, MetaMap runs a tokenisation, sentence boundary detection and acronym/abbreviation identification, followed by a part-of-speech tagging. Next, it applies a lexical lookup of input words using the UMLS SPECIALIST Lexicon. Finally, phrases and their syntactic heads are identified using a shallow parse. Each phrase is then further examined in the following way: First, variants of the phrase words are generated (normally using a lookup table), succeeded by an identification of possible UMLS Metathesaurus concepts mapping to words in the phrase. This resulting list of candidates are measured based on how well the input string is matched. In a next step different mappings (combinations) of MetaMap candidates of the given phrase are generated. Finally, a Word Sense Disambiguation (WSD) can be applied in order to find its favoured mapping of Metathesaurus concepts. MetaMap is highly configurable which influences the candidate mapping, the selection of the most appropriate concepts and also the output. An MetaMap example of the sentence ‘AIMS : To study the distribution of clinically important red cell antibodies in pregnancy, and the associated fetal and neonatal morbidity and mortality.’ (*PMID=9536844*) is provided in the following. Figure 2.7 shows how MetaMap segments the sentence into different phrases (phrases are separated by ‘—’). Next, MetaMap searches for possible CUIs for each phrase. Table 2.4 depicts the list of CUI candidates for the fifth phrase ‘the distribution of clinically important red cell antibodies’. Figure 2.8 presents the final UMLS-CUIs mapping according to

the highest ranked mapping results.

AIMS | : | To | study | the distribution of clinically important red cell antibodies | in pregnancy, | and | the associated fetal | and | neonatal morbidity | and | mortality.

Figure 2.7: MetaMap-segmented sentence

C0520511	distribution (Distributing) [Idea or Concept]
C1704711	Distribution [Functional Concept]
C2698777	DISTRIBUTION (Pharmacokinetics: Distribution) [Regulation or Law]
C0037775	Distributions (Spatial Distribution) [Spatial Concept]
C0014792	Red Cell (Erythrocytes) [Cell]
C0007634	THE CELL (Cells) [Cell]
C0003241	Antibodies [Amino Acid, Peptide, or Protein, Immunologic Factor]
C0332575	Red (Redness) [Finding]
C1260956	Red (Red color) [Qualitative Concept]
C1269647	Cell (Entire cell) [Cell]
C1704653	Cell (Cell Device Component) [Medical Device]
C1948049	Cell (Cell (compartment)) [Spatial Concept]
C0021027	Antibody (Immunoglobulins) [Amino Acid, Peptide, or Protein, Immunologic Factor, Pharmacologic Substance]
C3282337	Cells (Cells [Chemical/Ingredient]) [Cell]
C0178539	Cellular [Functional Concept]
C1516377	Cellularity [Qualitative Concept]

Table 2.4: CUI candidate list (including semantic types) provided by MetaMap

According to Aronson and Lang [2010] one of the weaknesses of MetaMap is its focus on English text. Another drawback of the system is its slow processing time (not appropriate for real-time use) and its reduced accuracy in context of ambiguous concepts. In the example in Figure 2.8, CUI *C0332575* connected to the colour red refers to the semantic type ‘Finding’ and has the following definition: ‘Coloration of the skin; sign of inflammation’. This does not seem to be appropriate for the given

---

[C1947946:AIMS]: To [C0008976:study] the [C0520511:distribution] of clinically important [C0332575:red] [C0007634:cell] [C0003241:antibodies] in [C0032961:pregnancy], and the [C0332281:associated] [C0015965:fetal] and [C1552240:neonatal] [C0026538:morbidity] and [C0026565:mortality].

Figure 2.8: MetaMap annotation of UMLS-CUIs

context. Moreover, Oellrich et al. [2015] shows that tools such as the NCBO Annotator [Jonquet et al., 2009] and the clinical Text Analysis and Knowledge Extraction System (cTakes) [Savova et al., 2010] often provide better results (in terms of F-Score) compared to MetaMap.

However, MetaMap has been used for this thesis due to a range of positive aspects. First, MetaMap is freely available<sup>11</sup>, widely used [Simpson and Demner-Fushman, 2012] and a popular tool to annotate UMLS concepts to natural language (used e.g. by Hanauer et al. [2014]; Liu et al. [2014a]; Preiss et al. [2015]). In this work one million Medline abstracts are used in order to generate distantly labelled training data, which might be a problem, due to the slow processing time of MetaMap. However, a Medline subset which contains already MetaMap annotations was found on the NLM webpage<sup>12</sup> and was used for this work<sup>13</sup>.

## 2.4 Summary

This chapter presented an overview of related work and relevant resources. Firstly, the chapter provided an introduction into supervised relation extraction, presenting a range of different biomedical state-of-the-art approaches to detect relations in natural language. It showed that SVMs are successful and popular approaches to the relation extraction task. Next, distantly supervised relation extraction was presented, including the introduction of the general idea, some approaches and knowledge sources

---

<sup>11</sup><http://metamap.nlm.nih.gov/>

<sup>12</sup>The subset can be downloaded here: [http://mbr.nlm.nih.gov/Download/MetaMapped\\_Medline/](http://mbr.nlm.nih.gov/Download/MetaMapped_Medline/) and further information can be found here: <http://skr.nlm.nih.gov/resource/MetaMappedBaselineInfo.shtml>

<sup>13</sup>MetaMap version 13 with UMLS 2013AA is used.

---

used. This section highlighted that multi-instance learning is a successful and popular technique to handle distantly labelled data. Then, different evaluation techniques for distantly supervised relation extraction were presented. Finally, various important resources for this work, including UMLS and MetaMap were presented.

## Chapter 3

# Bootstrapping Limited Training Data

Supervised machine learning techniques have proved to be the most effective approach to detect relations between entities in natural language (see Chapter 2). However, they require labelled training data which may not be available in sufficient quantity (or at all) and is expensive to produce. This chapter proposes a technique that can be applied when only limited training data is available. The approach uses a form of bootstrapping or distant supervision (according to the given definition in Chapter 2.1.2) and does not require an external knowledge base. Instead, it uses information from the training set to acquire new labelled data and combines it with manually labelled data. The described technique has similarities to minimally supervised bootstrapping techniques, semi-supervised learning and techniques described in Section 2.1.3 (e.g. Nguyen and Moschitti [2011b] and Kordjamshidi et al. [2015]). The approach can be considered also as a kind of distant supervision.

The goal of this chapter is to identify under which circumstances distantly labelled data can be used to support supervised learning. Using a small set of manually labelled instances (gold standard), a supervised classifier is trained to detect adverse-drug effects. In parallel, seeds (known positive and negative facts) are extracted from the training data to automatically label further positive and negative examples using distant supervision (see Section 2.1.2). In addition to the supervised classifier, two further classifiers are trained, one using the noisy (but large set of) distantly labelled data as input and the other one a mixture of the manually labelled and the distantly labelled data. The chapter examines, when gold standard data set is better than the larger distantly labelled data set and vice versa.

---

The chapter is structured as follows. The next section introduces the ADE data set which is used for the experiments. The techniques for generating the distantly supervised training data and relational classifier are described in Section 3.2, followed by an introduction of the relation extraction system in Section 3.3. Section 3.4 describes the experiment with a data analysis and the results. The chapter ends with a conclusion in Section 3.5.

## 3.1 ADE Data

The experiments in this chapter use the ADE data set [Gurulingappa et al., 2012b] which contains examples of adverse drug effects (ADE). An ADE can be defined as follows: ‘Adverse drug effect is a response of a drug which is noxious and unintended, and which occurs at doses normally used in humans for the prophylaxis, diagnosis, therapy of disease, or for the modification of physiological function.’<sup>1</sup> [Gurulingappa et al., 2012b]. ADEs are responsible for one of the most common causes of death in industrialised nations and are the fourth leading cause of death in the U.S. [Giacomini et al., 2007]. To reduce this risk, the side-effects of drugs need to be detected and made publicly available as quickly as possible. Relation extraction can be used to support the detection of adverse drug effects.

The ADE data set has been used by Gurulingappa et al. [2012a] and Kang et al. [2014] in context of relation extraction. Gurulingappa et al. [2012a] address the problem by using a SVM-based classifier, as described in Section 3.3. Kang et al. [2014] instead, rely on a data-driven method, using the shortest path between candidate concepts within UMLS. Both methods provide very promising results, however, a direct comparison is not possible, since generation of parts of the data and exact split into training and test data remains unknown.

### 3.1.1 Corpus overview

The ADE data set consists of Medline case reports examined by three human annotators. Sentences in these case reports containing adverse effects between *drugs* and

---

<sup>1</sup>World Health Organization (WHO) glossary of terms used in Pharmacovigilance.

---

*conditions* were extracted and entities annotated to generate the data set. An example relation between a drug and a condition from this data set is shown in Figure 3.1. According to the given sentence the condition *pseudoporphyria* is caused by the two drugs *naproxen* and *oxaprozin*.

```
METHODS: We report two cases of [CONDITION:pseudoporphyria]
caused by [DRUG:naproxen] and [DRUG:oxaprozin].
(PMID=10082597)
```

Figure 3.1: Example of a drug-related adverse effect

The ADE corpus only contains examples of positive relations. Negative examples are also required to set-up a meaningful ADE prediction task and to train a supervised ADE classifier. A set of negative examples are generated in a similar way as in Kang et al. [2014]:

First, named entity recognition is used to detect drugs and conditions. Thus, MetaMap (see Section 2.3.3) was run on the unannotated sentences of the ADE corpus to detect biomedical concepts from UMLS. Each annotated UMLS-CUI in the sentence can be mapped to a semantic type (see Section 2.3.3). Only those CUIs which refer to a semantic type which belongs to one of the two groups “Chemicals & Drugs” and “Disorders” (according to the definition of Bodenreider and McCray [2003], see Table 2 in Appendix 1) are kept; all other CUIs are removed. Negative data is generated between a CUI referring to the group “Chemicals & Drugs” and a CUI referring to the group “Disorders”. Similar as in the classification tasks of Kang et al. [2014] and Gurulingappa et al. [2012a], nested relations<sup>2</sup> are not considered in this generation process.

An example of the negative relation generation process is illustrated in Figures 3.2-3.4. Figure 3.2 presents a sentence with UMLS-CUI annotations provided by MetaMap. Figure 3.3 shows the different semantic types of each CUI. As seen in the figure, some CUIs can match to various semantic types. Next, entities with semantic types that do not belong to the semantic groups “Chemicals & Drugs” or “Disorders”

---

<sup>2</sup>In case of a nested relation, one of the entities contains the other one. In the following example *caffeine intoxication*, which is the adverse effect of *caffeine*, embeds the other entity: ‘Severe rhabdomyolysis following massive ingestion of oolong tea: **caffeine intoxication** with coexisting hyponatremia.’ (PMID=10592946)



---

```
[C0060483:Fludarabine monophosphate], a [C1268902:purine analogue] is [C1524063:used] in the [C0001554:treatment] of [C1518071:lymphoid] [C0006826:malignancies]. (PMID=12111771)
```

Figure 3.2: Mapping UMLS annotations to a sentence using MetaMap

```
[Nucleic Acid, Nucleoside, or Nucleotide, Pharmacologic Substance:Fludarabine monophosphate], a [Biologically Active Substance, Nucleic Acid, Nucleoside, or Nucleotide:purine analogue] is [Functional Concept:used] in the [Occupational Activity:treatment] of [Qualitative Concept:lymphoid] [Neoplastic Process:malignancies]. (PMID=12111771)
```

Figure 3.3: Replacing UMLS-CUIs with semantic types

(see Table 2 in Appendix 1) are removed. The names of the remaining entities are replaced with the name of the semantic group, as shown in Figure 3.4. Finally, negative data is generated by creating new combinations between entities from the semantic group “Chemicals & Drugs” with entities from the semantic group “Disorders”. According to the given example the two negative combinations (*fludarabine monophosphate*, *malignancies*) and (*purine analogue*, *malignancies*) can be created.

The resulting set of positive and negative instances is used to generate training and evaluation sets. A set of 1644<sup>3</sup> ADE abstracts was used for the experiment. 500 abstracts were used to create training data and the remainder used to form the evaluation set.

## 3.2 Automatic Generation of Additional Training Data

Distant supervision uses information about related instances (e.g. drugs and known adverse effects) to automatically generate training data. In the majority of cases this information is obtained from a knowledge base (see Section 2.1.2). This chapter presents

---

<sup>3</sup>That is the amount of abstracts available containing at least one positive instance.

[**Chemicals & Drugs:Fludarabine monophosphate**], a [**Chemicals & Drugs:purine analogue**] is used in the treatment of lymphoid [**Disorders:malignancies**]. (PMID=12111771)

Figure 3.4: Keep entities that match to semantic groups

an approach that makes use of information from a small set of abstracts. For example, the sentence shown in Figure 3.1 suggests that there are cases when the drugs *oxaprozin* and *naproxen* cause *pseudoporphyria*. Consequently unlabelled sentences containing these two *drug-condition* entity pairs (i.e. *oxaprozin-pseudoporphyria* and *naproxen-pseudoporphyria*) can be treated as positive examples.

The automatically labelled data used for the experiment in this chapter is generated by applying a three stage process (see Figure 3.5):

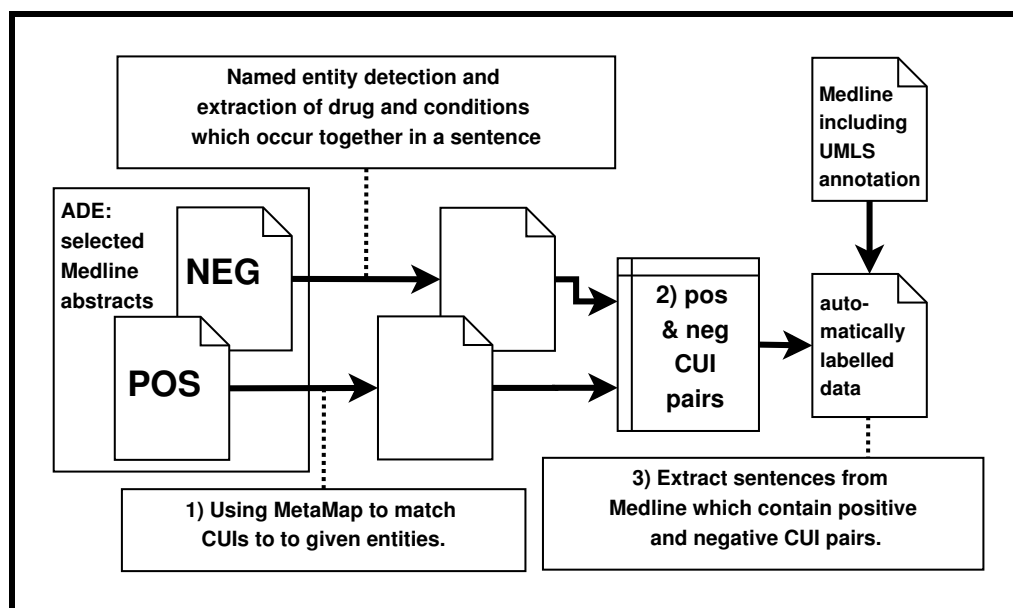


Figure 3.5: Automatic generation of training data for ADE relations

1) *Map CUIs to related entities in the training data set.* First positive sentences in the manually labelled data are normalised<sup>4</sup>. As mentioned already (Section 2.3.2), medical terms can occur in literature with different names, using a different spelling

<sup>4</sup>This step is not necessary for the negative examples, because they already include CUI information for each entity (see Section 3.1).

or abbreviations. For instance *Naproxen* can be also described as *Methoxypropioicin*, *MNPA* or *6-Methoxy-alpha-methyl-2-naphthaleneacetic Acid*. UMLS maps these different names to the same CUI, *C0027396*. MetaMap is used to annotate sentences containing positive examples to obtain UMLS-CUIs for each relevant entity. In many cases it is possible to assign a MetaMap annotation to the existing related entities. Figure 3.6 shows the original ADE sentence and its MetaMap CUI annotation. In the original training data, *pseudoporphyria* is defined as *condition* and *naproxen* and *oxaprozin* as *drug*. MetaMap provides UMLS concepts for each of these entities.

```
METHODS: We report two cases of [CONDITION:pseudoporphyria]
caused by [DRUG:naproxen] and [DRUG:oxaprozin].

[C0025663:METHODS]: We [C0684224:report] [C0205448:two]
[C0868928:cases] of [C0521616:pseudoporphyria]
[C0015127:caused] by [C0027396:naproxen] and
[C0069739:oxaprozin]. (PMID=10442258)
```

Figure 3.6: Positive ADE sentence (top) and its MetaMap annotation (below)

Only CUIs that can be mapped to the entity in its full length (not only a substring) are used. An example for incomplete matches is given in Figure 3.7. The original entity ‘*life-threatening complications*’ cannot be completely matched against a concept in UMLS. Instead, MetaMap provides two UMLS concepts for the entity (C2826244 and C0009566). To reduce the possibility of false data the entities in Figure 3.7 are not used to generate automatically labelled data.

2) *Extract a set of positive and negative seed instance pairs.* In the next step, all CUI pairs from the positive ADE examples are extracted and added to a set of positive instance pairs  $P$ . Furthermore, CUI pairs from negative ADE examples are extracted and added to a negative instance pair set  $N$ . CUI pairs which occur in both sets ( $P$  and  $N$ ), are removed from  $N$ . Considering the example in Figure 3.6 it is possible to extract the positive CUI pairs (*C0027396,C0521616*) and (*C0069739,C0521616*) from the ‘known’ ADE pairs (*naproxen, pseudoporphyria*) and (*oxaprozin, pseudoporphyria*).

3) *Extract sentences containing positive and negative seed instances from abstracts.* The automatically labelled training data is generated using 3,000,000 Medline ab-

---

We now report the first known cancer patient who developed **[CONDITION:life-threatening complications]** after treatment with topical **[DRUG:5-FU]** and was shown subsequently to have profound DPD deficiency.

We **[C1948052:now]** **[C3273238:report]** the **[C1552608:first known]** **[C1516213:cancer patient]** who developed **[C2826244:life-threatening]** **[C0009566:complications]** **[C0001758:after treatment]** with **[C0332237:topical]** **[C0016360:5-FU]** and was **[C1547282:shown]** subsequently **[C1883351:to]** have **[C0439808:profound]** **[C1959620:DPD deficiency]**. (PMID=10473079)

Figure 3.7: Example of an insufficient entity normalisation

The phenylpropionic acid derivative group of nonsteroidal anti-inflammatory drugs, especially **[C0027396:naproxen]**, is known to cause **[C0521616:PP]**. (PMID=17266758)

Figure 3.8: Automatically labelled positive example

stracts with MetaMap-UMLS annotations<sup>5</sup> (see Section 2.3.3). Then sentences from this subset containing positive and negative CUI pairs are extracted and labelled as positive and negative examples. The sentence in Figure 3.8 for instance is a distantly labelled sentence containing the positive CUI pair (C0069739, C0521616).

The automatically generated data is biased towards negative examples. Using a much larger number of negative instances than positive ones has several disadvantages. First, it can lead to a high precision but to a relatively low recall. Second, using approximately 8-10 times more negative instances than positive ones increases the training time required. If the positive data is relatively large it might also lead to memory errors. As seen in Thomas et al. [2011] for instance, the ratio of positive to negative instances influences the classification results. For this reason the bias of the distantly labelled data will be always adjusted to the same ratio as the manually labelled training data being used.

---

<sup>5</sup>Downloaded from [http://mbr.nlm.nih.gov/Download/MetaMapped\\_Medline/](http://mbr.nlm.nih.gov/Download/MetaMapped_Medline/).

### 3.3 Relation Extraction System

This chapter focuses on improving relation extraction when only a small set of manually labelled instances is available. A successful way to detect relations in natural language (using manually labelled data) is supervised learning using a SVM, as seen in the related work (Section 2.1.1). Gurulingappa et al. [2012a], a previous approach using the ADE data set to detect adverse drug effects, used the Java Simple Relation Extraction<sup>6</sup> (JSRE) system to train a supervised classifier. The system bases on the SVM implementation LibSVM [Chang and Lin, 2011]. The classifier includes an implementation of the shallow linguistic (SL) kernel and is a combination of two kernel methods, the *global context kernel* and the *local context kernel*. The global context kernel considers n-grams of the words (and other information such as stemmed words and part of speech tags) between the two entities of the whole sentence. The local context kernel considers only a limited amount (window-size) of information around each entity.

JSRE has been used in different publications [Airola et al., 2008; Giuliano et al., 2006; Segura-Bedmar et al., 2011a]. It has proved to be an effective classifier and is freely available. Furthermore also Gurulingappa et al. [2012a] used JSRE to detect adverse drug effects. Unfortunately a direct comparison is not possible since training split and the exact generation of negative data could not be reproduced. For these reasons however, the experiment in this chapter uses the JSRE system.

Token	Unaccountable severe hypercalcemia in a patient treated for hypoparathyroidism with dihydrotachysterol .
Stemmed Data	unaccount sever hypercalcemia in a patient treat for hypoparathyroid with dihydrotachysterol .
POS	VBP JJ NN IN DT NN VBN IN NN IN NN .
Named Type	O O DISORDER O O O O O O O DRUG O
Entity Label	O O A O O O O O O T O

Table 3.1: Data input for classifier of a positive sentence from publication PMID 10048291, taken from ADE data set.

Sentences are first stemmed using the Porter Stemmer [Porter, 1997]. Next the

<sup>6</sup><https://hlt.fbk.eu/technologies/jsre>

---

Charniak-Johnson Parser [Charniak and Johnson, 2005] is run on each sentence to generate part-of speech (POS) tags. An example of the input data of JSRE is given in Table 3.1. The entity label defines the agent (*A*) and target (*T*) of the relation.

## 3.4 Experiment

The goal of this experiment is to examine the impact of a large set of automatically labelled training data in comparison to a small set of manually labelled instances. Thus, three different methods with a different amount of training data are examined: supervised relation extraction, distantly supervised relation extraction and relation extraction using a mixture-model. The supervised model uses a set of manually labelled abstracts (1-500) as input (seed abstracts; SA). The distantly supervised model takes the bootstrapped data as input (see Section 3.2). Its size varies depending on the size of the manually labelled data. The mixture model merges the automatically generated and manually labelled training data to form a combined training set. The JSRE system, as described in previous section, is used as the classifier.

Starting with a single abstract, the number of seed abstracts is gradually increased to 500. In parallel the automatically labelled data set is generated for each training set, using the given ADE seed facts of the training data for distant supervision. The more information the manually labelled data contains, the more different seeds can be extracted which increases the size of the distantly labelled data. Thereafter both data sets are combined to a mixture-model.

### 3.4.1 Analysis of Generated Data

A maximum of 500 abstracts are used for training. In order to provide reliable results the experiment is repeated five times with a different set of 500 abstracts (randomly selected). In the following this training data is analysed more in detail.

Each setup (of those five runs) contains different positive and negative seeds, thereby producing different data. In the first run, the manually labelled data contains 4064 positive and 4660 negative sentences. Those sentences contain 812 different positive and 1975 different negative CUI pairs, but only 595 different positive and 1018 different negative CUI pair combinations occur within the Medline subset used for this exper-

iment. The combination (*Cyclosporine*, *Nephrotoxicity*), for instance, occurs 1199 times as positive example in the distantly labelled data. Conversely, the same CUI pair occurs only 2 times within the 500 manually labelled training abstracts. 551 of the positive CUI pairs occur fewer than 100 times in the distantly labelled data and 371 fewer than 10 times.

#SA	#seeds		manually lab.		distantly lab.	
	pos	neg	pos	neg	pos	neg
10	12	49	57	108	484	990
25	33	111	168	246	801	1271
50	72	209	378	480	1404	1799
100	154	403	773	921	3444	4036
150	226	610	1162	1378	7278	8591
200	317	805	1573	1840	9140	10683
250	408	1002	2027	2305	11080	12646
300	496	1205	2459	2783	13335	15115
400	673	1615	3332	3760	16795	18961
500	827	2028	4139	4731	20124	23024

Table 3.2: ADE training data size (mean across five runs)

Table 3.2 shows the size of the different sets of training data averaged across five runs. *#seeds* indicates the number of different positive and negative entity pairs extracted from the given abstracts, *manually lab.* indicates the number of manually labelled (gold standard) positive and negative sentences extracted from the seed abstracts and containing the given seeds and *distantly lab.* indicates the number of automatically labelled positive and negative sentences using the given seeds. *#SA* describes the number of abstracts used for training.

Table 3.2 shows that the amount of distantly labelled data is much larger than the manually labelled data at each classification step. Larger amounts of manually labelled data increase the number of ADE seed instances that can be extracted which leads to more distantly supervised examples. 50 training abstracts (SA) contain approximately 72 different positive and 209 different negative seed instances (seeds). Using those seed instances, it is possible to generate approximately 1404 distantly labelled positive and 1799<sup>7</sup> distantly labelled negative instances.

<sup>7</sup>Ratio adjusted to the same as of manually labelled data.

### 3.4.2 Results

The results of the ADE experiments using manually and distantly labelled data are presented in Table 3.3. The results reported are the mean of all five runs. The table presents the results of all three different models (supervised model, distantly supervised model, mixed model), including precision, recall and F1-Score, using a varying amount of training data. The size of the training data is connected to the number of seed abstracts used (see Section 3.4.1).

The results show that the performance for all models improves as the amount of data increases. Performance of the supervised classifier increases sharply as the number of abstracts is increased from 1 to 10. Increasing the size of the training data to 50 abstracts produces a further improvement of approximately 30%. These results demonstrate that even small amounts of training data are sufficient to provide reasonable results on the ADE data set.

#SA	supervised model			distant supervision			mixture model		
	prec.	rec.	F1	prec.	rec.	F1	prec.	rec.	F1
1	63.40	34.91	39.32	52.32	46.40	<b>47.92</b>	53.54	41.14	44.95
5	64.15	49.61	51.52	66.82	40.64	44.44	69.42	51.08	<b>53.65</b>
10	63.85	59.81	57.72	70.29	67.41	68.16	70.19	71.73	<b>69.26</b>
25	70.13	68.11	68.93	71.28	76.93	73.93	73.13	78.72	<b>75.80</b>
50	72.57	78.60	75.46	64.96	91.76	75.98	69.88	89.63	<b>78.47</b>
100	74.56	81.78	77.95	67.17	92.99	77.94	72.64	91.73	<b>81.06</b>
150	76.24	84.69	80.23	70.53	90.82	79.39	74.93	90.91	<b>82.11</b>
200	77.51	85.47	81.29	68.51	93.34	78.99	74.38	92.13	<b>82.28</b>
250	78.75	86.52	82.45	68.75	93.86	79.31	75.02	92.64	<b>82.89</b>
300	79.44	87.55	<b>83.30</b>	68.18	94.90	79.31	74.71	93.71	83.13
350	79.95	87.83	<b>83.70</b>	68.76	94.93	79.73	75.31	93.65	83.48
400	79.93	88.63	<b>84.05</b>	68.63	95.08	79.70	75.64	93.78	83.74
450	80.42	88.38	<b>84.20</b>	68.86	94.66	79.71	76.25	93.78	84.12
500	80.49	88.85	<b>84.46</b>	69.25	94.57	79.94	76.72	93.64	84.34

Table 3.3: Effect of varying size of training data set

Performance of the distantly supervised classifier shows a similar pattern. Increasing the number of seed abstracts results in a larger distantly labelled training data set which improves classification results. The distantly supervised classifier outperforms



---

the supervised one when there are fewer than 100 seed abstracts. The reason for this is the supervised classifier does not have access to a sufficient volume of training data while the distant supervision is able to generate more. As the number of seed abstracts increases the situation is reversed with the supervised classifier outperforming the distantly supervised one. When more than 100 abstracts are available the supervised classifier has the advantage of having access to sufficient accurately labelled examples to train a relation extraction system. The distantly supervised classifier still has access to more data but it is not as accurate.

The mixture model produces the best results of all approaches when between 5 and 250 abstracts are used. The mixture model tends to achieve higher precision but lower recall than the distantly supervised approach, possibly because the training data used by the mixture model is more accurate and contains fewer “false positive” examples. On the other hand, the precision and recall of the mixture model are often higher than the supervised model. The increase in recall is presumably caused by having access to additional training data and the precision scores suggest that the classifier is not harmed by some of these containing noisy labels.

The difference in performance between the supervised and the mixture-models gets smaller as the number of seed abstracts increases. Using 300 or more abstracts, the supervised classifier outperforms the mixture model. At this point the supervised classifier has access to sufficient amount of training data (more than 5000 manually labelled instances versus 28000 distantly labelled instances) and provides better results than using the larger set of distantly labelled data in addition. However the results for the mixture model are significantly better than the results of the supervised classifier in terms of F-Score to a seed abstract size of 150 (one-tailed paired t-test,  $p \leq 0.05$ ). Using 200 training abstracts the result of mixture model shows only a weak evidence to be better than the supervised model ( $p \leq 0.1$ ).

### **3.5 Conclusion**

This chapter explored of the effect using automatically labelled data to improve classification results for adverse-drug effects. The presented method showed that if only a small set of manually labelled training instances is available, a bootstrapping technique to generate a large set of distantly labelled data can improve the classification

---

results. However, the benefit of using distantly labelled data decreases as the amount of manually labelled data available increases.

## Chapter 4

# Detecting relations from the UMLS Metathesaurus in Medline abstracts

This chapter describes several important aspects of this thesis. First, the chapter describes the generation of distantly labelled data from Medline abstracts using related information of the UMLS Metathesaurus. The data will be used for all the remaining experiments in this thesis. Section 2.3 introduced Medline, UMLS and MetaMap which are used in this chapter. However, this chapter analyses UMLS in detail in order to select a subset of relations which are used to train classifiers.

Furthermore, this chapter also presents results using a set of UMLS Metathesaurus relations for distant supervision. Two vocabularies of UMLS are selected: *NDF-RT* (National Drug File - Reference Terminology) and *NCI* (National Cancer Institute Thesaurus). A classifier is trained and evaluated using held-out evaluation at entity-level. In addition an evaluation on the manually labelled gold standard is also conducted.

The chapter is structured as follows. First, a more detailed overview of the UMLS Metathesaurus is provided and relations selected for the experiments. Section 4.2 describes the generation of distantly labelled data from Medline abstracts using the selected relations. Then, in Section 4.4, filter techniques are presented to improve the quality of the data. The relation extraction system used for the following experiments is introduced in Section 4.5. Section 4.6 describes the distantly supervised classification experiments using the UMLS Metathesaurus in combination with held-out evaluation. First, an overview about the experimental setup is given. Then, the results of

---

the held-out experiment are presented. The end of Section 4.6 presents the results of the classifier using manually labelled data. The chapter finishes with a conclusion in Section 4.7.

## 4.1 Selection of UMLS Metathesaurus relations

This section provides more details about the relations in the UMLS Metathesaurus in order to identifying those which are potentially useful for distant supervision. UMLS contains a large number of relations from different vocabularies. Some relations contain hundreds of thousands of related instance pairs (CUI pairs) while other ones involve only a few hundred instances. In general, each vocabulary includes medical terms and relationships associated with a particular topic. For example, the FMA (Foundational Model of Anatomy) contains biomedical relationships associated with anatomical knowledge (e.g. *has\_muscle\_attachment*, *has\_nerve\_supply* and *has\_physical\_state*) and GO (Gene Ontology) contains information about gene products, cellular components and molecular functions (e.g. *occurs\_in*, *positively\_regulates* and *part\_of*).

Information about relations in the UMLS Metathesaurus is stored within MRREL (see Section 2.3.2). Table 4.1 presents the number of instances<sup>1</sup> of the 25 most frequent relations (according to RELA labels) within MRREL. Even if some relations and relation instances can be defined across several vocabularies, this table depicts the situation of frequencies of the related instances. The most frequent kind of relations are those defined between two CUIs without any further definition and without any RELA label, followed by the *isa* relation.

The largest set of relations, those ones without RELA label, do not seem to be useful for detecting relationships between entities in natural language, since it is not clear which relationship they describe. Another large set of related instance pairs defines taxonomical relations such as *isa* (e.g. SNOMEDCT contains 542,485 of those instance pairs) and sibling relations (e.g. GO contains 1,487,492 sibling instance pairs). This information might be very valuable for a different use case or even in combination with other relations. However, similar relations have been already explored in a range

---

<sup>1</sup>Note, the frequency describes how often a relation (instance pairs per relation) occurs within MRREL and not the number of distinct instance pairs for each relation. Multiple occurrences of instance pairs per relation can be possible.

---

<b>RELA</b>	<b>frequency</b>
<i>no RELA label</i>	29293384
isa	2170609
has_translation	1181911
has_ingredient	984373
classifies	658781
has_member	598785
sib_in_isa	553804
mapped_to	550486
has_expanded_form	413523
same_as	235628
has_finding_site	194754
has_component	190256
has_clinical_course	167613
has_episodicity	167558
severity_of	166759
method_of	160810
part_of	150052
has_dose_form	149039
sib_in_part_of	136240
has_associated_morphology	123075
has_tradename	118251
has_priority	106209
constitutes	105857
has_permuted_term	105695
sib_in_branch_of	96316

Table 4.1: Relation frequencies in the UMLS Metathesaurus

of different publications, such as Snow et al. [2004], Pantel and Pennacchiotti [2006], Wu and Weld [2007] or Nikolova and Angelova [2011]. Also other relations such as meronyms (e.g. *part-of*) or localisation relations also show a high frequency. Those relations have also been addressed in a different context e.g. Bossy et al. [2013] or Ling et al. [2013]. For this reason, relations such as *isa* or *part-of* are less interesting for this thesis. One goal of this thesis is to investigate, whether the UMLS Metathesaurus is a useful source for distant supervision. Relations which have not been used previously are more interesting than relations which have already been used. If relations have been successfully used in a different context, it is likely that they are also useful for the context of UMLS. So this problem does not need to be explored again.

UMLS vocabulary, <i>RELA-label</i>	example: related instance pair
<b>RXNORM</b> , <i>dose_form_of</i>	C0709583='Exelderm 1 % Topical Cream' C0991551='Topical Cream'
	C1245694='Cyclacillin Oral Suspension' C0991537='Oral Suspension'
<b>SNOMEDCT</b> , <i>active_ingredient_of</i>	C2585241='Oral form calcitriol' C0006674='Calcitriol'
	C2315140='Oral form cyanocobalamin' C0042845='Cyanocobalamin'
<b>GO</b> , <i>regulates</i>	C1817638='Schwann cell differentiation' C1817639='regulation of Schwann cell differentiation'
	C2247665='maintenance of sister chromatid cohesion' C2247671='regulation of maintenance of sister chromatid cohesion'
<b>UWDA</b> , <i>branch_of</i>	C0503881='Right anteromedial basal pulmonary artery' C0226062='Right anterior basal segmental artery'
	C0735513='Meningeal branch of right fourth sacral nerve' C0735548='Transverse branch of meningeal branch of right fourth sacral nerve'

Table 4.2: UMLS Metathesaurus relation examples

UMLS also defines nested relations (examples are given in Table 4.2, line 1-3) which do not appear useful nor interesting to focus on.

MRREL contains a large range of different vocabularies with different amounts of relation instances. Figure 4.3 shows the different vocabularies with the largest number of relation instances. The first column presents the SAB label, which is shown within MRREL, the second column shows the name of the source vocabulary and the last column the frequency within UMLS. SNOMEDCT contains more than 3 million instances within MRREL. However, not all vocabularies appear to be useful or interesting for this work. Some of the vocabularies containing the most frequent relations also contain redundant information. Vocabularies such as MDRFRE or MDRGER represent a language specific version (French and German) of MDR (Medical Dictionary for Regulatory Activities (MedDRA)). The defined relation instances are all identical

<b>SAB label</b>	<b>Name</b>	<b>Relation Instances</b>
SNOMEDCT	Systematized Nomenclature of Medicine (SNOMED) Clinical Terms US Edition	3075104
SCTSPA	SNOMED Clinical Terms, Spanish Language Edition	2907412
MSH	Medical Subject Headings	2853848
RXNORM	RXNORM Vocabulary	2448434
LNC	Logical Observation Identifier Names and Codes	2340736
GO	Gene Ontology	1808194
NCBI	NCBI (National Center for Biotechnology Information) Taxonomy	1626382
MDRJPN	MedDRA Japanese	1588956
MDRSPA	MedDRA Spanish	1577052
MDRPOR	MedDRA Portuguese	1577052
MDRITA	MedDRA Italian	1577052
MDRHUN	MedDRA Hungarian	1577052
MDRGER	MedDRA German	1577052
MDRFRE	MedDRA French	1577052
MDRDUT	MedDRA Dutch	1577052
MDRCZE	MedDRA Czech	1577052
MDR	MedDRA (Medical Dictionary for Regulatory Activities)	1394706
MTHSPL	FDA Structured Product Labels	1325370
ICD10PCS	International Classification of Diseases, 10th Revision, Procedure Coding System	1174716
MEDCIN	MEDCIN	1169784
UWDA	Digital Anatomist	1149564
FMA	Foundational Model of Anatomy	1104998
MTH	Metathesaurus Names	1047578
NCI	NCI (National Cancer Institute) Thesaurus	883070
ICD10CM	International Classification of Diseases, 10th Revision, Clinical Modification	788992
NDFRT	National Drug File - Reference Terminology	645942
MSHRUS	MeSH (Medical Subject Headings) Russian	642102
MSHSPA	MeSH Spanish	642072
MSHPOR	MeSH Portuguese	641046
MSHPOL	MeSH Polish	590948

Table 4.3: Top-30 vocabularies of the UMLS Metathesaurus based on relation instances

---

within the language specific vocabularies. Some other vocabularies consist of either mainly relations without RELA labels (such as MTH) or taxonomical relations (such as ICD10CM or UWDA).

In order to show that UMLS Metathesaurus is a useful source for distant supervision, relations from two vocabularies are selected for the experiments in the following chapters. First, some relations of NDFRT (National Drug File - Reference Terminology) are selected. NDFRT defines relations connected to drugs and diseases such as *may-treat* or *may-prevent* which are easy to understand for a biomedical non-expert. This simplifies the analysis of the instance pairs and distantly labelled annotations later on. An overview about different NDF-RT relations is given in Table 4.4. The table shows a range of potentially useful relations. On the other hand, some other relations appear to be less useful, such as *has\_print\_name* (relation with itself) or *has\_participant* (many instances are nested relations).

Additionally, relations from NCI are selected. NCI contains information related to genes and cancer. The advantage of this vocabulary is that it often contains a large number of instance pairs for each relation. Furthermore, many sentences containing NCI instance pairs can be found in Medline abstracts allowing distantly labelled data to be generated (as seen e.g. in Table 4.5). Other vocabularies such as SNOMEDCT, GO or FMA seem to contain also useful relations for distant supervision. Although, this thesis examines relations from two vocabularies of UMLS, the approach could be extended to other UMLS subsets with similar characteristics (large number of related information, large number of positively labelled sentences using known CUI pairs).

## 4.2 Data Generation

A set of relations from NDFRT and NCI are used to generate distantly labelled data from sentences in Medline abstract. Relations are selected according to factors such as number of instance pairs in the Metathesaurus, number of distantly labelled sentences found or promising results in preliminary experiments (the list of relations can be found in Table 4.5). The generation is divided into three steps and has similarities to the process described in Section 3.2:

- 1) **Generation of positive and negative instance pairs:** CUI pairs which occur multiple times within the same relation set are removed since this information is re-



<b>freq.</b>	<b>RELA</b>	<b>example instance</b>
48298	may_treat	C0010200='Coughing', C0976389='DIHYDROCODEINE BITARTRATE 16MG CAP'
36782	has_product_component	C0004261='Hyoscyamine Sulfate', C0974688='Belladonna Alkaloids 0.00672 MG/ML Oral Suspension'
36251	has_contraindicated_drug	C0304925='Human serum albumin preparation' C0018802='Congestive heart failure'
25823	has_physiologic_effect	C1371672='Decreased Immunologically Active Molecule Activity', C0057992='diflorasone diacetate'
23630	has_ingredient	C0065864='medroxyprogesterone acetate', C1648910='medroxyprogesterone acetate 400 MG/ML Injectable Suspension'
15007	has_mechanism_of_action	C2757023='Norepinephrine Uptake Inhibitors', C0011812='Dextroamphetamine'
14610	has_participant	C0043031='Warfarin', C2926987='ROMIDEPSIN/WARFARIN'
12387	has_dose_form	C1154181='Inhalant Solution', C0980200='Sodium Chloride 0.45% Inhalation So- lution'
12337	has_print_name	C0975677='Ciprofloxacin 500 MG Oral Tablet' C0975677='Ciprofloxacin 500 MG Oral Tablet'
6048	may_prevent	C0030193='Pain', C0282232='Levomethadyl Acetate Hydrochloride'
2228	has_contraindicating_class	C0033858='Psoralens', C0025684='Methoxsalen'
967	may_diagnose	C0242350='Erectile dysfunction', C0979270='Papaverine 150 MG Oral Capsule'
762	has_pharmacokinetics	C1373187='Renal Excretion', C1533036='Ipratropium bromide 17mcg HFA in- haler'
714	induces	C0026961='mydriasis', C0977488='homatropine ophthalmic 5% oph- thalmic solution'
674	has_contraindicating_mechanism_of_action	C1373090='Cholinergic Nicotinic Antagonists', C0001714='Polymyxin B Sulfate'

Table 4.4: NDF-RT relations

---

dundant. Next, self-relations (CUIs which are related to themselves) are removed from the list of positive candidates (see Example 5 in Table 2.3). The same technique is also applied by Bobic et al. [2012], who argue that self-relations are likely to produce false positives. Instance pairs which occur within different UMLS Metathesaurus relations are also removed from the list of positive candidates (also used by e.g. Tymoshenko et al. [2012]). Those entity pairs can express different relations so it is likely that they will produce false positives.

Unfortunately, the majority of UMLS Metathesaurus relations do not contain negative entity pairs for a target relation. Thus, negative data for each target relation is generated using the closed world assumption (see Section 2.1.2). Given all entity pairs for each relation, negative entity pairs are generated by producing new combinations between all entities (without changing the order of the entities in the pairs). All of those entity pairs which are not known to be related within UMLS are then considered as a negative instance pair of that relation. This step follows the closed world assumption.

2) **Preparing textual data:** Distantly labelled data is generated from Medline abstracts. A corpus of 1,000,000 Medline abstracts is split into sentences and annotated with UMLS-CUIs using MetaMap.

3) **Selecting sentences containing positive and negative seed instances:** Finally, UMLS-annotated Medline sentences are extracted to generate distantly labelled data. Sentences containing positive seed instances of a target relation are labelled as positive instance of the target relation and sentences containing negative seed instances are labelled as negative examples.

### 4.3 Corpus Statistics

Table 4.5 shows the results of the automatic generation process. The upper part of the table shows the NDF-RT relations and the lower part the NCI relations. The table shows the number of CUI pairs for the relations that are useful to generate distantly labelled data. The middle part “*instance pairs*” presents the generation process of the seed instances. The Column *#given* shows the number of positive instances originally defined within the UMLS Metathesaurus. The Column *#pos* shows the number of positive instances remaining after the first filtering process (reduction of doubles, self-relation, occurrences in multiple relations, see Section 4.2). The Column *#neg* presents

MRREL relation	instance pairs		CUI pairs in baseline data			
	#given	#pos	#neg (g.)	#pos	#u-pos	#u-neg
may_treat	48,298	35,283	8,826,791	54,857	2,072	339,408
mechanism_of_action_of	15,007	12,434	3,266,398	845	138	7,793
may_prevent	6,048	2,347	722,593	6,476	156	55,516
contraindicating_class_of	2,228	1,777	91,160	557	82	15,081
biological_process_involves_gene_product	10,267	10,178	1,244,340	54,681	2,502	349,911
disease_has_normal_cell_origin	13,494	13,348	1,139,450	7,750	736	70,005
gene_product_has_associated_anatomy	1,488	1,481	144,021	13,696	549	149,842
gene_product_has_bio_chemical_function	5,266	3,977	518,168	22,674	1,046	141,715
process_involves_gene	19,258	19,079	2,260,050	19,291	1,792	95,554

Table 4.5: Amount of CUI pairs for the generation of the baseline data (all); #given: amount of CUI pairs in MRREL, #pos: amount of positive CUI pairs, #neg (g): amount of negative CUI pairs generated from the positive pairs, #u-pos: unique positive pairs, #u-neg: unique negative pairs

---

the amount of negative seed instances generated from *#pos* using the closed world assumption (see Section 4.2). The columns on the right hand side of the table “*CUI pairs in baseline data*” present the statistics for the distantly labelled data generated from the Medline abstracts. Column *#pos* defines the number of sentences of each target relation that are found. *#u-pos* presents the number of different (unique) CUI pairs found within *#pos*.

The table shows that only a small number of CUI pairs can be found within the Medline abstracts. Likewise, Column *#neg* and Column *#u-neg* show the number of sentences containing negative seed instances and the number of different unique negative seed instances found.

## 4.4 Further filtering steps

Distantly labelled data often contains false positive and false negative examples for different reasons. In order to remove some false labels, various filtering steps (which have been used by previous approaches) are applied:

1) **Reduction of sentences containing both positive and negative instances:** In biomedical literature medical concepts are often mentioned within a list (e.g. *a,b,c, and d*), such as seen in the two sentences in Figure 4.1 (“*generalized and focal epilepsies, including special epileptic*” and “*myocardial infarction and stroke*”). In both examples one of the entities within the list is an instance of a positive relation (*may-treat*(“*antiepileptic drugs*”, “*epileptic*”), *may-treat*(“*aspirin*”, “*myocardial infarction*”)), while the other entities are either detected incorrectly by MetaMap (entity should be “*generalized epilepsies*” rather than just “*generalized*”) or entities describe an unknown (and therefore negative) relation (“*focal epilepsies*” and “*stroke*”) according to UMLS. To reduce the risk of annotating false negatives, sentences containing both positive and negative relations are removed from the set of candidates. Other authors address the issue of entity lists, including Liu et al. [2014b] and Bing et al. [2015].

2) **Reducing word distance:** Distantly labelled sentences may contain a large number of words between the candidate entities. The assumption is that the more words that occur between two target entities the less likely they express the relation of interest. For this reason candidate sentences are deleted if they contain more than five words between candidate entities. A setting of five words provided the best results in

---

Valproic acid (VPA) is considered to be a drug of first choice and one of the most frequently-prescribed [C0003299:antiepileptic drugs] worldwide for the therapy of [C0205246:generalized] and [C0014547:focal epilepsies], including [C0205555:special] [C0014544:epileptic]. (PMID=18201150)

BACKGROUND AND PURPOSE : [C0004057:Aspirin] reduces the risk of [C0027051:myocardial infarction] and [C0038454:stroke] by inhibiting thromboxane production in platelets. (PMID=19466986)

Figure 4.1: Group of CUIs might cause false negatives

preliminary experiments. However, an optimal word distance might differ for each relation. Filtering according to a restricted number of words between entities is used by a range of different approaches, including Takamatsu et al. [2012], Abad and Moschitti [2014], Poon et al. [2015] and Zheng and Blake [2015].

3) **Remove entity pairs with low frequency:** Entity pairs with a low frequency are removed from the candidate list. An entity pair which occurs only a few times within one million documents might not be useful. In particular in context of positive examples, it is less likely that entity pairs occurring only within a few sentences contain at least one (true) positive example than entity pairs with a higher frequency. For this reason entity pairs (positives and negatives) which occur fewer than five times in a sentence are deleted.

Activation of hypothalamic-pituitary-adrenal (HPA) axis [C0018790:inhibits] development of [C0026549:morphine] tolerance. (PMID=18053645)

Figure 4.2: Using non-nouns for distant supervision might cause errors

4) **Restricting entities to noun phrases:** All CUIs that refer to non-nouns are removed from the set of candidates to accommodate for errors in MetaMap's annotations. Restriction to noun phrases has been also used by authors such as Augenstein

---

et al. [2015]. The example in Figure 4.2 describes the (negative) relation *NOT-may-treat*(“*cardiac arrest*”, “*morphine*”). The concept C0018790=“*cardiac arrest*” was mapped to the verb “*inhibits*” by MetaMap.

## 4.5 MultiR - Relation Extraction

In Chapter 3 (Bootstrapping Limited Training Data) a SVM-based (JSRE) relation extraction system was used to detect biomedical relations. In previous context JSRE was a good choice for the following reasons: The classifier has been used by other authors using the same data set (ADE). The system is easy to setup, experiments can be carried out quickly and results are relatively efficiently. Furthermore, the presented technique had the objective to use DL data support and improve a supervised classifier. However, SVM-based systems tend to be a popular choice for supervised learning.

In contrast to Chapter 3, multi-instance learning will be used for this and the following chapters, rather than a classical supervised learning method. In context of distantly supervised relation extraction, a range of systems transform distant supervision with noisy input data to traditional supervised learning by using a single-instance single-label technique<sup>1</sup> [Surdeanu et al., 2011], such as in the work of Bellare and McCallum [2007], Mintz et al. [2009] or Nguyen and Moschitti [2011a]. In the more recent years however, multi-instance learning became much more popular to train relational classifiers with automatically labelled data, such as in Surdeanu et al. [2012], Ritter et al. [2013] or Liu et al. [2014b]. Multi-instance learning bases on the idea of using bags of labelled instances/sentences as input, rather than single instances, in order to better deal with noisy input data. A bag labelled as positive contains at least one positive example. Conversely a bag labelled as negative contains no positive instance at all (see Section 2.1.2). The relation extraction system used in the next chapters is exchanged for the following reasons:

This work focusses on exploring a new domain in context of distantly supervised relation extraction. Introducing a new system (which has not been used for distantly labelled data) would raise the question whether an already established method would do better on this new domain. In this case a direct comparison to existing methods of

---

<sup>1</sup>as usually used for supervised learning; each single (noisy) example is used as input

---

that area is required. However, this work does not focus on introducing new machine learning methods for distantly supervised relation extract. Instead it focuses on detecting biomedical relation using distant supervision, in particular for relations of UMLS. For this reason, this and the following chapters will rely on MultiR [Hoffmann et al., 2011], a well established and frequently used method for distantly supervised relation extraction. In comparison to the previous chapter, the goal was to examine whether it is possible to support a supervised classifier with distantly labelled data if the number of manually labelled training instances is small. In the following chapters instead, no manually labelled data is available for training.

Moreover, Ray and Craven [2005] compare different multi-instance learning methods with supervised learning algorithms on a range of different domains. First of all, the authors show, that ordinary supervised methods are doing well on different tasks in comparison to multi-instance learning. However, the authors also show that multi-instance approaches often outperform their supervised counterparts. This might be a reason why multi-instance learning became popular in recent years to train a classifier with noisy data. Moreover, also in context of distantly supervised relation extraction, Surdeanu et al. [2011] for instance show, that their multi-instance method outperform an approach trained on single instances, such Mintz et al. [2009].

MultiR is a probabilistic, graphical model which bases on multi-instance learning. The method takes positive and negative ‘bags’ (sets) of instances as input rather than single instances. A bag labelled as positive contains at least one positive example and a bag labelled as negative contains no positive instance at all. For this use case it means that all sentences containing a particular entity pair are considered as one bag (labelled according to the label of the entity pair in UMLS). Assuming that an entity pair is known as an instance of a relation, the approach expects that at least one of the sentences containing the entity pair is a true positive (see also Section 2.1.2). MultiR is a frequently used state-of-the-art approach for distantly supervised relation extraction (see Section 2.1.2). For this reason it appears to be more appropriate to explore the usage of UMLS in context of distantly supervised relation extraction. Software to implement the system is freely available<sup>1</sup>.

MultiR features described by Surdeanu et al. [2011] are used for all further experi-

---

<sup>1</sup>MultiR can be downloaded here: <http://aiweb.cs.washington.edu/ai/raphaelh/mr/index.html>.

---

Argument Features	<ul style="list-style-type: none"> <li>- Head words (CUIs) of related entities, their combination and order</li> <li>- Entity mention and words around the entity</li> <li>- Semantic type of entity mentions and their combination</li> </ul>
Syntactic Features	<ul style="list-style-type: none"> <li>- Sequence of labels in the dependency path connecting the two entities</li> <li>- Stemmed words in the dependency path</li> </ul>
Surface Features	<ul style="list-style-type: none"> <li>- Sequence of words and POS tags between entities</li> <li>- Distance between related entities</li> <li>- Counts of semantic types in sentence</li> </ul>

Table 4.6: Selection of features used for distant supervision

ments. These features are adjusted to the biomedical context by using Semantic Types. Features can be divided into three groups: argument features, syntactic features and surface features. An overview of the different features is listed in Table 4.6. Sentences in the training and test data are processed with the Porter Stemmer [Porter, 1997] and the Charniak-Johnson Parser [Charniak and Johnson, 2005]. In addition the Stanford Parser [Klein and Manning, 2003] is used to generate dependency tree features.

## 4.6 Experiment

In this section a distantly supervised classifier for relations of the UMLS Metathesaurus is presented. The presented classifier is trained on relations of the UMLS vocabularies NDF-RT and NCI and will be evaluated using two different techniques. First, the evaluation is carried out using held-out evaluation, followed by an evaluation against a small manually labelled gold standard containing annotations of two different relations. Following the example of most other distantly supervised relation extraction systems, the experiment will be conducted at entity level (against the KB). In context of using automatically labelled evaluation data, entity level reduces the risk of false labels.

The objective of the following experiments is to show that the UMLS Metathesaurus is a useful source for distant supervision. Even though relations of only two subsets are used for the following experiment, it can be expected that other relations



---

with similar characteristics (such as similar frequencies in UMLS or similar number of distantly labelled sentences) provide similar results. NDF-RT has been used for distant supervision before by Nikolova and Angelova [2011]. However, the distantly supervised classifier was only considered as a lower baseline for a different method. Furthermore, the DS system was implemented with a SVM and used with an unnatural bias of 1:1 (positives/negatives). Usually the number of negative instances in distantly labelled data is much larger, as seen in Table 4.5, which makes the classification task more challenging. The relations of NCI have not previously been used for distant supervision.

### 4.6.1 Setup

In the following experiment a classifier is trained and evaluated for a range of different UMLS relations. In order to provide reliable results, the experiment is carried out using a 4-fold held-out cross validation. Results are then averaged across the four steps and compared to a naive baseline. This baseline is a simple technique which predicts each instance as positive. Hence, the baseline achieves a low precision and a perfect recall. For the experiments the multi-instance learning method MultiR is used (see Section 4.5).

In the following the generation and the usage of the data in the experiment is described in more detail:

#### 4-fold held-out cross validation

Evaluation data of an  $n$ -fold cross validation is usually generated by splitting data into  $n$  equally sized subsets (folds). Then, each of the single folds is selected once as test set and using all the remaining sets for training. For this experiment however, a random split into four equally sized sets would be not appropriate, since the evaluation is carried out at entity level (and evaluated against the knowledge base). Using a random split, it is very likely that entities of the test set might also occur within the training data and would be known in advance. This can make the evaluation less reliable and possibly less efficient. For this reason the split into different evaluation sets is carried out by assigning the different CUI pairs into different folds to ensure a held-out evaluation setup (see Section 2.2). In this way, each fold contains a different set of positive

---

and negative CUI pairs. However, this n-fold split based on entity occurrences makes it difficult to generate equally sized folds (some CUI pairs occur more often in distantly labelled sentences than other ones). In order to generate folds with a similar size, the data for the cross validation is generated in the following way:

Each positive entity pair occurring within the distantly labelled data of a relation is sorted by its frequency. Then, according to this frequency (and starting with the most frequent pair), entity pairs are assigned sequentially to the different folds. The most frequent pair is assigned to the first fold, the second most frequent one is assigned to the second fold and so on. Next, all sentences containing a CUI pair which is assigned to a particular fold are then sorted into that fold. This step is repeated for positive and negative CUI pairs for all relations.

### **Bias adjustment**

As illustrated in Table 4.5, the distantly labelled data of each relation contains a larger number of negative sentences than positive ones. In context of relation extraction and other machine learning tasks, highly unbalanced training data is a known issue (e.g. Chawla et al. [2004]; Swampillai and Stevenson [2011]) and might cause problems to train a classifier. Using a much larger number of negative training instances might result in a restrictive model (high precision, low recall), whereas a larger number of positive training examples, might result in a less precise model (high recall, low precision). In context of using distantly labelled data, it can happen that negative examples outperform positive examples by more than 10 times (see e.g. *contraindicating\_class\_of* in Table 4.5).

In addition, using a large number of training instances can significantly increase the runtime of a classifier and might result in memory errors. Hence, if the training data is very large and unbalanced (towards negative instances), it appears to be reasonable to reduce the number of negatives.

Adjusting the ratio of positive to negative training instances can have a influence on the classification results, as seen in Thomas et al. [2011]. Authors such as Surdeanu et al. [2012] or Riedel et al. [2010] for instance, randomly reduce the number of negative instances in their data (under-sampling). In order to deal with the large number of negative instances in this experiment, negative instances are randomly reduced as

well. For the following experiment a bias of 1:2 (positives:negatives) has been chosen. This setup provided the most promising results. However, a further examination how to deal with imbalanced data in the best way, is not carried out and its exploration is not in the focus of this thesis.

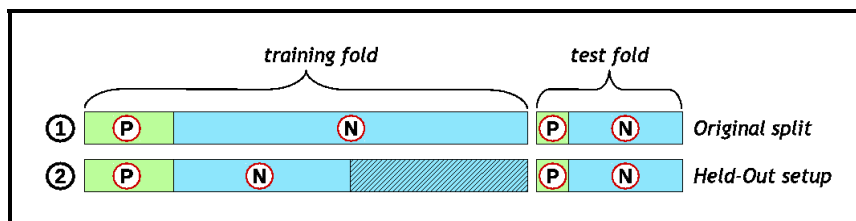


Figure 4.3: Adjusting bias of training data (Held-Out Setup)

Figure 4.3 visualises the bias adjustment (P:positives; N:negatives). Line 1 shows a given training and test fold of a cross validation step including the ratio of positive and negative instances. Line 2 represents the setup of the following experiment. Using the data from line 1, the number of negative instances is randomly reduced to a bias of 1:2 (the grey area represents the removed instances). However, instances from the test fold are not removed.

## 4.6.2 Held-Out Evaluation

Table 4.7 presents the held-out classification results in comparison to the naive baseline. In the upper part of the table the overall classification result (micro-avg. f1-score) is presented, followed by results for NDF-RT. The lower part of the table presents results for NCI relations. The naive baseline is located in the rightmost part of the table.

The results show that the classifier is able to extract related instances of the UMLS Metathesaurus with a higher f1-score than the baseline. In particular the relations from NCI are detected well. All NCI relations provide an f1-score above 40 and the relation *gene\_product\_has\_biochemical\_function* results in a f1-score above 60. The relations of NDF-RT seem to be more difficult to detect, in particular a relation such as *contraindicating\_class\_of*. The classification result is close to the naive baseline. These low results might be connected to the fact that related entity pairs of NDF-RT occur more often in a context outside the relation. Likewise, *contraindicating\_class\_of* con-

MRREL relation	Results			Naive Baseline		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Overall (micro-avg.)	44.72	54.23	<b>49.02</b>	15.41	100.0	26.71
	NDF-RT relations					
contraindicating_class_of	16.84	32.21	<b>21.96</b>	8.76	100.0	16.10
may_prevent	29.17	26.47	<b>27.57</b>	7.79	100.0	14.45
may_treat	51.47	42.68	<b>46.29</b>	20.45	100.0	33.95
mechanism_of_action_of	48.88	65.63	<b>55.56</b>	21.05	100.0	34.78
	NCI relations					
biological_process_involves_gene_product	49.71	46.97	<b>48.17</b>	18.36	100.0	31.02
disease_has_normal_cell_origin	34.32	80.41	<b>48.04</b>	13.54	100.0	23.85
gene_product_has_associated_anatomy	35.67	61.41	<b>45.10</b>	8.91	100.0	16.36
gene_product_has_biochemical_function	60.07	73.06	<b>65.81</b>	21.67	100.0	35.63
process_involves_gene	54.03	52.68	<b>52.83</b>	21.61	100.0	35.55

Table 4.7: Best results using held-out

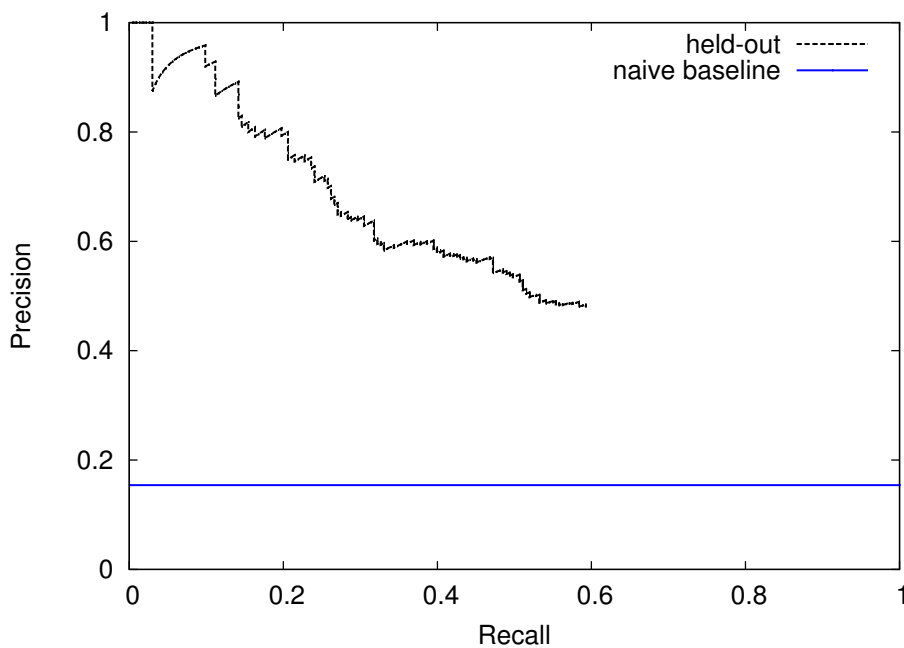


Figure 4.4: Precision/Recall Curve for Held-out data

tains a much larger amount of negative instances compared to positive ones (see Table 4.5). Using a strongly unbalanced test set often leads to lower results. Furthermore,

---

relations with a smaller training/test set (such as in case of *contraindicating\_class\_of* or *may\_prevent*) tend to achieve lower results than relations with a larger number of instances (e.g. *may\_treat* or *biological\_process\_involves\_gene\_product*).

The overall classification result takes all predicted relations into account, including their evaluation data size. Those results are very encouraging. The distantly supervised classifier of the overall result achieves an improvement of f1-score of more than 80% compared to the naive classifier.

Figure 4.4 presents a precision/recall curve computed using MultiR’s output probabilities for all different relations. The graph highlights that instances with a high output probability tend to have a high precision. With the increasing amount of positive predictions the precision decreases.

### 4.6.3 Manual Evaluation

In order to support the held-out classification results, a distantly supervised classifier is additionally evaluated on a manually labelled data set. The evaluation set consists of 800 sentences containing annotated *may-treat* and *may-prevent* relations of the UMLS Metathesaurus. Each relation consists of 400 sentences each, including a varying number of positive and negative entity pairs. A detailed overview about the data set is provided in the Section 5.1 (next chapter).

For the experiment the same setup (bias 1:2 for training) as in the previous held-out experiment is used. However, as already mentioned in Section 4.6.1 (‘Setup’), results are assumed to be more reliable if entity pairs in the evaluation data are not already known from the training data. To ensure a held-out setup, training data is generated in a different way as in the previous experiment. Instead of assigning CUI pairs into different folds, sentences containing CUI pairs from the manually labelled test data are removed from the distantly labelled data. Then, the bias of this data is adjusted to 1:2 (positives:negatives) by randomly reducing the number of negative sentences. The remaining instances are used to train a single classifier for the manual evaluation experiment. The classification results are then compared to a naive classification baseline.

The results of the manual evaluation are presented in Table 4.8. In case of *may-prevent* the naive baseline achieves an f1-score of 21.50 with a precision of 12.05.

---

*May-treat* contains a larger number of different positive entity pairs, thus, the naive classification results in a precision of 28.11 and a f1-score of 43.88. In comparison with the *may-prevent* baseline (and the baseline of the held-out experiment) the naive f1-score of *may-treat* is relatively high.

relation	DS Classification			Naive Classification		
	Prec.	Rec.	F1	Prec.	Rec.	F1
may_prevent	58.06	30.00	<b>39.56</b>	12.05	100.00	21.50
may_treat	45.21	47.14	<b>46.15</b>	28.11	100.00	43.88
<b>Overall</b> (micro-avg.)	47.46	42.00	<b>44.56</b>	20.41	100.00	33.90

Table 4.8: Evaluation of distantly labelled classifier using manually labelled data

However, the results of the distantly supervised classifier outperform the naive baseline in all cases. In case of *may-prevent* precision increases by more than four times the precision of the naive classifier. The f1-score of distantly supervised classifier nearly doubles, even though the recall is only 30. The DS result for the *may-treat* relation also outperforms the baseline. The precision increases by approximately 60%. However, due to the fact the *may-treat* baseline is already very strong, the f1-score of 46.15 only represents an improvement of less than 3 points.

The overall classification result (micro average) takes into account that *may-treat* contains more instances. Also in this case the distantly supervised classifier outperforms the naive baseline by more than 10 points compared to the naive baseline.

## 4.7 Summary

This chapter presented a distantly supervised classifier to detect relations from two different UMLS Metathesaurus vocabularies. The system has been evaluated in two different ways. A first experiment evaluated the system using a held-out portion of the data within a 4-fold cross-validation. A second experiment used a small gold standard as additional evaluation measure.

The system has been trained with a state-of-the-art relation extraction system MultiR, evaluated on entity level and compared with a simple baseline method. However, the results are encouraging and show that parts of UMLS are a useful knowledge source from which to generate distantly labelled data for a relational classifier.

## Chapter 5

# Comparison of Evaluation Strategies for Distantly Supervised Relation Extraction

Distant supervision is an effective and commonly used technique when manually labelled data is not available. However, without manually labelled data it is difficult to carry out an evaluation to estimate the quality of the classifier. Held-out evaluation is an alternative method which is able estimate the quality of a distantly supervised classifier and does not require any manually labelled data. The approach splits facts from a knowledge base into two parts. One part is used to generate distantly supervised training data and the other one to generate distantly supervised evaluation data [Hoffmann et al., 2010; Mintz et al., 2009; Riedel et al., 2010] (see Chapter 2.2). Held-out evaluation using distantly labelled data is a simple and quick technique for estimating the accuracy of distantly supervised relation extraction systems. Notwithstanding, this evaluation data is noisy and it is unclear what effect this has on the accuracy of performance estimates. The objective of this chapter is to investigate the question how useful held-out evaluation is without any additional manually labelled evaluation set and without any prior knowledge about the target domain.

The issue is explored in this chapter by evaluating relation extraction systems for two biomedical relations using both distantly and manually labelled data. Held-out data is automatically generated and then manually annotated by medical experts to al-

---

low direct comparison. Then, a distantly supervised classifier is trained and evaluated on both data sets. This chapter shows that a large portion of the labels generated by distant supervision for the two relations are incorrect. However, this chapter makes two contributions in context of held-out evaluation. Firstly, the chapter finds that evaluating classifiers using distantly supervised held-out data tends to overestimate performance compared to manually labelled data. Secondly, the experiments show that improvements in performance observed in evaluation against distantly supervised data are not necessarily reflected in improved results when measured against manually labelled data. This is the first direct comparison of evaluating distantly supervised classifiers against distantly and manually labelled gold standards. Analysis in previous work has been restricted to determining the true labels for a set of positively predicted labels (such as in Takamatsu et al. [2012]; Zhang et al. [2013]).

The remainder of this chapter is structured as follows. Section 5.1 describes the creation of the distantly supervised data and a manually labelled subset. A comparison of the automatically and manually generated labels is carried out in Section 5.2. A relation extraction system is introduced in Section 4.5. Section 5.3 evaluates a relation extraction system using distantly labelled and manually labelled data sets and compares the performance obtained. The chapter ends with a conclusion in Section 5.4.

## 5.1 Data Generation

This section describes the generation of the (gold standard) data used for the experiments. The following Section 5.1.1 describes how distantly labelled data is generated (for more detailed information see previous Chapter 4). Then, in Section 5.1.2 a small portion is extracted as held-out test data and manually re-annotated.

### 5.1.1 Distant labelling

For this experiment two biomedical relations (*may-treat* and *may-prevent*) are selected, which are easy to understand even for biomedical non-experts. These relations describe connections between a pharmacological substance (e.g. drug) and a disease (or symptom) and are taken from the UMLS vocabulary NDF-RT (National Drug File -



---

Reference Terminology). For example, the following sentence expresses a *may-prevent* relationship between the entities *fluoride* and *dental caries*:

*“Although **fluoride** is clearly a major reason for the decline in the prevalence of **dental caries**, there are no studies of the incremental benefit of in-office fluoride treatments for low-risk patients exposed to fluoridated water and using fluoridated toothpaste.” (PMID=10698247)*

Distantly labelled data for the two relations was generated from biomedical abstracts from Medline annotated with UMLS concepts by MetaMap. The detailed overview about the generation process is already presented in the previous Chapter 4.

### 5.1.2 Manually labelled Test Data

A set of 400 distantly labelled sentences were randomly selected for each relation to generate held-out test data. Although the distantly labelled data contains more negatively labelled sentences than positive ones, equal numbers of positive and negative examples (200 of each) are selected to ensure that a sufficient number of positive instances are included in the data set. This data set is referred to as **DL** (Distantly Labelled).

The DL data set was then manually annotated. Two annotators were recruited, both of whom were studying graduate degrees in subjects related to medicine at the University of Sheffield. Both annotators were instructed to follow an annotation guideline to carry out the labelling task. The guideline can be found in the Appendix 2. Both annotators did not have any experience with annotating medical text nor natural language processing.

The annotation task was carried out as follows: Given a sentence with a highlighted pharmacological substance and a highlighted disease, the annotators had to determine whether a sentence expresses the relationship of interest between two presented entities or not. The annotators were not shown the labels generated by the distant supervision process. The annotators were asked to only label sentences as positive if it contains a clear indication that the pharmacological substance either treats or prevents the disease. For example, the following sentence mentions that a study has been carried out to determine whether the drug *voriconazole* treats *paracoccidioidomycosis*:

---

*“A pilot study was conducted to investigate the efficacy, safety, and tolerability of [DRUG:voriconazole] for the long-term treatment of acute or chronic [DISEASE:paracoccidioidomycosis], with itraconazole as the control treatment.” (PMID=17990229)*

However, the sentence does not contain any indication that the drug successfully treats the disease and should therefore be annotated as a negative example of the relation.

### **Annotation Process**

The annotation task can be divided into four stages. Within the first stage each annotator was asked to label all 400 sentences for each relation. The first stage resulted in 325 agreements for *may-prevent* and 324 agreements for *may-treat*.

Next, each annotator was assigned to one relation and re-examined the disagreements of that relation. Each annotator had the opportunity to either change the previously given label or to write an explanation why the given label should remain different to the label of the other annotator. The second stage resulted in 364 agreements for *may-prevent* and 366 agreements for *may-treat*.

Within the third annotation stage each annotator again re-examined the remaining disagreements for the other relation. Again, the annotator had to decide whether the previous label should be changed or not. Unlike the previous stage, the annotator was able to take the comment of the other annotator into consideration. The third stage resulted in 390 agreements for *may-prevent* and 383 agreements for *may-treat*. Inter-annotator agreement [Cohen, 1960] after this stage was of  $\kappa = 0.94$  for *may-prevent* and  $\kappa = 0.91$  for *may-treat*. Each annotator spend at least 10 hours on the three annotation stages.

Finally, the remaining disagreements were resolved (stage four) by the author based on comments provided by both annotators and the annotation guidelines. The manually annotated version of the data set is referred to as **ML** (Manually Labelled).<sup>1</sup>

---

<sup>1</sup>The annotated corpus is available here: <https://sites.google.com/site/umlscorpus/home>.

---

## 5.2 Label Comparison

This section presents the differences between distantly labelled and manually labelled annotation. The experiments in this chapter (Section 5.3) are carried out at sentence and at entity level (see Section 2.2). For this reason annotation differences at both levels are considered. First, the labels for each single sentence are examined (Section 5.2.1). The analysis presents the amount of falsely labelled sentences in the DL data, including the number of false positives and false negatives. Then the differences between DL and ML at entity level are examined (Section 5.2.2).

### 5.2.1 Sentence Level Labels

Table 5.1 shows differences in annotations generated by the two labelling techniques evaluated at sentence level. The DL data set contains for both relations 200 positive and 200 negative sentences. Conversely, the ML data set for *may-treat* contains 173 positive and 227 negative examples, whereas the ML data set for *may-prevent* contains 139 positives and 261 negatives examples. A comparison of the DL and ML data sets shows that 40.25% of the labels changed for *may-treat* and 39.75% for *may-prevent*. The distant supervision process generated a larger number of false positives (*may-treat*=94, *may-prevent*=115) than false negatives (*may-treat*=67, *may-prevent*=54) for both relations. First of all, this shows that (as expected) distantly labelled data is noisy. Not every sentence containing a positive entity pair expresses the target relation. Moreover, instances known to be negatives, turn out to be positive (in a given sentence) less frequently.

		distantly labelled (DL)					
		may-treat			may-prevent		
		pos	neg	#	pos	neg	#
manually labelled (ML)	pos	<b>106</b>	67	173	<b>85</b>	54	139
	neg	94	<b>133</b>	227	115	<b>146</b>	261
		200	200		200	200	

Table 5.1: Comparison of manual and distantly labelled annotations

Assuming that a classifier is able to identify the *may-treat* and *may-prevent* relations with perfect accuracy then performance on the ML data sets would be preci-

---

sion=1.0, recall=1.0 and f1-score=1.0. However, the false labels on the DL data sets would lead to performance of the same classifiers being estimated as precision=0.61, recall=0.53 and f1-score=0.57 for *may-treat* and precision=0.61, recall=0.43 and f1-score=0.50 for *may-prevent*. Hence, the two data sets may provide quite different estimates of system performance. It also means that an improvement in terms of f1-score using distantly labelled data does not necessarily provide improvements on the ‘real’ annotation labels. This issue will be explored in detail in the next section.

## 5.2.2 Entity Level Labels

The analysis in Section 5.2.1 (Sentence Level Labels) presented annotation differences between distantly labelled and manually labelled sentences. The following analysis instead, presents annotation differences for the different entity pair (entity level labels). Each relation in the ML data set contains 200 positively and 200 negatively labelled sentences. Those sentence are labelled as positive, if the given entity pair is known as positive instance within UMLS, and labelled as negative according to the closed world assumption (not known to be related according to UMLS). However, an entity pair might occur several times within different sentences. While the previous analysis examined the correct labelling of each single sentence, this analysis examines the correct labelling of entity pairs. This means, that entity pairs which are considered to be negative according to the closed world assumption might be (manually) labelled at least once as positive within one of the sentences. In this case the original negative label has to be reconsidered. In context of entity level labels this means, that the negative DL label of an entity pair will be re-labelled as positive, if it occurs within at least once positive sentence (in the manually labelled data). Conversely it can happen that a positive entity pair which has been extracted from UMLS (DL), never occurs within a positive sentence. Although the number of selected sentences is relatively small, we cannot prove that the entity pair will ever occur within a positive sentence. For this reason, positive entity pairs (DL) which do not occur within at least one positive sentence will be re-labelled as negative in the ML set. The situation of different positive and negative entity pairs according to UMLS (DL) and according to their occurrence in the sentences (ML) is presented in the following.

Table 5.2 presents the number of different positive and negative entity pairs within

the set of sentences. Within the distantly labelled data, *may-treat* contains 124 positives and 186 negatives and *may-prevent* 41 positives and 147 negatives. The number of positives and negatives for both relations is lower compared to sentence level labels. This can be explained by the fact, that entity pairs can occur within multiple sentences. Furthermore the table shows that the data of *may-treat* contains a larger number of different CUI pairs than *may-prevent*. This might be connected to the fact that *may-treat* contains a larger number of CUI pairs within UMLS and generates a larger number of distantly labelled sentences (as seen in Table 4.5 in previous Chapter 4.3). Thus, a random selection of sentences from a larger data set with a larger number of different CUI pairs leads to a larger number of positive and negative instance pairs in the ML data set.

		distantly labelled (DL)					
		may-treat			may-prevent		
		pos	neg	#	pos	neg	#
manually labelled (ML)	pos	<b>76</b>	64	140	<b>22</b>	38	60
	neg	48	<b>122</b>	170	19	<b>109</b>	128
		124	186		41	147	

Table 5.2: Comparison of manual and distantly labelled annotations at entity level

Table 5.2 shows the difference between DL and ML entity level labels. For instance, sentences of *may-treat* contain 124 different positive entity pairs according to UMLS and 186 negative entity pairs according to the closed world assumption. These entities are labelled as DL. Entity pairs are then examined and re-labelled according to the fact, whether the entity pair occurs at least once within a (manually labelled) positive sentence. An analysis of the different sentences reveals, that 64 (of 186) different negative entity pairs occur at least once within a positive sentence. Those entity pairs are then manually re-labelled as positive. Furthermore, 48 (of the 124) positive entity pairs extracted from UMLS never occur within at least one positive sentence are then manually re-labelled as negative entity pairs.

In contrast to the comparison at sentence level, the data contains a larger number of false negatives than false positives. As seen in Table 5.2 some CUI pairs which are known to be correct never occur within a sentence expressing the relation of interest (*may-treat*=48, *may-prevent*=19). Conversely, a larger number of negative (unknown)

---

CUI pairs are re-labelled as positives (*may-treat*=64, *may-prevent*=38). Using a larger number of sentences for each positive CUI pair would probably further decrease the false positives. The number of false negatives shows that CUI pairs which are not known to be related may also occur in a positive sentence. This shows that UMLS is not necessarily complete.

## 5.3 Experiment

In this section a distantly supervised relation classifier is trained for two UMLS Metathesaurus relations *may-treat* and *may-prevent* using MultiR. The classifier is trained for both relations and is evaluated using manually and distantly labelled versions of the test data.

This section is structured as follows. First, the experimental setup is described. Then, two experiments are presented. The first experiment is an evaluation carried out using entity level evaluation, i.e. precision and recall are computed based on the proportion of correctly identified entity pairs which occur in sentences labelled as positive examples (according to the annotations contained within DL or ML). Entity level evaluation is commonly used to evaluate distantly supervised relation extraction systems (see Section 2.2). Results at sentence level are then presented following the same experimental setup. Sentence level evaluation measures precision and recall by examining the correct prediction of each sentence (see Section 2.2).

### 5.3.1 Experimental Setup and Overview

In order to examine different classification results on the DL and ML evaluation data, several experiments are carried out with a different number of training instances. Since machine learning methods tend to provide better results using more data, the experiment is conducted with a different amount of training data. Starting with 2,000 instances (1,000 instances of *may-treat* and 1,000 instance of *may-prevent*), the number of training examples is increased to 12,000 in increments of 2,000. Note, no instance of the training data occurs in the evaluation data (held-out). For all experiments in this chapter, the same proportion of positive and negative sentences is used for training (1:1) in order to use the same proportion as given in the DL test data (1:1). However,

---

the important aspect of this chapter is that improvements in performance observed in evaluation against distantly supervised data are not necessarily reflected in improved results when measured against manually labelled data. This aspect does not depend on the bias of the training data.

The following experiments present results in different ways. First, classification results are examined for each relation separately. This analysis shows how efficiently *may-treat* and *may-prevent* can be detected independently. Then, the overall classification results are presented. These results analyse how well instances can be predicted. This analysis takes into account that one relation (*may-treat*) contains a larger number of positive instances. Results are always reported for both labels (DL and ML) from the test data, in order to have a direct comparison between held-out evaluation and evaluation using manually labelled data.

The highlighted (bold) f1-score results in the tables indicate which classifier obtains better results on the data set (DL or ML) for each training data. Underlined results indicate that an increase in f1-score on DL data (compared to the previous result using less data) leads to a decrease of f1-score on ML data (compared to the previous result) and vice versa. The underline shows that an improved result in terms of f1-score on distantly labelled data does not imply necessarily imply an improvement on the (usually not known) true labels of the manually labelled data.

### 5.3.2 Entity level evaluation

This section presents the results using entity level evaluation. Table 5.3 shows the results for *may-prevent* and Table 5.4 the results for *may-treat*. The highlighted (bold) f1-scores indicate which classifier obtains better results on the data set (DL or ML). The underlined results indicate that an increase in f1-score using DL data, compared to the previous result using less data, leads to a decrease of f1-score using ML data compared to the previous result and vice versa<sup>1</sup>.

In general, increasing the amount of training data leads to improved results on the DL data. In particular, an increase in recall is observed when there is a larger amount of training data. However, a different pattern is observed for the ML data and increasing the amount of training data does not always lead to an improvement in the f1-score.

---

<sup>1</sup>a decrease compared to the previous result on the DL data leads to an increase on the ML data

Results also show that the performance estimates obtained using the DL and ML data sets are only loosely associated. The results are more similar for smaller training data sets but diverge as the amount of training data increases.

<i>may-prevent</i>						
	evaluation on DL			evaluation on ML		
#	prec	rec	f1	prec	rec	f1
2000	64.71	26.83	<b>37.93</b>	58.82	16.67	25.97
4000	43.33	31.71	<u>36.62</u>	66.67	33.33	<b>44.44</b>
6000	48.84	51.22	<b>50.00</b>	46.51	33.33	<u>38.83</u>
8000	47.17	60.98	<b>53.19</b>	49.06	43.33	46.02
10000	50.00	58.54	<b>53.93</b>	45.83	36.67	<u>40.74</u>
12000	47.37	65.85	<b>55.10</b>	43.86	41.67	42.74

Table 5.3: Results for relation extraction system evaluated against DL and ML data sets (may-prevent)

<i>may-treat</i>						
	evaluation on DL			evaluation on ML		
#	prec	rec	f1	prec	rec	f1
2000	51.28	48.39	<b>49.79</b>	52.99	44.29	48.25
4000	57.58	45.97	<b>51.12</b>	56.57	40.00	<u>46.86</u>
6000	60.00	43.55	<b>50.47</b>	53.33	34.29	41.74
8000	58.20	57.26	<b>57.72</b>	45.90	40.00	42.75
10000	63.89	55.65	<b>59.48</b>	51.85	40.00	45.16
12000	63.00	50.81	<b>56.25</b>	48.00	34.29	40.00

Table 5.4: Results for relation extraction system evaluated against DL and ML data sets (may-treat)

In addition, the tables highlight that for both relations the performance estimates using the DL data are in general higher than those obtained using ML. This trend becomes more pronounced as the amount of training data used increases. The most likely reason for this difference is that the classifiers are trained using distantly supervised data and therefore model the labels in the DL data set more closely than those found in ML.

These results demonstrate that evaluation using distantly labelled data tends to



overestimate performance compared to gold standard. In some cases the discrepancy is large (up to 13.19 for *may-prevent* and 16.25 for *may-treat*). However, it does not seem to be consistent or particularly predictable. Consequently, improving the performance of a relation extraction system relative to distantly labelled evaluation data does not necessarily imply an increase in performance when measured against a manually annotated gold-standard.

<i>Overall</i>						
	evaluation on DL			evaluation on ML		
#	prec	rec	f1	prec	rec	f1
2000	52.99	43.03	<b>47.49</b>	53.73	36.0	43.11
4000	54.26	42.42	<b>47.62</b>	58.91	38.0	46.20
6000	56.39	45.45	<b>50.34</b>	51.13	34.0	40.84
8000	54.86	58.18	<b>56.47</b>	46.86	41.0	43.73
10000	59.62	56.36	<b>57.94</b>	50.00	39.0	43.82
12000	57.32	54.55	<b>55.90</b>	46.50	36.5	40.90

Table 5.5: Results for relation extraction system evaluated against DL and ML data sets (Overall)

Table 5.5 presents the classification results of both relations together (overall results) and takes the frequency of different entity pair of each relation into account (micro-avg. f1-score). The results indicate similar characteristics as the *may-prevent* and *may-treat* tables. The maximum difference between results on DL and ML is 15.00 (for 12,000 examples).

### 5.3.3 Sentence level evaluation

A second experiment was conducted to examine the suitability of using held-out data for sentence level evaluation. Results of *may-prevent* and *may-treat* are presented in Table 5.6 and Table 5.7 and show a similar pattern of results to those obtained using entity level evaluation. Results obtained using the distantly labelled data tend to be higher than those for the manually labelled data in terms of f1-score (up to 7.45 for *may-prevent* and 17.72 for *may-treat*), precision and occasionally also in terms of recall. Similar to entity level evaluation, improving performance on distantly labelled

evaluation data does not imply an increase in performance when measured against a manually annotated gold-standard.

<i>may-prevent</i>						
	evaluation on DL			evaluation on ML		
#	prec	rec	f1	prec	rec	f1
2000	72.73	8.0	<b>14.41</b>	45.45	7.19	12.42
4000	61.40	17.5	27.24	59.65	24.46	<b>34.69</b>
6000	50.72	17.5	26.02	43.48	21.58	<b>28.85</b>
8000	53.57	22.5	<b>31.69</b>	36.90	22.30	27.80
10000	59.41	30.0	<b>39.87</b>	46.53	33.81	39.17
12000	54.00	27.0	<b>36.00</b>	40.00	28.78	33.47

Table 5.6: Sentence-level results for relation extraction system evaluated against DL and ML data sets (may-prevent)

<i>may-treat</i>						
	evaluation on DL			evaluation on ML		
#	prec	rec	f1	prec	rec	f1
2000	57.75	41.0	<b>47.95</b>	48.59	39.88	43.81
4000	64.29	40.5	<b>49.69</b>	50.00	36.42	42.14
6000	64.29	36.0	<b>46.15</b>	44.64	28.90	35.09
8000	63.95	47.0	<b>54.18</b>	44.22	37.57	40.63
10000	69.17	46.0	<b>55.26</b>	48.87	37.57	42.48
12000	70.87	45.0	<b>55.05</b>	44.09	32.37	37.33

Table 5.7: Sentence-level results for relation extraction system evaluated against DL and ML data sets (may-treat)

The overall results at sentence level are presented in Table 5.8 and show similar characteristics as the results of the single relations. Using the maximum number of training instances the discrepancy between DL and ML is 10.31. Surprisingly the f1-score discrepancy at sentence level is smaller than than the discrepancy at entity level.

---

<i>Overall</i>						
	evaluation on DL			evaluation on ML		
#	prec	rec	f1	prec	rec	f1
2000	59.76	24.50	<b>34.75</b>	48.17	25.32	33.19
4000	63.39	29.00	<b>39.79</b>	53.01	31.09	39.19
6000	59.12	26.75	<b>36.83</b>	44.20	25.64	32.45
8000	60.17	34.75	<b>44.06</b>	41.56	30.77	35.36
10000	64.96	38.00	<b>47.95</b>	47.86	35.90	41.03
12000	63.44	36.00	<b>45.93</b>	42.29	30.77	35.62

Table 5.8: Sentence-level results for relation extraction system evaluated against DL and ML data sets (Overall)

## 5.4 Conclusion

This chapter explored the effect of evaluating a biomedical relation extraction system using held-out data annotated via distant supervision. Test data for two biomedical relations was generated using distant supervision and then manually annotated. The manual and automatic labels differ for a large portion of the sentences. A distantly supervised relation extraction system was evaluated using both data sets. The experiments show, that evaluation using distantly labelled held-out data tends to overestimate performance. Likewise the experiments show that there is no clear connection between improved performance measured against distantly and manually labelled data.

At first glance the use of held-out data does not look promising nor reliable. The results at sentence level appear to be even less reliable (than at entity level) considering the fact that entities can express something different within each sentence. At entity level, false negatives (facts which are correct but not contained in the KB) are the main problem. Those instances might be detected by the relation extraction system but considered as a false prediction which decreases the precision of the results. The data set used for this experiment is relatively small. The re-annotation process changed the ‘truth’ of some entity pairs, because they never occur within a positive example in any sentence - even though the pair is positive according to the KB. Using more data might increase the probability to find a sentence with the entity pair expressing the relation of interest.

The use of distantly labelled held-out data is a cheap and quick method to evaluate

---

relation extraction systems. However the chapter demonstrates that results obtained should be considered with some caution and, ideally, systems should also be evaluated against manually labelled data as well. In the following chapter work will focus at entity level evaluation for the following reasons: Even though distantly labelled data at entity level suffered the problem of false positives (entities known to be positive but never occur as positives in given sentences), the evaluation only at entity level appears to be more reliable. An evaluation at sentence level is always connected to uncertainty whether a sentence is labelled correctly. However, at entity level this uncertainty also exists, but using a sufficient amount of sentences for each positive entity pair increases the probability that at least once sentence is a true positive. This supports the use of filtering step 3) in Section 4.4. In Section 6 (Reduction of Falsely Labelled Data) an inference technique will be introduced in order to reduce the number of false negatives in the data.

# Chapter 6

## Reduction of Falsely Labelled Data

This chapter focuses on increasing the quality of training data to improve classification results. Distantly labelled data contains a large number of falsely labelled instances (see Section 5.2). This chapter addresses the problem of noisy training data by detecting potential false negatives using a knowledge inference method, an approach motivated in Section 6.1. Section 6.2 introduces the inference learning method PRA (Path Ranking Algorithm) which is used to identify potentially false negatives. The experimental setup is described in Section 6.3, followed by results in Section 6.4. Section 6.5 presents an analysis of the relation paths and the reduced data. The chapter ends with a conclusion in Section 6.6.

### 6.1 Motivation

Falsely labelled data reduces the effectiveness of distantly supervised relation extraction systems. A system trained with false labels might be less accurate and an evaluation on noisy labels less meaningful. False positives occur more frequently than false negatives in distantly labelled data, as seen in Table 5.1. This thesis has attempted to address the issue of false positives in two ways: a) using a multi-instance classifier MultiR (see Section 4.5) and b) restriction to entity level evaluation (see Section 5.4) in combination with the reduction of low frequency entity pairs (see Section 4.4). MultiR is able to deal with noisy input data and the latter approach increases the probability that at least one sentence expresses the target relation. This chapter addresses the issue

---

of false negatives.

UMLS is a knowledge source with millions of related entities (see Section 2.3.2 and Section 4.1). Nonetheless, knowledge bases tend to be incomplete. There are different reasons for the lack of information. One is certainly the fact that many knowledge bases are created and updated manually. At the same time new discoveries are made and are hidden in the large amount of medical papers which are published every year. Consequently, it might take some time until new information are included into a new UMLS version, which is published twice a year.

The issue of false negatives is approached by detecting missing information in the knowledge base and removing them from the negative data. The detection of missing data is tackled with a preliminary processing step using an inference learning method applied on UMLS. This technique is able to detect new connections (relations) between entities, based on existing (and similar) connections seen in the data. The inference learning and reduction technique is described in Section 6.2. This section instead, provides a motivation and idea of the reduction of potentially false negatives<sup>1</sup>.

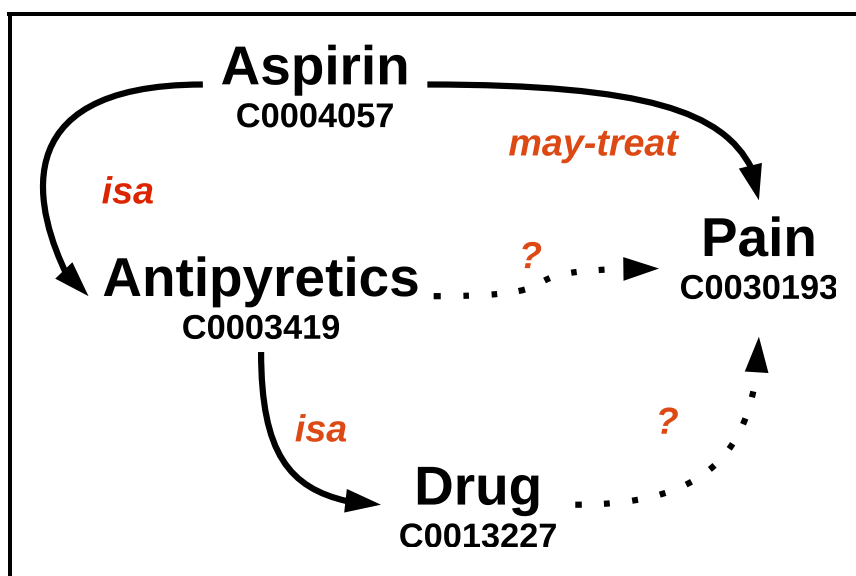


Figure 6.1: Using isa-relations to reduce false negatives

---

<sup>1</sup>The described entity pairs are called ‘potentially false negatives’ (often just ‘false negatives’), because it is difficult for individuals without biomedical expert knowledge to determine whether entities express a target relation or not. Moreover, the main objective is to improve the classification results, not the prediction of false negatives.

---

In this work negative training data is generated using the closed world assumption (see Section 2.1.2). Negative data is generated by creating new entity pair combinations from the set of related entity pairs. However, this process might produce false negative entity pairs since the knowledge base is incomplete.

Figure 6.1 and Figure 6.2 present two examples of potentially false negatives. Figure 6.1 shows that *Aspirin* can be used to treat *pain* according to the UMLS Metathesaurus. Moreover, *Aspirin* is defined as an *antipyretics* and *antipyretics* is defined as a *drug*, according to *isa* relations of UMLS. However, UMLS does not contain any relation between (*pain*, *antipyretics*) and (*pain*, *drug*). Using the closed world assumption it might happen that these entity pairs are constructed as negatives of *may-treat*. Conversely, it might be likely that the mentioned entity pairs occur in context of *may-treat*, due to its close connection.

The example in Figure 6.2 goes a step further. According to the Metathesaurus, *Oxazepam* can be used to treat *anxiety disorders*. *Oxazepam* is a benzodiazepine drug which is contained in *Serax 10mg Capsule*<sup>2</sup>. Furthermore, a tradename of *Oxazepam* is *Serax*<sup>3</sup>. In both cases (*Serax 10mg Capsule* and *Serax*) entities are closely related to *Oxazepam*. On the other hand, neither of the two entities includes any relationship to the treatment of anxiety disorder.

Both examples show entity pairs which are not defined within UMLS Metathesaurus. However, according to common-sense it is possible that those entity pairs occur together within a positive relation. Consequently, those entity pairs might decrease the quality of the distantly labelled data. UMLS is a linked knowledge base and it is likely that further useful connections, similar as those presented in the examples, exist. The idea of this chapter relies on using a technique which automatically detects potential relation paths (e.g. if *may-treat*(*Y,Z*) then also *may-treat*(*X,Z*) if the instance *has-ingredient*(*X,Y*) exists) with the aim to remove new connections (entity pairs) from the negative data set. An inference learning method such as PRA might be an ideal technique to detect those connections to detect false positives.

---

<sup>2</sup>Also known as e.g. *Oxazepam 10 MG Oral Capsule [SERAX]*.

<sup>3</sup>There are a range of different other brand names (not all are covered within UMLS) according to Wikipedia, such as e.g. *Alepam*, *Murelax* or *Opamox* (<https://en.wikipedia.org/wiki/Oxazepam>, date: 2nd of November 2015).

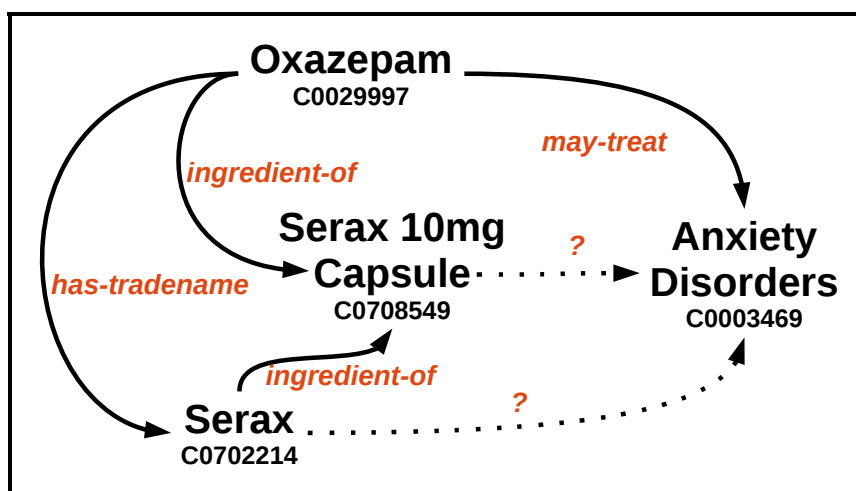


Figure 6.2: Using more relations to reduce false negatives (example 2)

## 6.2 PRA-Reduction

The path ranking algorithm PRA [Lao and Cohen, 2010; Lao et al., 2011] is an algorithm that infers new relation instances from knowledge bases. By considering a knowledge base as a graph, where nodes are connected through typed relations, it performs random walks over it and finds bounded-length relation paths that connect graph nodes. These paths are used as features in a logistic regression model, which predicts new relations in the graph. Although initially conceived as an algorithm to discover new links in the knowledge base, PRA can also be used to learn relevant relation paths for any given relation. For instance, if  $x$  and  $y$  are related via a *sibling* relation, the model trained by PRA would learn that the relation path  $parent(x,a) \wedge inverse-parent(a,y)$ <sup>4</sup> is highly relevant, as siblings share the same parents.

PRA was used for this thesis because it can construct inference methods that scale to large knowledge bases, such as UMLS. Furthermore, the results provided by PRA appear to address the task of detecting potentially false negatives in a good way. Finally, the software is freely available<sup>5</sup> and could be easily adjusted to the UMLS task.

<sup>4</sup> $\wedge$  represents a path composition and *inverse-parent* defines the inverse relation of *parent*.

<sup>5</sup>Available on <http://www.cs.cmu.edu/~nlao/>.



## 6.2.1 Removing False Negative Instances

In this chapter knowledge graphs were extracted from the NDF-RT and NCI vocabularies generating approximately 200,000 related instance pairs for NDF-RT and 400,000 for NCI (only relations with a RELA label were considered). PRA is then run on both graphs in order to learn paths for each target relation. Table 6.1 and Table 6.2 show examples of the paths PRA generated for the relation *biological-process-involves-gene-product* and *contraindicating-class-of* together with their weights. Paths with a higher weight are considered to be more useful than those ones with a lower weight. For this reason only relation paths with positive weights generated by PRA are considered for this work.

path	weight
gene-encodes-gene-product( $x,a$ ) $\wedge$ <b>inverse</b> -gene-plays-role-in-process( $a,y$ )	10.53
<b>inverse</b> -isa( $x,a$ ) $\wedge$ biological-process-involves-gene-product( $a,y$ )	6.17
isa( $x,a$ ) $\wedge$ biological-process-involves-gene-product( $a,y$ )	2.80
gene-encodes-gene-product( $x,a$ ) $\wedge$ <b>inverse</b> -gene-plays-role-in-process( $a,b$ ) $\wedge$ isa( $b,y$ )	-0.06

Table 6.1: Example PRA-induced paths and weights for the NCI relation *biological-process-involves-gene-product*.

path	weight
<b>inverse</b> -contraindicating-class-of( $x,a$ ) $\wedge$ <b>inverse</b> -CHD( $a,y$ )	0.32556
CHD( $x,a$ ) $\wedge$ <b>inverse</b> -ingredient-of( $a,b$ ) $\wedge$ contraindicating-class-of( $b,y$ )	1.64169
product-component-of( $x,a$ ) $\wedge$ contraindicating-class-of( $a,y$ )	4.59762
<b>inverse</b> -CHD( $x,a$ ) $\wedge$ CHD( $a,b$ ) $\wedge$ ingredient-of( $b,y$ )	-0.12166
may-treat( $x,a$ ) $\wedge$ may-treat( $a,b$ ) $\wedge$ contraindicating-class-of( $b,y$ )	1.76046

Table 6.2: Example PRA-induced paths and weights for the NDF-RT relation *contraindicating-class-of*.

The paths induced by PRA are used to identify potential false negatives in the negative training examples. Each negative training example is examined to check whether the entity pair is related in UMLS by following any of the relation paths extracted by PRA for the relevant target relation. Examples containing related entity pairs are assumed to be false negatives, since the relation can be inferred from the knowledge base,

---

and removed from the set of negatives training examples. For instance, using the path in the top row of Table 6.1, sentences containing the entities  $x$  and  $y$  would be removed if the path  $gene-encodes-gene-product(x,a) \wedge \mathbf{inverse-gene-plays-role-in-process}(a,y)$  could be identified within UMLS.

The CUI pair (C0072916, C0014806) ('Cisapride', 'Erythromycin'), a negative example of the relation *contraindicating-class-of*, will be removed according to the first line in Table 6.2. The following CUI instances can be found within UMLS:  $\mathbf{inverse-contraindicating-class-of}(C0072916='Cisapride', C0014809='Erythromycin\ Estolate')$   $\wedge \mathbf{inverse-CHD}(C0014809='Erythromycin\ Estolate', C0014806='Erythromycin')$ .

## 6.3 Experimental Setup

The experiment is conducted using the MultiR system with the same features as described in Chapter 4.5. Overall, the following experiment uses the same relations and same configuration as Chapter 4.

### 6.3.1 Training Data Sets

Three datasets were created to train MultiR and evaluate performance: **Unfiltered**, **Pra-Reduced** and **Random-Reduced**. The (**Unfiltered**) training set uses the data obtained using distant supervision without removing any examples identified by PRA. It is exactly the same training data as used for the experiment in Chapter 4. The **PRA-reduced** dataset is created by applying PRA reduction (Section 6.2) to the *Unfiltered* dataset to remove a portion of the negative training examples. Removing these examples produces a dataset that is smaller than *Unfiltered* and with a different bias. Changing the bias of the training data can influence the classification results. Consequently the **Random-reduced** dataset was created by removing randomly selected negative examples from *Unfiltered* to produce a dataset with the same size and bias as *PRA-reduced*. The random reduction step is repeated four times. Results are then averaged across the four steps.

Figure 6.3 visualises the construction of the data. Line numbers 3-5 present the data used for the following experiment and line numbers 1-2 present the data from Chapter 4 (see also Figure 4.3). Line 2 and Line 3 show that the training data of

*Held-Out* and *Unfiltered* are the same, only the test data changes. For the training of the *PRA-reduced* classifier further sentences of the negative data are removed. Line 5 shows, that the same number of sentences are also randomly removed from the *Random-reduced* training data.

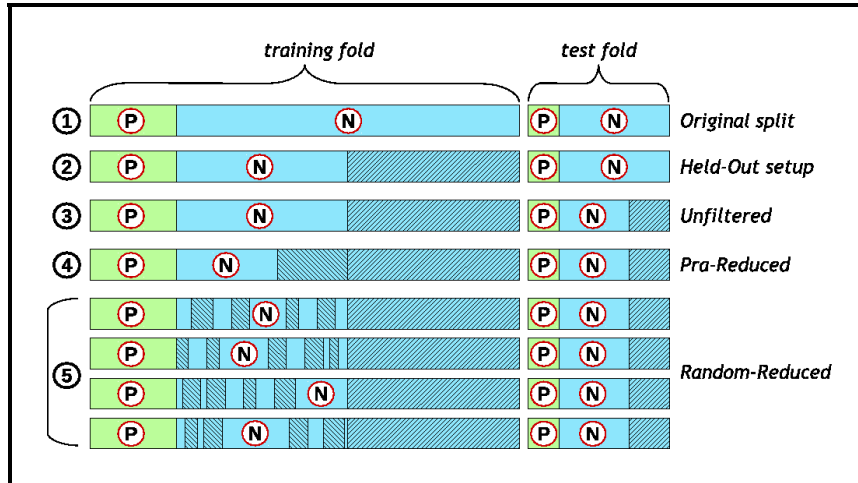


Figure 6.3: Bias adjustment

## 6.3.2 Evaluation

Two approaches were used to evaluate performance: held-out evaluation and evaluation on the gold standard. The **Held-out** datasets consist of the *Unfiltered*, *PRA-reduced* and *Random-reduced* data sets. The set of entity pairs obtained from the knowledge base is split into four parts and a process similar to 4-fold cross validation applied. In each fold the automatically labelled sentences obtained from the pairs in 3 of the quarters are used as training data and sentences obtained from the remaining quarter used for testing. Each training split is adjusted according to the description in Section 6.3.1. The average ratio of positive to negative sentences in the Held-out evaluation set is 1:5.1 (after applying further filtering steps from Section 4.4 to the data in Table 4.5). However, this changes to an average bias of 1:2.3 after removing examples identified by PRA.

The **Manually labelled** dataset is described in Section 5.1.2. This dataset is more balanced than the held-out data with a ratio of 1:1.3 for *may-treat* and 1:1.8 for *may-prevent*.

---

As in previous experiment, entity level evaluation is used since this is the most appropriate approach to determine suitability for database population. Precision and recall are computed based on the proportion of entity pairs identified. For the held-out data the set of correct entity pairs are those which occur in sentences labelled as positive examples of the relation and which are also listed as being related in UMLS. For the manually labelled data it is simply the set of entity pairs that occur in positive examples of the relation.

## 6.4 Results

### 6.4.1 Held-out data

Table 6.3 shows the results obtained using the held-out data. Overall results, averaged across all relations, are shown in the top portion of the table and indicate that applying PRA improves performance. Although the highest precision is obtained using the *Unfiltered* classifier, the *PRA-reduced* classifier leads to the best recall and F1-score. Performance of the *Random-reduced* classifier indicates that the improvement is not simply due to a change in the bias in the data but that the examples it contains lead to an improved model. The results of the *PRA-reduced* classifier are significantly better than the results of the *Unfiltered* classifier in terms of F1-Score (one-tailed paired t-test,  $p \leq 0.05$ ).

The lower part of Table 6.3 shows results for each relation. The *PRA-reduced* classifier produces the best results for the majority of relations and always increases recall compared to *Unfiltered*.

It is perhaps surprising that removing false negatives from the training data leads to an increase in recall, rather than precision. False negatives cause the classifier to generate an overly restrictive model of the relation and to predict positive examples of a relation as negative. Removing them leads to a less constrained model and higher recall.

There are two relations where there is also an increase in precision (*contraindicating-class-of* and *mechanism-of-action-of*) and these are also the ones for which the fewest training examples are available. The classifier has access to such a limited amount of data for these relations that removing the false negatives identified by PRA allows it to

MRREL relation	Unfiltered			Random-reduced			PRA-reduced		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Overall (micro avg.)	65.02	54.23	59.14	46.33	73.21	56.75	58.88	77.33	<b>66.85</b>
	NDF-RT relations								
contraindicating_class_of	60.42	32.21	41.40	40.17	78.77	52.86	54.27	82.05	<b>64.89</b>
may_prevent	47.38	26.47	33.92	24.74	38.60	28.82	47.22	47.06	<b>46.94</b>
may_treat	56.76	42.68	48.22	42.78	50.15	45.16	52.04	52.44	<b>51.97</b>
mechanism_of_action_of	62.30	65.63	62.40	52.17	71.09	59.88	54.96	75.00	<b>63.38</b>
	NCI relations								
biological_process_involves_gene_product	92.60	46.97	61.56	68.73	77.65	72.30	74.81	85.61	<b>79.83</b>
disease_has_normal_cell_origin	57.54	80.41	<b>66.78</b>	41.99	94.40	57.95	43.19	93.65	58.98
gene_product_has_associated_anatomy	36.04	61.41	45.39	25.02	70.27	36.74	39.76	67.98	<b>49.94</b>
gene_product_has_biochemical_function	85.75	73.06	78.87	64.90	91.02	75.70	73.09	95.76	<b>82.72</b>
process_involves_gene	81.55	52.68	63.69	54.92	81.49	65.30	68.04	84.74	<b>75.22</b>

Table 6.3: Evaluation using held-out data

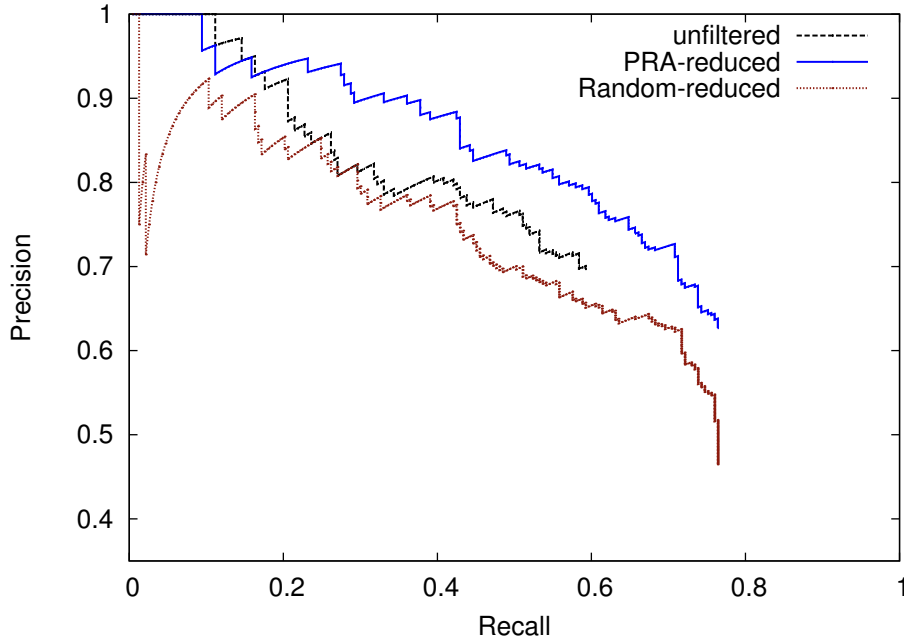


Figure 6.4: Precision/Recall Curve for Held-out data using PRA

learn a more accurate model.

Figure 6.4 presents a precision/recall curve computed using MultiR’s output probabilities. Results for the *PRA-reduced* and the *Random-reduced* classifiers show that reducing the amount of negative training data increases recall. However, using *PRA-reduced* generally leads to higher precision, indicating that PRA is able to identify suitable instances for removal from the training set. The *Unfiltered* classifier produces good results but precision and recall are lower than *PRA-reduced*.

## 6.4.2 Manually labelled

Table 6.4 shows results of evaluation on the manually labelled data set. The best overall performance is once again obtained using the *PRA-reduced* classifier. There is an increase in recall and a slight decrease in precision for both relations. Performance of the *Random-reduced* classifier does not lead to any improvements. The precision achieves comparable results to those of the *PRA-reduced* classifier. However, the recall is much lower than using the other two classifiers. These results confirm that removing examples identified by PRA improves the quality of training data. This supports the

---

initial hypothesis that removing potential false negatives from training data improves classifier predictions.

relation	Unfiltered			Random-reduced			PRA-reduced		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
may_prevent	58.06	30.0	39.56	56.62	25.67	35.32	56.76	35.0	<b>43.30</b>
may_treat	45.21	47.14	46.15	42.59	45.57	44.03	42.25	56.43	<b>48.32</b>
<b>Overall (micro)</b>	47.46	42.00	44.56	44.75	39.60	42.02	44.64	50.00	<b>47.17</b>

Table 6.4: Evaluation using manually labelled data

Further analysis indicated that the *PRA-reduced* classifier produces the fewest false negatives in its predictions on the manually annotated dataset. It incorrectly labels 82 entity pairs (45 *may-treat*, 37 *may-prevent*) as negative while *Unfiltered* predicts 120 (73, 47) and *Random-reduced* 114 (69, 45).

## 6.5 Data Analysis

The following section investigates why the PRA-reduction step tends to achieve improved classification results. First, the differences (in terms of classification results) between the PRA-reduced and the Random-reduced data set are examined. Then, a range of paths for different relations are examined.

### 6.5.1 Examination of PRA-reduced data

Negative instances removed by using PRA are compared to those which are randomly reduced. The entity pairs and sentences which occur in the Random-reduced but not in the PRA-reduced data might provide an answer why the usage of PRA leads to better classification results. It is likely that exactly those sentences contain some false negatives, otherwise the result would not improve. The following examination serves only as example and focusses on one cross-validation step of *may-prevent*.

Within the selected cross-validation step 319 different CUI pairs are removed from the set of negative *may-prevent* instances. Considering one Random-reduced setup of the same cross-validation step 278 CUI pairs (of the 319) occur as negatives in the data set. 20 of those CUI pairs were randomly selected and examined including all

---

Does the prophylactic administration of [C0001047:N-acetylcysteine] prevent [C0022660:acute kidney injury] following cardiac surgery? (PMID=20570977)

Figure 6.5: PRA-reduced example 1

[C0052796:Azithromycin] for treating uncomplicated [C0041466:typhoid] and paratyphoid fever (enteric fever). (PMID=18843701)

Figure 6.6: PRA-reduced example 2

sentences containing those entities. 4 different CUI pairs include sentences expressing a prevention in a wider sense such as shown in Figure 6.5. Furthermore, seven CUI pairs occur in sentences in context of a treatment such as shown in Figure 6.6. For the given example of 20 CUI pairs, 20% of them occur in a close context of the target relation. Another 35% of entity pairs occur as falsely labelled data in context of a relation similar to *may-prevent*.

The given example shows that PRA was able to identify entity pairs that are closely connected to the target relation. Removing those entity pairs from the negative data can help to generate distantly labelled data of a higher quality (less noise). Thus, PRA-induced paths are able to decrease the amount of false negatives in the training data. Since *may-treat* and *may-prevent* are closely connected, also false negatives of *may-treat* are removed from the data set.

## 6.5.2 Pattern Analysis

A subset of relation paths taken from two relations are examined to demonstrate that paths generated by PRA appear to be very useful for the reduction of false negatives.

Table 6.5 presents a selection of PRA-induced patterns generated for *biological-process-involves-gene-product*. Many relation paths appear to be reasonable according to the examples presented in the motivation (see Figure 6.1 and Figure 6.2). The table shows that many paths involve the main target relation in combination with re-



lations such as *isa*, *inverse-gene-product-is-physical-part-of* or *inverse-gene-product-has-associated-anatomy*.

path	weight
<b>inverse-CHD-isa</b> $\wedge$ biological-process-involves-gene-product	6.11759
<b>inverse-gene-product-is-physical-part-of</b> $\wedge$ biological-process-involves-gene-product	0.46255
<b>inverse-gene-product-has-associated-anatomy</b> $\wedge$ gene-product-has-associated-anatomy $\wedge$ biological-process-involves-gene-product	1.21214
<b>inverse-gene-product-affected-by-chemical-or-drug</b> $\wedge$ <b>inverse-chemical-or-drug-has-physiologic-effect</b> $\wedge$ <b>inverse-CHD-isa</b>	0.01531
gene-encodes-gene-product $\wedge$ CHD-isa $\wedge$ process-involves-gene	9.67983
gene-product-has-biochemical-function $\wedge$ biological-process-involves-gene-product	0.84485

Table 6.5: Example PRA-induced paths and weights for the NCI relation *biological-process-involves-gene-product*.

Table 6.6 presents a selection of PRA-induced patterns generated for *may-prevent*. At the first glance the results are surprising. Many relation paths contain the relation *may-treat* and only a few paths contain *may-prevent* relations. However, CUI pairs of *may-treat* and *may-prevent* are closely connected. Many instances of *may-prevent* also occur as instance within *may-prevent*, such as CUIs connected to *Vitamin E* and *Alzheimer*. An analysis of UMLS reveals, that *may-prevent* contains 4838 different CUI pairs. 2416 of those CUI pairs also occur in *may-treat* which expresses a close connection of both relations. Accordingly, it seems reasonable that *may-treat* occurs in the relation paths of *may-prevent*.

Taking the similarity between both relations into account, relation paths also appear to be very reasonable. Many PRA-induced paths generated for *may-prevent* combine relations such as *CHD* (child/parent relations), *ingredient-of* or *product-component* which also support the motivation of this chapter.

## 6.6 Conclusions

This chapter proposed a novel approach to identifying potentially incorrectly labelled instances generated using distant supervision. The method targets in particular false

---

<b>path</b>	<b>weight</b>
may-treat	5.64337
may-treat $\wedge$ <b>inverse</b> -CHD $\wedge$ CHD	2.81085
may-treat $\wedge$ <b>inverse</b> -CHD $\wedge$ <b>inverse</b> -ingredient-of	1.76814
may-treat $\wedge$ <b>inverse</b> -CHD $\wedge$ <b>inverse</b> -product-component-of	0.66923
may-treat $\wedge$ ingredient-of	0.78277
may-treat $\wedge$ ingredient-of $\wedge$ <b>inverse</b> -product-component-of	0.58369
may-treat $\wedge$ <b>inverse</b> -may-prevent $\wedge$ may-treat	4.66815
may-treat $\wedge$ mechanism-of-action-of $\wedge$ <b>inverse</b> -mechanism-of-action-of	0.56310
may-treat $\wedge$ physiologic-effect-of $\wedge$ <b>inverse</b> -physiologic-effect-of	1.78688
may-treat $\wedge$ drug-contraindicated-for $\wedge$ drug-contraindicated-for	0.22314

Table 6.6: Example PRA-induced paths and weights for the NCI relation *may-prevent*.

negative entity pairs and sentences in the negative data which are closely connected to the target relations. The presented method bases on the idea of using an inference learning method applied to the UMLS knowledge graph in order to detect potentially falsely labelled instances in the negative data.

The presented experiments in this chapter showed that removing those inferred instances from the negative data improves significantly the classification results for many UMLS relations. The method has been evaluated using held-out and gold standard data. Furthermore, an analysis of the removed data supported the assumption that inference learning can be used to clean distantly labelled training data for relation extraction.

# Chapter 7

## Conclusions

The final chapter of this thesis provides a summary of the work described in this thesis and possible future directions.

### 7.1 Summary of thesis

This thesis explored the detection of biomedical relations using distant supervision. Distantly supervised relation extraction is very useful when no manually labelled training data is available. The technique uses a knowledge base and labels training data automatically using given facts. Thus, large amount of training data can be generated easily but it is usually of lower quality than manually labelled instances.

In a first step the thesis showed that when only a small set of manually labelled data is available, distantly labelled data can significantly improve the classification results. This technique has been tested in the context of adverse-drug effects.

The main focus of this work is the examination whether UMLS is a useful source for distant supervision. Based on two example UMLS source vocabularies NDF-RT and NCI, this thesis shows that it is indeed a very useful source to detect similar information in natural language.

The thesis also explores evaluation of distantly supervised relation extraction when no manually labelled data is available. A very popular method for this use case is using held-out data. Held-out data is generated by splitting facts from a knowledge base into two sets to generate a distantly labelled training and a distantly labelled evaluation

---

set. This means that the classifier is trained and evaluated against noisy data. Many approaches use held-out evaluation to measure the efficiency of their system. However, this work examined how useful an evaluation against held-out data is and what exactly the results tell us. Overall, the results show that evaluation using held-out data tends to overperform an evaluation on the same data using gold labels. Furthermore, the results show that system improvements based on held-out data do not necessarily imply an improvement on gold data.

Finally, a new method was introduced with the aim to improve the quality of distantly labelled data and increasing the classification results of UMLS relations. The method presented an inference learning method with the goal to detect potentially missing information in the knowledge base to reduce the amount of false negatives in the data. The method is able to detect false negatives but also true negatives. However, results show that using the inference learning method to reduce ‘potentially’ falsely labelled instances improves the classification results.

## 7.2 Future directions

The thesis explores a range of different methods in the context of distantly supervised relation extraction in the biomedical domain. Future directions include the following:

### **Detecting and examining new instances**

This work examined whether UMLS is a useful source for distantly supervised relation extraction. The results have been carried out on gold standard and on held-out data. Evaluation was carried out using precision, recall and f1-score and the quality of a system measured with numbers between 0-100. The presented work shows that related instance pairs of UMLS can be detected within sentences of publications. However, in future work it would be interesting to put focus on new predictions itself (exploration of facts which are not known according to UMLS).

Figure 7.1 for instance, presents a sentence that expresses the usage of *dapsone* for the prophylaxis of *pneumocystis carinii pneumonia* (but may have a negative effect). The CUI pair in this example does not occur as related in UMLS. In future work it would be desirable to examine positively predicted instance pairs in more detail.

---

Second, the inadvertent simultaneous administration of low doses of oral iron with [C0010980:dapsone] for the prophylaxis of [C0032305:Pneumocystis carinii pneumonia] in HIV-positive patients may have been associated with excess mortality. (PMID=11166657)

Figure 7.1: Example sentence of a positively predicted, but unknown entity pair

Which new entity pairs can be detected? Why are those pairs are not within UMLS? Do entities occur in contraindicating sentences or do entities describe novel findings in biomedical research?

The examination could be interesting for the biomedical community and could be potentially used to support the development of UMLS. The advantage is, that further system extensions are not necessarily required. The system used in Chapter 6 seems to be already sufficient to address this task. However, in order to find and evaluate new instance pairs of a relation predicted information need to be examined manually. This requires a collaboration with a biomedical expert who might be able to provide a new perspective to new detected entity pairs.

## Processing of relations

This thesis shows that relations from UMLS can be detected in natural language. In a next step it might be useful to use the extracted information for further processing steps.

Particular relations could be extracted from a large number of Medline publications published in different years or published in different countries. For instance lung cancer treatments might have changed over the years. Using relation extraction it could be possible to explore differences of treatments across years and countries - also in combination with side effects of those treatments. It could be possible to explore questions such as ‘How did lung cancer treatment change in the last 10 years?’.

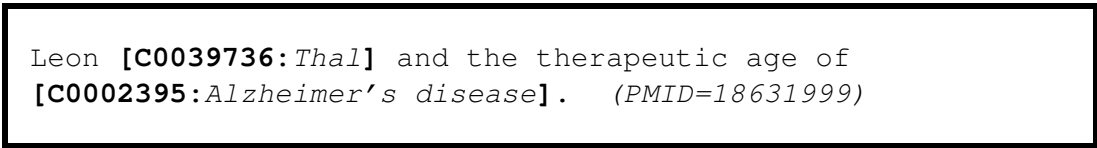
The implementation of a system capable of solving this task is relatively simple and could be addressed with the previously reported relation extraction system. The relation extraction system could run on PubMed subsets in order to extract a relation containing a target entity pair (e.g. ‘lung cancer’). Next, statistics of treatments can be

---

extracted and used for further analysis. Furthermore, this data could be used to identify trends of medical treatments using methods similar as to Lukasik et al. [2015].

## Improving selection of named entities

As Augenstein et al. [2015] already mentioned, “An important first step in distant supervision is to identify named entities (NEs) and their types to determine if a pair of NEs is a suitable candidate for the relation.” In this thesis MetaMap has been used to detect medical concepts in natural language. Named entities provided by MetaMap were in general considered as correct. However, some named entities are predicted (or selected) incorrectly. A closer analysis of the data actually reveals a range of different errors and problems connected to MetaMap. An example is given in figure 7.2. The sentence annotates a surname as *pharmacological substance (C0039736)*.



```
Leon [C0039736:Thal] and the therapeutic age of  
[C0002395:Alzheimer's disease]. (PMID=18631999)
```

Figure 7.2: False named entity recognition

This work focuses on biomedical relation extraction and not on named entity recognition. Named entities were taken as granted. Only little attempt has been taken to reduce false labels (such as reducing non-nouns). Biomedical named entity recognition is an ongoing challenge and an important task. Improving the quality of biomedical named entity recognition might positively influence tasks which rely on a correct named entity recognition, such as relation extraction, summarisation or sentiment analysis.

In order to improve the current MetaMap-NER task various possibilities exist. One possibility could take the mappings of MetaMap as baseline and apply some expert-rules to discard some falsely labelled entities. Using additional POS information for instance might help to reduce errors such as mentioned in Section 2.3.3 (a colour is annotated as disease) or Figure 7.2 (e.g. other noun phrases are next to the annotation).

Another possibility to improve the annotation could be the disambiguation of the MetaMap candidate list. As mentioned in Section 2.3.3, MetaMap provides a large list of candidates for different words and medical concepts in text. In order to reduce noise

---

(many instances were not useful) only the highest ranked mappings has been chosen. Approaches such as Agirre et al. [2014] or Weissenborn et al. [2015] use random walks over a knowledge base to disambiguate the list of candidates. This technique could be applied for UMLS as well, maybe in combination with some background knowledge such as a restriction to a domain vocabulary or a focus on particular target semantic types.

## References

- Azad Abad and Alessandro Moschitti. Creating a standard for evaluating distant supervision for relation extraction. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014*, Pisa, Italy, 2014. University Press. 21, 28, 67
- Eugene Agichtein and Luis Gravano. Snowball: extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, New York, NY, USA, 2000. ACM. 21
- Eneko Agirre, de Oier López Lacalle, and Aitor Soroa. Random Walks for Knowledge-Based Word Sense Disambiguation. *Computational Linguistics*, 17(3), 2014. 109
- Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. A Graph Kernel for Protein-Protein Interaction Extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, Columbus, Ohio, USA, 2008. Association for Computational Linguistics. 51
- Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D. Manning. Combining Distant and Partial Supervision for Relation Extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014. Association for Computational Linguistics. 21, 23, 30
- Alessio Palmero Aprosio, Claudio Giuliano, and Alberto Lavelli. Extending the Coverage of DBpedia Properties using Distant Supervision over Wikipedia. In *Proceedings of the 1st workshop on NLP & DBpedia (ISWC)*, Sydney, Australia, 2013. Springer. 28



## REFERENCES

---

- Alan Aronson and Francois-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Association*, 17(3), 2010. 40, 41
- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference*, Busan, Korea, 2007. Springer. 28
- Isabelle Augenstein, Diana Maynard, and Fabio Ciravegna. Relation Extraction from the Web using Distant Supervision. In *Proceedings of the 19th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2014)*, Linköping, Sweden, 2014. Springer. 27, 28, 32
- Isabelle Augenstein, Andreas Vlachos, and Diana Maynard. Extracting Relations between Non-Standard Entities using Distant Supervision and Imitation Learning. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015. Association for Computational Linguistics. 67, 108
- Michele Banko and Eric Brill. Scaling to Very Very Large Corpora for Natural Language Disambiguation. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, 2001. Association for Computational Linguistics. 14, 23
- Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction for the web. In *Twentieth International Joint Conference on Artificial Intelligence*, volume 7, Hyderabad, India, 2007. AAAI Press. 21
- David S. Batista, Bruno Martins, and Mário J. Silva. Semi-Supervised Bootstrapping of Relationship Extractors with Distributional Semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015. Association for Computational Linguistics. 21

## REFERENCES

---

- Kedar Bellare and Andrew McCallum. Learning Extractors from Unlabeled Text using Relevant Databases. In *Proceedings of the Sixth International Workshop on Information Extraction on Web*, Vancouver, Canada, 2007. AAAI Press. 68
- Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1), 2000. 29
- Lidong Bing, Sneha Chaudhari, Richard Wang, and William Cohen. Improving Distant Supervision for Information Extraction Using Label Propagation Through Lists. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015. Association for Computational Linguistics. 66
- Jari Björne and Tapio Salakoski. TEES 2.1: Automated Annotation Scheme Learning in the BioNLP 2013 Shared Task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 21, 22
- Jari Björne, Filip Ginter, and Tapio Salakoski. University of Turku in the BioNLP'11 Shared Task. *BMC Bioinformatics*, 13(Suppl 11), 2012. 23
- Jari Björne, Suwisa Kaewphan, and Tapio Salakoski. UTurku: Drug Named Entity Recognition and Drug-Drug Interaction Extraction Using SVM Classification and Domain Knowledge. In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics. 23
- Tamara Bobic, Roman Klinger, Philippe Thomas, and Martin Hofmann-Apitius. Improving Distantly Supervised Extraction of Drug-Drug and Protein-Protein Interactions. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, Avignon, France, 2012. Association for Computational Linguistics. 27, 29, 31, 64
- Olivier Bodenreider and Alexa T McCray. Exploring semantic groups through visual approaches. *Journal of Biomedical Informatics*, 36(6), 2003. 10, 46, 130

## REFERENCES

---

- Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Free-base: a collaboratively created graph database for structuring human knowledge. In *ACM SIGMOD International Conference on Management of Data*, Vancouver, Canada, 2008. ACM. 26
- Robert Bossy, Wiktorina Golik, Zorana Ratkovic, Philippe Bessières, and Claire Nédellec. BioNLP shared Task 2013 – An Overview of the Bacteria Biotope Task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 22, 59
- Thorsten Brants, Ashok C. Papat, Peng Xu, Franz J. Och, and Jeffrey Dean. Large Language Models in Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, 2007. Association for Computational Linguistics. 14, 23
- Sergey Brin. Extracting Patterns and Relations from the World Wide Web. In *Selected papers from the International Workshop on The World Wide Web and Databases (WebDB '98)*, London, UK, 1999. Springer-Verlag. 21
- Quoc-Chinh Bui, David Campos, Erik van Mulligen, and Jan Kors. A fast rule-based approach for biomedical event extraction. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 23
- Razvan Bunescu and Raymond Mooney. Learning to Extract Relations from the Web using Minimal Supervision. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, 2007. Association for Computational Linguistics. 26, 27, 28
- Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 2011. 51
- Eugene Charniak and Mark Johnson. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. 52, 70

## REFERENCES

---

- Nitesh Chawla, Nathalie Japkowicz, and Aleksander Kołcz. Editorial: Special Issue on Learning from Imbalanced Data Sets. *ACM SIGKDD Explorations Newsletter*, 6, 2004. 72
- Xue-wen Chen, Jong Cheol Jeong, and Patrick Dermeyer. KUPS: constructing datasets of interacting and non-interacting protein pairs with associated attributions. *Nucleic Acids Research*, 39(suppl 1), 2011. 29
- Md. Faisal Mahbub Chowdhury and Alberto Lavelli. Impact of Less Skewed Distributions on Efficiency and Effectiveness of Biomedical Relation Extraction. In *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India, 2012. The Coling 2012 Organizing Committee. 21
- Md. Faisal Mahbub Chowdhury and Alberto Lavelli. FBK-irst : A Multi-Phase Kernel Based Approach for Drug-Drug Interaction Detection and Classification that Exploits Linguistic Information. In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA, 2013a. Association for Computational Linguistics. 23
- Md. Faisal Mahbub Chowdhury and Alberto Lavelli. Exploiting the Scope of Negations and Heterogeneous Features for Relation Extraction: A Case Study for Drug-Drug Interaction Extraction. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 2013b. Association for Computational Linguistics. 23
- Md. Faisal Mahbub Chowdhury and Pierre Zweigenbaum. A controlled greedy supervised approach for co-reference resolution on clinical text . *Journal of Biomedical Informatics*, 46(3), 2013. 20
- Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 1960. 80
- Mark Craven and Johan Kumlien. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. In *Proceedings of the Seventh Inter-*

## REFERENCES

---

- national Conference on Intelligent Systems for Molecular Biology (ISMB)*. AAAI Press, 1999. 26, 27, 28, 31
- Aron Culotta. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, 2004. Association for Computational Linguistics. 22
- James R. Curran and Miles Osborne. A Very Very Large Corpus Doesn't Always Yield Reliable Estimates. In *Proceedings of the 6th conference on Natural language learning*, Taipei, Taiwan, 2002. 14
- Allan Peter Davis, Cynthia G. Murphy, Cynthia A. Saraceni-Richards, Michael C. Rosenstein, Thomas C. Wieggers, and Carolyn J. Mattingly. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Research*, 37(suppl 1), 2009. 29
- Thomas G. Dietterich, Richard H. Lathrop, Tomas Lozano-Perez, and Arris Pharmaceutical. Solving the Multiple-Instance Problem with Axis-Parallel Rectangles. *Artificial Intelligence*, 89, 1997. 26, 27
- Jon O. Ebbert, Denise M. Dupras, and Patricia J. Erwin. Searching the Medical Literature Using PubMed: A Tutorial. *Mayo Clinic Proceedings*, 78(1), 2003. 13
- Tilia Ellendorff, Fabio Rinaldi, and Simon Clematide. Using Large Biomedical Databases as Gold Annotations for Automatic Relation Extraction. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014. ELRA. 24, 29
- Kathleen M. Giacomini, Ronald M. Krauss, Dan M. Roden, Michel Eichelbaum, Michael R. Hayden, and Yusuke Nakamura. When good drugs go bad. In *Nature*, 466(7139), 2007. 45
- Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 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)*, Trento, Italy, 2006. Association for Computational Linguistics. 22, 23, 51

## REFERENCES

---

- Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Leif Hanlen, Aurlie Nvol, Cyril Grouin, Joo Palotti, and Guido Zuccon. Overview of the CLEF eHealth Evaluation Lab 2015. In Josiane Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, GarethJ.F. Jones, Eric SanJuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 9283 of *Lecture Notes in Computer Science*. Springer International Publishing, 2015. 22
- Harsha Gurulingappa, Abdul MateenRajput, and Luca Toldo. Extraction of potential adverse drug events from medical case reports. *Journal of Biomedical Semantics*, 3(1), 2012a. 21, 45, 46, 51
- Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5), 2012b. Text Mining and Natural Language Processing in Pharmacogenomics. 45
- Kai Hakala, Sofie Van Landeghem, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 22
- David A Hanauer, Mohammed Saeed, Kai Zheng, Qiaozhu Mei, Kerby Shedden, Alan R Aronson, and Naren Ramakrishnan. Applying MetaMap to Medline for identifying novel associations in a large clinical dataset: a feasibility analysis. *Journal of the American Medical Informatics Association*, 21(5), 2014. 42
- Peter E. Hodges, William E. Payne, and James I. Garrels. The Yeast Protein Database (YPD): A curated proteome database for *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 26(1), 1998. 26
- Raphael Hoffmann, Congle Zhang, and Daniel S. Weld. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 28, 31, 77

## REFERENCES

---

- Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, USA, 2011. 27, 69
- Ander Intxaurreondo, Mihai Surdeanu, Oier Lopez de Lacalle, and Eneko Agirre. Removing Noisy Mentions for Distant Supervision. *Procesamiento del Lenguaje Natural*, 51, 2013. 28, 32
- Thorsten Joachims. Making Large-scale Support Vector Machine Learning Practical. *Advances in Kernel Methods*, 1999. 23
- Clement Jonquet, Nigam H. Shah, and Mark A. Musen. The Open Biomedical Annotator. *Summit on Translational Bioinformatics*, 2009. 42
- Ning Kang, Bharat Singh, Chinh Bui, Zubair Afzal, Erik van Mulligen, and Jan Kors. Knowledge-based extraction of adverse drug events from biomedical text. *BMC Bioinformatics*, 15(1), 2014. 21, 45, 46
- Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. Introduction to the Bio-entity Recognition Task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (JNLPBA '04)*, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. 21
- Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1):10, 2008. 23
- Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. Overview of BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, Portland, Oregon, USA, 2011. Association for Computational Linguistics. 14, 22
- Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. The Genia Event Extraction Shared Task, 2013 Edition - Overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 22

## REFERENCES

---

- Dan Klein and Christopher D. Manning. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 2003. Association for Computational Linguistics. 70
- Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, An Chi Guo, and David S. Wishart. DrugBank 3.0: a comprehensive resource for ‘Omics’ research on drugs. *Nucleic Acids Research*, 39(1), 2011. 29
- Parisa Kordjamshidi, Dan Roth, and Marie-Francine Moens. Structured Learning for Spatial Information Extraction from Biomedical Text: Bacteria Biotopes. *BMC Bioinformatics* 16.1 (2015): 129. *PMC. Web. 20*, 2015. 30, 44
- Sebastian Krause, Hong Li, Hans Uszkoreit, and Feiyu Xu. Large-Scale Learning of Relation-extraction Rules with Distant Supervision from the Web. In *Proceedings of the 11th International Semantic Web Conference. International Semantic Web Conference (ISWC-2012)*, Boston, USA, 2012. Springer-Verlag. 23
- Ni Lao and William W. Cohen. Relational Retrieval Using a Combination of Path-constrained Random Walks. *Machine Learning*, 81(1), 2010. 18, 94
- Ni Lao, Tom Mitchell, and William W. Cohen. Random Walk Inference and Learning in A Large Scale Knowledge Base. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK, 2011. Association for Computational Linguistics. 18, 94
- Lishuang Li, Yiwen Wang, and Degen Huang. Improving feature-based biomedical event extraction system by integrating argument information. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 22
- Wesley T. Lindsey and Bernie R. Olin. PubMed Searches: Overview and Strategies for Clinicians. *Nutrition in Clinical Practice*, 28(2), 2013. 13
- Xiao Ling, Peter Clark, and Daniel Weld. Extracting Meronyms for a Biology Knowledge Base Using Distant Supervision. In *Proceedings of the 3rd Workshop on Auto-*



## REFERENCES

---

- matic KB Construction (AKBC'13)*, San Francisco, USA, 2013. ACM. 25, 27, 29, 59
- Chun-Chi Liu, Yu-Ting Tseng, Wenyuan Li, Chia-Yu Wu, Ilya Mayzus, Andrey Rzhetsky, Fengzhu Sun, Michael Waterman, Jeremy J. W. Chen, Preet M. Chaudhary, Joseph Loscalzo, Edward Crandall, and Xianghong Jasmine Zhou. DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. *Nucleic Acids Research*, 42(W1), 2014a. 42
- Mengwen Liu, Yuan Ling, Yuan An, Xiaohua Hu, Alan Yagoda, and Rick Misra. Relation Extraction from Biomedical Literature with Minimal Supervision and Grouping Strategy. In *Proceedings of IEEE Conference on Bioinformatics and Biomedicine (BIBM14)*, Belfast, UK, 2014b. 27, 29, 66, 68
- Xiao Liu, Antoine Bordes, and Yves Grandvalet. Biomedical Event Extraction by Multi-class Classification of Pairs of Text Entities. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 21
- Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. Point Process Modelling of Rumour Dynamics in Social Media. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China, 2015. Association for Computational Linguistics. 108
- Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. Distant Supervision for Relation Extraction with an Incomplete Knowledge Base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 2013. Association for Computational Linguistics. 27, 28
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 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*, Suntec, Singapore, 2009. Association for Computational Linguistics. 21, 24, 26, 27, 28, 31, 68, 69, 77

## REFERENCES

---

- Alessandro Moschitti. A Study on Convolution Kernels for Shallow Statistic Parsing. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, Barcelona, Spain, 2004. Association for Computational Linguistics. 23
- Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. Overview of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 14, 22
- Truc-Vien T. Nguyen and Alessandro Moschitti. 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*, Portland, Oregon, USA, 2011a. Association for Computational Linguistics. 25, 27, 28, 68
- Truc-Vien T. Nguyen and Alessandro Moschitti. Joint distant and direct supervision for relation extraction. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, Chiang Mai, Thailand, 2011b. Association for Computational Linguistics. 30, 44
- Ivelina Nikolova and Galia Angelova. Identifying relations between medical concepts by parsing UMLS definitions. In *Conceptual Structures for Discovering Knowledge*. Springer, 2011. 29, 59, 71
- Thiago Nunes and Daniel Schwabe. Building Distant Supervised Relation Extractors. In *Proceedings of the 2014 IEEE International Conference on Semantic Computing*. IEEE Computer Society, 2014. 21, 23
- Anika Oellrich, Nigel Collier, Damian Smedley, and Tudor Groza. Generation of Silver Standard Concept Annotations from Biomedical Texts with Special Relevance to Phenotypes. *PloS one*, 10(1), 2015. 42
- Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006. Association for Computational Linguistics. 59

## REFERENCES

---

- Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. Infusion of Labeled Data into Distant Supervision for Relation Extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, Maryland, 2014. Association for Computational Linguistics. 24, 30
- Hoifung Poon, Kristina Toutanova, and Chris Quirk. Distant Supervision for Cancer Pathway Extraction from Text. In *Biocomputing 2015: Proceedings of the Pacific Symposium*, Hawaii, USA, 2015. World Scientific Publishing. 24, 29, 67
- M. F. Porter. An Algorithm for Suffix Stripping. In *Readings in Information Retrieval*, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. 51, 70
- Judita Preiss, Mark Stevenson, and Robert Gaizauskas. Exploring Relation Types for Literature-based Discovery. *Journal of the American Medical Informatics Association*, 2015. 42
- Sampo Pyysalo, Tomoko Ohta, and Sophia Ananiadou. Overview of the Cancer Genetics (CG) task of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 22
- K.E. Ravikumar, Haibin Liu, J.D. Cohn, M.E. Wall, and K. Verspoor. Pattern Learning through Distant Supervision for Extraction of Protein-Residue Associations in the Biomedical Literature. In *10th International Conference on Machine Learning and Applications and Workshops (ICMLA)*, volume 2. IEEE, 2011. 29
- Soumya Ray and Mark Craven. Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of 22nd International Conference on Machine Learning (ICML-2005)*, pages 697–704. ACM Press, 2005. 69
- Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling Relations and Their Mentions without Labeled Text. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD '10)*. Springer, 2010. 23, 25, 26, 27, 31, 32, 72, 77

## REFERENCES

---

- Ellen Riloff. Automatically Constructing a Dictionary for Information Extraction Tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, 1993. 21
- Ellen Riloff and Rosie Jones. Learning Dictionaries for Information Extraction by Multi-level Bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, Menlo Park, CA, USA, 1999. American Association for Artificial Intelligence. 21
- Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. Modeling Missing Data in Distant Supervision for Information Extraction. In *Association for Computational Linguistics Vol. 1 (TACL)*, 2013. 24, 27, 28, 68
- Angus Roberts, Robert Gaizauskas, and Mark Hepple. Extracting Clinical Relationships from Patient Narratives. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, Columbus, Ohio, 2008. Association for Computational Linguistics. 20
- Roland Roller and Mark Stevenson. Identification of Genia Events using Multiple Classifiers. In *Proceedings of BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 17
- Roland Roller and Mark Stevenson. Self-Supervised Relation Extraction using UMLS. In *Proceedings of the Conference and Labs of the Evaluation Forum 2014*, Sheffield, England, September 2014a. Springer. 17
- Roland Roller and Mark Stevenson. Applying UMLS for Distantly Supervised Relation Detection. In *Proceedings of the Fifth International Workshop on Health Text Mining and Information Analysis (Louhi 2014)*, Gothenburg, Sweden, 2014b. Association for Computational Linguistics. 17
- Roland Roller and Mark Stevenson. Making the most of limited training data using distant supervision. In *Proceedings of the BioNLP 2015 Workshop*, Beijing, China, 2015a. 16

## REFERENCES

---

- Roland Roller and Mark Stevenson. Held-out versus Gold Standard: Comparison of Evaluation Strategies for Distantly Supervised Relation Extraction from Medline abstracts. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, Lisbon, Portugal, 2015b. Association for Computational Linguistics. 17
- Roland Roller, Eneko Agirre, Aitor Soroa, and Mark Stevenson. Improving distant supervision using inference learning. In *Proceedings of the Joint Conference of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference of the Asian Federation of Natural Language Processing*, Beijing, China, 2015. 17
- Barbara Rosario and Marti A. Hearst. Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. 20, 21
- Benjamin Roth and Dietrich Klakow. Combining Generative and Discriminative Model Scores for Distant Supervision. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, USA, 2013. Association for Computational Linguistics. 24
- Benjamin Roth, Tassilo Barth, Michael Wiegand, and Dietrich Klakow. A Survey of Noise Reduction Methods for Distant Supervision. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, New York, NY, USA, 2013. ACM. 28
- Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5), 2010. 42
- Carl F. Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H. Buetow. PID: the Pathway Interaction Database. *Nucleic Acids Research*, 37(suppl 1), 2009. 29

## REFERENCES

---

- Isabel Segura-Bedmar, Paloma Martínez, and Cesar de Pablo-Sánchez. Using a shallow linguistic kernel for drug-drug interaction extraction . *Journal of Biomedical Informatics*, 44(5), 2011a. 51
- Isabel Segura-Bedmar, Paloma Martínez, and Daniel Sánchez-Cisneros. The 1st DDI Extraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. In *Proceedings of DDI Extraction-2011 challenge task.*, 2011b. 22
- Isabel Segura-Bedmar, Paloma Martínez, and María Herrero Zazo. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. 14, 22
- Matthew S. Simpson and Dina Demner-Fushman. Biomedical Text Mining: A Survey of Recent Progress. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*. Springer US, 2012. 21, 42
- Pawel Smialowski, Philip Wong, Barbara Brauner, Irmtraud Dunger, Gisela Fobo, Goar Frishman, Corinna Montrone, Thomas Rattei, and Dmitrij Frishman. The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res*, 2010. 25
- Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems (NIPS 2004)*, 2004. 59
- Stephen Soderland, David Fisher, Jonathan Aseltine, and Wendy Lehnert. CRYSTAL: Inducing a Conceptual Dictionary. In Chris Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, San Francisco, 1995. Morgan Kaufmann. 21
- Fabian M. Suchanek, Gerhard Weikum, Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th international World Wide Web conference (WWW 2007)*, Banff, Canada, 2007. CEUR-WS.org. 28

## REFERENCES

---

- Mihai Surdeanu, David McClosky, Mason Smith, Andrey Gusev, and Christopher Manning. Customizing an Information Extraction System to a New Domain. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. 68, 69
- Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. Multi-instance Multi-label Learning for Relation Extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12)*, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. 26, 27, 32, 68, 72
- Kumutha Swampillai and Mark Stevenson. Extracting Relations Within and Across Sentences. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, Hissar, Bulgaria, 2011. RANLP 2011 Organising Committee. 20, 72
- Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL '12)*, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. 25, 28, 32, 67, 78
- Philippe Thomas, Illés Solt, Roman Klinger, and Ulf Leser. Learning Protein Protein Interaction Extraction using Distant Supervision. In *Proceedings of Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing*, 2011. 14, 25, 29, 31, 50, 72
- Philippe Thomas, Mariana Neves, Tim Rocktäschel, and Ulf Leser. WBI-DDI: Drug-Drug Interaction Extraction using Majority Voting. In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics. 23
- Mai-Vu Tran, Nigel Collier, Hoang-Quynh Le, Van-Thuy Phi, and Thanh-Binh Pham. Exploring a Probabilistic Earley Parser for Event Composition in Biomedical Texts.

## REFERENCES

---

- In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 23
- Kateryna Tymoshenko, Swapna Somasundaran, Vinodkumar Prabhakaran, and Vinay Shet. Relation Mining in the Biomedical Domain using Entity-level Semantics. In *ECAI 2012: 20th European Conference on Artificial Intelligence*, Montpellier, France, 2012. IOS Press. 29, 64
- Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5), 2011. 22
- Erik M. van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A. Kors, and Laura I. Furlong. The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships . *Journal of Biomedical Informatics*, 45(5), 2012. Text Mining and Natural Language Processing in Pharmacogenomics. 23
- Andreas Vlachos and Stephen Clark. Application-Driven Relation Extraction with Limited Distant Supervision. In *Proceedings of the First AHA!-Workshop on Information Discovery in Text*, Dublin, Ireland, 2014. Association for Computational Linguistics and Dublin City University. 24
- Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wieggers, and Zhiyong Lu. Overview of the BioCreative V Chemical Disease Relation (CDR) Task. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, Sevilla, Spain, 2015. 22
- Dirk Weissenborn, Leonhard Hennig, Feiyu Xu, and Hans Uszkoreit. Multi-Objective Optimization for the Joint Disambiguation of Nouns and Named Entities. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China, 2015. Association for Computational Linguistics. 109



## REFERENCES

---

- Fei Wu and Daniel S. Weld. Autonomously Semantifying Wikipedia. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, New York, USA, 2007. ACM. 26, 28, 59
- Fei Wu and Daniel S Weld. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010. 21
- Feiyu Xu, Hans Uszkoreit, and Hong Li. Task driven coreference resolution for relation extraction. In *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI)*, Patras, Greece, 2008. IOS Press. 21
- Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 28
- Roman Yangarber, Ralph Grishman, and Pasi Tapanainen. Automatic Acquisition of Domain Knowledge for Information Extraction. In *Proceedings of the 18th International Conference on Computational Linguistics*, 2000. 21
- Limin Yao, Sebastian Riedel, and Andrew McCallum. Collective Cross-document Relation Extraction Without Labelled Data. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 20, 21, 28
- Dmitry Zelenko, Chinatsu Aone, Anthony Richardella, Jaz K, Thomas Hofmann, Tomaso Poggio, and John Shawe-taylor. Kernel Methods for Relation Extraction. *Journal of Machine Learning Research*, 3, 2003. 22
- Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015. Association for Computational Linguistics. 20, 27
- Xingxing Zhang, Jianwen Zhang, Junyu Zeng, Jun Yan, Zheng Chen, and Zhifang Sui. Towards Accurate Distant Supervision for Relational Facts Extraction. In *Proceed-*

## REFERENCES

---

- ings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, 2013. Association for Computational Linguistics. 28, 32, 78
- Wu Zheng and Catherine Blake. Using distant supervised learning to identify protein subcellular localizations from full-text scientific articles . *Journal of Biomedical Informatics*, 57, 2015. 13, 67
- GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. Exploring Various Knowledge in Relation Extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. 22
- Fei Zhu and Bairong Shen. Combined SVM-CRFs for Biological Named Entity Recognition with Maximal Bidirectional Squeezing. *PLoS ONE*, 7(6), 06 2012. 21
- Xiaojin Zhu. Semi-supervised learning literature survey. In *University of Wisconsin-Madison, Tech. Rep.*, 2006. 30

# Appendix 1

<b>REL label</b>	<b>Definition</b>
AQ	allowed qualifier
CHD	has child (narrower hierarchical term)
DEL	deleted concept
PAR	has parent (broader hierarchical term)
QB	can be qualifier by
RB	has a broader relationship
RL	has similar or like relationship
RN	has narrower relationship
RO	has relationship other than synonymous, narrower or broader
RQ	related and possibly synonymous
SIB	has sibling
SY	source-asserted synonymy

Table 1: Relationship definitions of MRREL

Semantic Group	Semantic Type (STY)
Chemicals & Drugs	Amino Acid, Peptide, or Protein Antibiotic Biologically Active Substance Biomedical or Dental Material Carbohydrate Chemical Chemical Viewed Functionally Chemical Viewed Structurally Clinical Drug Eicosanoid Element, Ion, or Isotope Enzyme Hazardous or Poisonous Substance Hormone Immunologic Factor Indicator, Reagent, or Diagnostic Aid Inorganic Chemical Lipid Neuroreactive Substance or Biogenic Amine Nucleic Acid, Nucleoside, or Nucleotide Organic Chemical Organophosphorus Compound Pharmacologic Substance Receptor Steroid Vitamin
Disorders	Acquired Abnormality Anatomical Abnormality Cell or Molecular Dysfunction Congenital Abnormality Disease or Syndrome Experimental Model of Disease Finding Injury or Poisoning Mental or Behavioral Dysfunction Neoplastic Process Pathologic Function Sign or Symptom

Table 2: Semantic groups according to Bodenreider and McCray [2003].

# Appendix 2

## Annotation Guidelines

Please read these instructions carefully before starting the annotation:

### Annotation of relationships between medical concepts

The goal of this annotation task is to determine whether two medical concepts express a particular target relationship within the given sentence or not. The annotated sentences will be used to evaluate a relation extraction system.

The annotation task provides sentences with two highlighted medical concepts, one labelled as DRUG and the other as DISEASE. You have to decide whether the two highlighted concepts describe the target relation or not.

There will be two different annotation tasks with two different relations to annotate: may-prevent and may-treat. Each target relation includes 400 sentences, split into 20 sub-tasks with 20 sentences each. If you start a new sub-task, it is recommended that you work through the complete sub-task and finish it, otherwise your results will not be saved in the database. If you are unsure about a sentence just make the best guess.

### Target relation MAY-PREVENT:

This task will focus on the annotation of the relation may-prevent, according the following definition:

You should decide for each given pair of highlighted entities, whether they describe

---

'preventative use or indication of a generic ingredient preparation or drug'

the relation according the definition. The annotation is always either true or false. Some positive and negative examples which will help you to annotate the data set are provided below.

The two medical concepts are highlighted with brackets and enhanced with the concepts **DRUG** or **DISEASE** in capital letters. The sentences are taken from medical abstracts and processed with different scripts. Therefore, sometimes the format might look incorrect. Furthermore, since the detection of the two entities was carried out with a program, it might contain errors. Highlighted entities which do not make sense should be annotated as false. If you spot an error (e.g. entities are highlighted in a wrong way), you can write down the number of the sentence and send it to me.

#### **Examples MAY-PREVENT:**

“It is unlikely that [**DRUG:dipyridamole**] leads to a permanent reduction in blood pressure and that this would explain why this drug might prevent [**DISEASE:strokes**] rather than coronary events.” (PMID=12958322)

→ The sentence expresses, that “dipyridamole” may be used to prevent “strokes”. Therefore it is a positive example and should be annotated as “true”.

“CONCLUSION: [**DRUG:Ondansetron**] and ondansetron plus dexamethasone were equally effective in preventing early [**DISEASE:nausea**] and vomiting in children following strabismus surgery.” (PMID=15089065)

→ The annotation should be “true”, since “Ondansetron” and “nausea” express the relation may-prevent within the sentence. It is important to note, that the sentence contains also another drug (“ondansetron”) and another disease (“vomiting”) which also describe the target relation. Since these words are not highlighted, they are not of interest at this point.

“There is no easily identifiable magic bullet for preventing [**DISEASE:caries**] in that

---

age group, but the use of evidence-based preventive interventions (such as [DRUG:fluoride]) should suffice.” (PMID=14739966)

→ “true”

“Etidronate increases the lumbar bone mineral density (BMD), and prevents new vertebral fractures, in patients with [DISEASE:osteoporosis], while alendronate and [DRUG:risedronate] increase the lumbar and femoral neck BMDs, and prevent new vertebral and femoral neck fractures.” (PMID=14584089)

→ the prevention of “osteoporosis” refers to “Etidronate” and NOT to the highlighted drug “risedronate”, therefore the annotation should be “false”

“A meta-analysis of randomised, placebo-controlled trials for the secondary prevention of [DISEASE:seizures] after alcohol withdrawal showed lorazepam to be effective, whereas [DRUG:phenytoin] was ineffective.” (PMID=14594442)

→ “false”

“An open and multicentric study was conducted with 66 patients with mild to severe diastolic arterial hypertension and echocardiographic [DISEASE:left ventricular hypertrophy], the evolution of diastolic function, by means of doppler transmitral flow echocardiography, under treatment with [DRUG:ramipril], an angiotensin converting enzyme inhibitor, at a dose of 2.5 and 5 mg/day, or combined with a diuretic, after three and six months of treatment.” (PMID=9580230)

→ “false”

“The efficacy of [DRUG:nedocromil sodium] (NED) (8mg twice daily in controlling the clinical symptoms of [DISEASE:asthma] (score symptoms), the pulmonary parameters (FEV1, FVC) and bronchial hyperreactivity to histamine was assessed.” (PMID=9489432)

→ “false”

“The [DISEASE:caries] resistance concept was shown to be erroneous 25 years ago, but the new paradigm is not yet fully adopted in public health dentistry, so we still await real breakthroughs in more effective use of [DRUG:fluorides] for caries prevention.”

---

(PMID=15153687)

→ “false”. The sentence indirectly expresses, that “fluorides” can be used to prevent “caries” - even if it is not the best method, but the word “fluorides” refers to another “caries” in the sentence. The highlighted “caries” expresses something else. For this reason the annotation should be “false”. It is important that you just consider the two highlighted words to determine the relation.

“Efficacy and safety of [DRUG:UFH] and enoxaparin are similar for the treatment [DISEASE:of deep vein thrombosis].” (PMID=15151480)

→ “false”. The highlighted entities in the sentence express a treatment and not a preventive usage. First of all, the may-treat relation will be annotated in another task and second, usually a preventive usage will be applied before a disease occurs.

### **Target relation MAY-TREAT:**

This task will focus on the annotation of the relation may-treat, according the following definition:

```
`therapeutic use or indication of a generic ingredient  
preparation or drug'
```

You should decide for each given pair of highlighted entities, whether they describe the relation according the definition. The annotation is always either true or false. Some positive and negative examples which will help you to annotate the data set are provided below.

### **Examples MAY-TREAT:**

“CONCLUSIONS: [DRUG:Tamoxifen] is safe and effective for the treatment of [DISEASE:gynecomastia].” (PMID=18357357)

→ “true”



---

“[**DRUG:Ivermectin**] is also affective in the treatment of [**DISEASE:ascariasis**] and cutaneous larva migrans.” (PMID=15318139)

→ “true”

“Comparison was made between pregnancies with severe [**DISEASE:lupus**] requiring [**DRUG:cyclophosphamide**] and those that did not.” (PMID=16175930)

→ “false”

“The frequency of [**DISEASE:eczema**] ( $p < 0.005$ ) and exhaled [**DRUG:nitric oxide**] levels ( $p < 0.001$ ) were higher among atopic patients.” (PMID=16683051)

→ false

“There is no easily identifiable magic bullet for preventing [**DISEASE:caries**] in that age group, but the use of evidence - based preventive interventions (such as [**DRUG:fluoride**]) should suffice.” (PMID=14739966)

→ false

## Further Instructions

Please annotate the two highlighted words as “true” only if the sentences express that the two highlighted concepts (and only those) express the relationship.

Does the sentence clearly expresses that a substance can be used to prevent or treat a disease or does the sentences just express whether it will be examined?

Even if the sentence expresses a prevention or a treatment of a disease, is this relation expressed by the two highlighted entities?

“The results show that mixing *X1* and *X2* can result in temporary blindness, while the intake of [**DRUG:X1**] has been successfully tested to prevent [**DISEASE:Y**].”

→ this case would be correct, since *X1* refers to *Y* and expresses the may-prevent relation, but:

---

“The results show that mixing [**DRUG:X1**] and *X2* can result in temporary blindness, while the intake of *X1* has been successfully tested to prevent [**DISEASE:Y**].”

→ In this case *X1* does not refer with a may-prevent relation to *Y*. The highlighted *X1* refers to “temporary blindness” and would be more a kind a side-effect

Sometimes, sentences just express, that they examine something, but it is not a fact, that it works: e.g. a title of a publication might be: [**DRUG:X1**] for the treatment of [**DISEASE:Y**].

Maybe the terms **DRUG** and **DISEASE** might be a bit confusing, depending on the context. **DRUG** can be for example concepts such as “Pharmacologic Substance”, “Clinical Drug”, “Organic Chemical” and **DISEASE** for example “Disease or Syndrome”, “Sign or Symptom”.