

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE INFORMÁTICA
Departamento de Ingeniería del Software e Inteligencia Artificial



TESIS DOCTORAL

**Mejorando la extracción automática de relaciones biomédicas
usando diferentes características lingüísticas de los textos**

**Enhancing automatic extration of biomedical relations using
different linguistic features extracted from text**

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Behrouz Bokharaeian

Director

Alberto Díaz Esteban.

Madrid, 2018

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMÁTICA



TESIS DOCTORAL

**Mejorando la extracción automática de relaciones biomédicas
usando diferentes características lingüísticas de los textos**

**Enhancing automatic extraction of biomedical relations using
different linguistic features extracted from text**

MEMORIA PARA OPTAR AL GRADO DE DOCTOR
PRESENTADA POR

Behrouz Bokharaeian

Director:

Dr. Alberto Díaz Esteban

Madrid, 2017

**Enhancing automatic extraction of biomedical relations using
different linguistic features extracted from text**



UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMÁTICA

Departamento de Ingeniería del Software e Inteligencia Artificial

*Requirements for the Ph.D. degree
presented by*

Behrouz Bokharaeian

Supervised by Professor

Alberto Díaz Esteban

**Programa de Doctorado en Ingeniería Informática
Junio 2017**

Acknowledgments

I would like to express my special appreciation and thanks to my advisor Profesor Dr. Alberto Diaz, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been priceless.

A special thanks to my family. Words cannot express how grateful I am to my mother, Fatemeh Mehdikhanian for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far. This one is for you mom!

Thank you to everyone who assisted with the study. Mariana Neves and Miguel Ballesteros were especially helpful for using related tools and corpora and Hamidreza Chitsaz and MT Pilehvar provided me with valuable advice on performing the experiments.

I would also like to thank all of my friends who supported me in writing, and incited me to strive towards my goal.

Table of Contents

LIST OF ABBREVIATIONS	5
LIST OF FIGURES	7
LIST OF TABLES	8
ABSTRACT	9
RESUMEN	11
I CONTENTS OF THE THESIS	14
1 INTRODUCTION	17
1.1 Motivation	17
1.2 This Thesis	19
1.2.1 Objectives.....	19
1.2.1.1 Improving the Performance of Methods of DDI Extraction from Text through Detecting Linguistic-Based Negation	19
1.2.1.2 Improving the Performance of DDI Extraction Methods from Text through Detecting and Discriminating Between Different Clauses	20
1.2.1.3 Preparing a Corpus For Extracting SNP-Phenotype Association from Text, Annotated With Negation, Modality and Ranked Associations	20
1.2.1.4 Developing A Method for Extracting Graded SNP-Phenotype Associations from Text through Recognizing Degree of Confidence and Negation.....	20
1.2.2 Thesis Structure.....	21
2 PRELIMINARIES.....	25
2.1 Biomedical Relations.....	25
2.2 Related Corpora.....	27
2.3 Biomedical Relation Extraction Methods.....	29
2.3.1 Amount of the Labeled Training Data.....	29
2.3.2 Kernel Based Methods.....	30
2.3.2.1 Sequence Kernels	30
2.3.2.2 Tree Kernels	31
2.3.2.3 Graph Kernels.....	32

2.3.2.4	Other Types of Kernels	33
2.4	Related NLP Tasks	33
2.4.1	Negation Detection Methods.....	33
2.4.2	Clause Dependency Detection in Relation Extraction.....	34
2.4.3	Level of Confidence and Neutral Candidate Detection in Relation Extraction	35
3	ENHANCING AUTOMATIC EXTRACTION OF BIOMEDICAL RELATIONS USING DIFFERENT LINGUISTIC FEATURES EXTRACTED FROM TEXT	39
3.1	Improving the Performance of Methods of the DDI Extraction from Text through Detecting Linguistic-Based Negation	39
3.1.1	Annotating the Drug-DDI Corpus with Negation	39
3.1.2	Proposing the Neutral Candidates Features and Linguistic-based Negation.....	41
3.2	Improving the Performance of DDI Extraction Methods from Text through Detecting and Discriminating Between Different Clauses	42
3.2.1	Enhancing the DDI Supervised Extraction Methods Using Clause Dependency Features.....	42
3.2.2	Proposing a New DDI Extraction Method through Combining Tree and Sequence Kernels	43
3.3	Preparing a Corpus For Extracting SNP-Phenotype Association from Text, Annotated With Negation, Modality and Ranked Associations	44
3.3.1	Producing the Ranked SNP-Phenotype (SNPPhenA) Association Extraction Corpus	44
3.3.2	Developing a Website for the SNPPhenA Corpus	46
3.4	Developing a Method for Extracting Graded SNP-Phenotype Associations from Text through Degree of Confidence and Negation.....	47
3.4.1	Suggesting the Criteria for Reliability of the SNP-Phenotype Association Extraction Method.....	48
3.4.2	Proposing a New Method for Extraction of SNP-Phenotype Association and Degree of Confidence of Association through Linguistic-Based Negation.....	48
3.4.3	Developing an Online Web Based Portal for Extracting SNP-Phenotypes from Text.....	49
4	CONCLUSIONS AND FUTURE WORKS	53
4.1	Conclusions	53
4.1.1	Improving the Performance of Methods of the DDI Extraction from Text through Detecting Linguistic-Based Negation	53
4.1.2	Improving the Performance of the DDI Extraction Methods from Text through Detecting and Discriminating between Different Clauses	54
4.1.3	Preparing a Corpus For Extracting SNP-Phenotype Association from Text, Annotated With Negation, Modality and Ranked Associations	54

4.1.4	Developing a Method to Extract Graded SNP-Phenotype Associations from Text through Recognizing Degree of Confidence and Negation	55
4.2	Future Works	56
4.2.1	DDI Extraction.....	56
4.2.2	SNP-Phenotype Association Extraction.....	56
	BIBLIOGRAPHY	58
	II PUBLICATIONS	67
	Extracting Drug-Drug Interaction From Text Using Negation Features	66
	Exploring Negation Annotations in the DrugDDI Corpus	77
	Extracting Drug-Drug Interactions From Text Through Combination of Sequence and Tree Kernels	88
	Enhancing DDI Extraction Using Neutral Candidates, Negation, and Clause Dependency	98
	SNPPhenA: A Corpus For Extracting SNP-Phenotype Associations From Literature	121
	Extraction of Ranked SNP-Phenotype Associations from Text Using Neural Candidates, Negation And Modality	137
	III APPENDICES	164
	APPENDICES	151
	Appendix 1: SNPPhenA Corpus Guidelines	153
	Appendix 2: Snapshots of the SNPPhenA Corpus in XML and Brat Formats	167
	Appendix 3: Snapshots of the SNPPhenA Website	173
	Appendix 4: Kappa Calculation for Analyzing the Reliability of the SNPPhenA Corpus	176

List of Abbreviations

BioNLP	Biomedical Natural Language Processing
BOW	Bag of Words
CL	Confident Level
CLA	Clause connector
DC	Dependent Clause
DNA	Dexoxribonucleic Acid
DDI	Drug-Drug Interaction
FM	F-Measure
FN	False negative
FP	False positive
GC	Global context
GCK	Global context kernel
GWA	Genome Wide Association
GWAS	Genome Wide Association Study
IDC	Independent Clause
IE	Information extraction
LC	Local context
LCK	Local context kernel
LOC	Localization (biological event)
MFC	Minimum Frequency of the Case MMF Minimum Matching Feature
NEG	Negative Regulation (biological event)
NER	Named Entity Recognition
NEU	Neutral candidate
NegSc	Negation scope
NLP	Natural Language Processing
P	Precision
Num	Number
POS	Part-of-Speech
POSSTEM	Part-Of-Speech Stem (Conjunction feature)
POSLEMMA	Part-Of-Speech Lemmatization (Conjunction feature)
PPI	Protein-protein interaction
Phent	Phenotype trait
RNA	Ribonucleic Acid
SNP	Single nucleotide polymorphism
SNPPhenA	SNP-Phenotype Association
SVM	Support vector machine
SubTK	Subtree kernel
SubSetTK	Subset tree kernel
D1 D2	(drug interaction arguments)
ST, ST2	(biological event's arguments)
TM	Text mining
TP	True positive
TRA	Transcription (biological event)

UIMA Unstructured Information Management Architecture

WWW World Wide Web

List of Figures

Figure 1: Number of SNP publications from 2000 to 2014 in PubMed.....	21
Figure 2: A sample of a sentence with a gene-disease relation.....	26
Figure 3: Two samples of sentences with disease-treatment association [21]	26
Figure 4: The unified XML format of a sentence in the DrugBank-DDI 2013 corpus	28
Figure 5 : A conceptual comparison between supervised and semi-supervised methods	30
Figure 6: Two samples of sequence kernels (local and global context) for a gen-disease relation.....	31
Figure 7: Subtrees of constituent parse tree of a sample sentence	32
Figure 8: List of subset trees of a constituent parse tree	32
Figure 9: Sentences as sample of directional dependency graphs	32
Figure 10: Graph representation generated from an example sentence with the candidate interaction pair is marked as PROT1 and PROT2	33
Figure 11: A sample of sentence annotated with negation scope and cue in the NegDDI corpus	40
Figure 12: The unified XML format of a sentence in the SNPPhenA corpus.....	45
Figure 13: The annotation of the sample sentences in brat format (*.ann)	45
Figure 14: A sample of a sentence in the produced SNPPhenA corpus drawn by brat.....	45
Figure 15 : A screenshot of the produced website corpus for the SNPPhenA	47
Figure 16 : A screenshot of the web based ranked association extractor.....	49
Figure 17 : A screenshot of the result of the web based ranked association extractor.....	50

List of Tables

Table 1: Statistics of the training and test datasets of the DDI-DrugBank 2013 corpus	27
---	----

Abstract

Extracting biomedical relations from texts is a relatively new, but rapidly growing research field in natural language processing (NLP). Due to the increasing number of biomedical research publications and the key role of databases of biomedical relations in biological and medical research, extracting biomedical relations from scientific articles and text resources is of utmost importance.

Drug-drug interactions (DDI) are, in particular, a widespread concern in medicine, and thus, extracting this kind of interactions automatically from texts is of high demand in BioNLP. A drug-drug interaction usually occurs when one drug alters the activity level of another drug. According to the reports prepared by the U. S. Food and Drug Administration (the FDA) and other acknowledged studies [1], over 2 million life-threatening DDIs occur in the United States every year. Many academic researchers and pharmaceutical companies have developed relational and structural databases, where DDIs are recorded. Nevertheless, most up-to-date and valuable information is still found only in unstructured research text documents, including scientific publications and technical reports.

In this thesis, three complementary, linguistically driven, feature sets, are studied: negation, clause dependency, and neutral candidates. The ultimate aim of this research is to enhance the performance of the DDI extraction task by considering the combinations of the extracted features with well-established kernel methods.

Our experiments indicate that the proposed features significantly improve the performance of the relation extraction task. We also characterize the contribution of each category of features and finally conclude that neutral candidate features have the most prominent role among all of the three categories.

Another biomedical relation studied is the association between Single Nucleotide Polymorphisms (SNPs) and Phenotypes (SNPPhenA). SNPs are referred to the most significant genetic changes contributing to common diseases. A SNP is a DNA sequence variation commonly occurring within a population with a single nucleotide — A, T, C, or G — in the genome varying between members of a biological species. The huge number of identified SNP-phenotype associations, implies the necessity of developing an automatic association extraction tool.

In this thesis, a corpus for extracting ranked associations of SNP and Phenotypes has been developed. It is the first relation extraction corpus annotated with degree of confidence, showing the strength of associations. The process of producing the corpus includes collecting abstracts, recognizing named entities, and annotating the ranked association,

negation scope and cues as well as modality markers. In addition, the confident level of positive association was annotated in three categories: strong, moderate and weak degree of confidence. The corpus has been generated in two formats: xml and standoff BRAT formats and a website has been enabled with all the relevant information.

Finally, a supervised method to extract SNP-Phenotypes association has been developed. The relation extraction method relied on linguistic-based negation detection and neutral candidates. The experiments have shown that negation detection as well as detecting neutral candidates can be employed to implement a superior relation extraction method outperforming the kernel-based counterparts. These results are mainly due to a uniform innate polarity of sentences and a small number of complex sentences in the corpus. Furthermore, we implemented a novel modality-based supervised method (MMS) to identify the level of confidence of the extracted association.

Resumen

La extracción de relaciones entre entidades es una tarea muy importante dentro del procesamiento de textos biomédicos. Cada vez hay más información sobre este tipo de interacciones almacenada en bases de datos, pero sin embargo la mayor cantidad de información relacionada con el tema está presente en artículos científicos o en recursos donde la información se almacena en formato textual.

Las interacciones entre fármacos son, en particular, una preocupación generalizada en medicina, por esa razón la extracción automática de este tipo de relaciones es una tarea muy demandada en el procesamiento de textos biomédicos. Una interacción entre 2 fármacos normalmente se produce cuando un fármaco altera el nivel de actividad de otro fármaco. De acuerdo a los informes presentados por la Administración Nacional de Alimentos y Fármacos de Estados Unidos y otros estudios reconocidos [1], cada año se producen más de 2 millones de interacciones mortales entre fármacos. Muchos investigadores y compañías farmacéuticas han desarrollado bases de datos donde estas interacciones son almacenadas. Sin embargo, la información más actualizada y valiosa sigue apareciendo sólo en documentos no estructurados en formato textual, incluyendo publicaciones científicas e informes técnicos.

En esta tesis se estudian 3 conjuntos de características lingüísticas de los textos: negación, dependencia clausal y candidatos neutros. El objetivo final de la investigación es mejorar el rendimiento de la tarea de extracción de interacciones entre fármacos considerando las combinaciones de las características lingüísticas extraídas de los textos con métodos de aprendizaje basados en kernel.

Nuestros experimentos indican que las características propuestas mejoran la tarea de extracción de relaciones de manera significativa. También se han caracterizado la contribución de cada una de las características por separado, lo que ha llevado a la conclusión de que los candidatos neutros juegan el papel más importante dentro de las 3 categorías.

Otra relación biomédica que ha sido estudiada es la asociación entre Polimorfismos de Nucleótido Simple (SNP) y Fenotipos (SNPPhenA). Los SNPs son considerados como los cambios genéticos más significativos que contribuyen a enfermedades comunes. Un SNP es una variación en la secuencia de ADN que afecta un nucleótido simple – A, T, C o G – de una secuencia del genoma y que varía dentro de una población significativa entre miembros de una especie biológica. El elevado número de asociaciones entre SNPs y fenotipos implica la necesidad del desarrollo de una herramienta de extracción automática de estas asociaciones.

En esta tesis se ha desarrollado un corpus para la extracción de asociaciones entre SNPs y fenotipos. Es el primer corpus anotado con el grado de confianza de la relación. El proceso de generación del corpus incluye la recopilación de resúmenes de artículos, reconocimiento de entidades, anotación de la asociación con su grado de confianza, así como anotación de negaciones y marcadores modales. La anotación del grado de confianza de las asociaciones positivas ha sido realizada en 3 niveles: fuerte, moderada y débil. El corpus ha sido generado en 2 formatos: xml y formato standoff para BRAT. También se ha habilitado un sitio web con toda la información relevante.

Por último, se ha desarrollado un método supervisado para la extracción de asociaciones entre SNPs y Fenotipos que utiliza la información asociada a la detección de la negación y la presencia de candidatos neutros. Los experimentos han mostrado que la detección de la negación y la detección de candidatos neutros pueden ser utilizadas para desarrollar un método mejor que los basados en kernel tradicionales. Estos resultados son debidos, principalmente, a la polaridad intrínseca de la mayoría de las sentencias del corpus, así como al pequeño número de sentencias complejas. Además, se ha implementado un método supervisado basado en modalidad para identificar el nivel de confianza de las asociaciones extraídas.

Part I

Contents of the thesis

If there is a book that you want to read, but it hasn't been written yet, you must be the one to write it.

Toni Morrison

Chapter

Introduction

1 Introduction

This chapter presents an overview of the tasks and problems we intend to solve in addition to the current status of the field where this thesis is placed. We start our work by presenting the motivations for the thesis, in which the emerging needs for the successful implementations and practices to extract Drug-Drug interactions and SNP-Phenotype associations from text using linguistic features are highlighted.

Due to the two (different, but completely connected) lines of research followed in the development of this thesis, we continue by introducing some basic notions on extracting Drug-Drug interactions and SNP-Phenotype associations from text in addition to some related linguistic features that we have applied to improve the performance of the tasks.

The chapter ends with a summary of the genesis of this thesis followed with a detailed discussion of our objectives. Finally, we conclude with the structure of the rest of the thesis.

1.1 Motivation

Unstructured text documents such as articles are the major source of knowledge in biomedical fields, and millions of biomedical papers are published every year. The MEDLINE 2015 database contains over 22 million records, and the database is currently growing at the rate of 500,000 new records every year. With this huge amount of information, staying up to date is very difficult. On the other hand, traditional keyword and indexing search methods cannot satisfy researches, and so there is a need for new knowledge discovery and text mining tools in this area. Extracting biomedical relations from text is a relatively new but fast growing research field in Natural Language Processing (NLP) [1]. Because of the increasing number of biomedical researches and the huge number of biomedical unstructured text resources, extracting biomedical relations from scientific articles and text reports is a highly demanding task. Formally, relation extraction is the task of finding semantic relations between entities from text.

Several supervised methods have been developed to extract biomedical relations from text, however, these methods do not consider all the linguistic information available in the text. One of the contributions of this thesis is to propose some new linguistic features to improve the extraction of biomedical relations from text: negation detection [2], clause identification [3] in the sentences and considering the degree of confidence of relations and modality in sentences [4] [5].

Although there are several relation extraction tasks, this thesis is concentrated in two types of them: Drug-Drug Interactions and SNP-Phenotype associations.

Drug-drug interaction (DDI) is, in particular, a widespread concern in medicine, and thus, extracting this kind of interaction automatically from texts is of high demand in BioNLP. Drug-drug interaction usually occurs when one drug alters the activity level of another drug. According to the reports prepared by the Food and Drug Administration (the FDA) and other acknowledged studies [6], over 2 million life-threatening DDIs occur in the United States

every year. Many academic researchers and pharmaceutical companies have developed relational and structural databases, where DDIs are recorded. Nevertheless, most up-to-date and valuable information is still found only in unstructured research text documents, including scientific publications and technical reports.

In this thesis, we first introduce the basics of three complementary, linguistically driven feature sets: negation, clause dependency, and neutral candidates. The ultimate aim of this research is to enhance the performance of the DDI extraction task by considering and employing the above-mentioned three operations and feature sets.

First, it is essential to detect negative assertions in most biomedical text-mining tasks, where the overall purpose is to derive factual knowledge from textual data. According to Loos et al. [2], negation is a morphosyntactic operation in which a lexical item denies or inverts the meaning of another lexical item or construction. Negation is commonly utilized in biomedical articles and is an important origin of low precision in automated information retrieval systems [8].

Second, identifying the role of clause dependency in complex sentences in DDI detection is another linguistically driven subject investigated in this research. According to Harris and Rowan [3], a dependent clause is a group of words with a subject and a verb that do not express a complete thought, cannot stand alone, and usually extend the main clause. An independent clause, or main clause, is one that can stand alone as a sentence and express a complete thought.

Finally, we study the role of neutral DDI candidates in the relation extraction. Most of the current relation extraction problems and the produced corpora are based on binary relations, that is, they decide a binary relation between two entities. Although detecting DDI interactions is the main target of a DDI task, there is a difference between a negative interaction candidate having been stated by the authors (distinguished candidate) and that which has not (neutral candidate). Both of these candidates are considered negative in the DrugDDI corpus.

The other biomedical relation studied is the association between Single Nucleotide Polymorphisms and Phenotypes (SNPPhenA). SNPs are referred to the most significant genetic changes contributing to common diseases. A SNP is a DNA sequence variation commonly occurring within a population with a single nucleotide — A, T, C, or G — in the genome varying between members of a biological species. Generally, there are two types of mutation that work in protein-level or DNA-level. An SNP is a single base mutation occurring in DNA-level, and variations in the DNA sequences can influence how humans develop diseases and respond to pathogens, drugs, and other agents. SNPs are also important for personalized medicine [7].

1.2 This Thesis

1.2.1 Objectives

This thesis can be included in the general study of biomedical relation extractions from literature. More specifically, it was concentrated on DDI and SNP-phenotype association as important relations, which are of high interest in biomedical researches. Although several methods have been developed for extracting DDIs and limited studies have been developed for exploring the association between mutations and phenotypes from text, none has comprehensively considered negation, clause dependency, degree of confidence of relations, and modalities markers.

The objectives of this thesis are mentioned in the rest of this section:

1.2.1.1 Improving the Performance of Methods of DDI Extraction from Text through Detecting Linguistic-Based Negation

Following the research on negation detection in the NIL group [9] and the results achieved by Chowdhury [10] in DrugDDI (2011) who employed the negation in a limited form, we aimed to study the negation more thoroughly. We started our research annotating the DDI corpus with negation scope and cues. Moreover, a superior method was proposed through employing the negation annotations.

During the analysis of the key factors affecting the use of negation, the importance of neutral candidates was realized, which has not been discussed in the literature so far.

The neutral interaction candidate in fact is a co-mention of two drugs with no remarks by the author in the sentence or the discussed clause, while the distinguished interaction candidate is exactly the opposite (with remarks by the author). Actually, neutral candidates are a particular subclass of non-positive candidates whose lack of interaction cannot be exactly determined by a confident level above zero. For instance, consider the following sentence:

- Studies in healthy volunteers have shown that acarbose has no effect on either the pharmacokinetics or pharmacodynamics of digoxin, nifedipine, propranolol, or ranitidine.

There is no remark by the author about the interaction between propranolol and ranitidine. Therefore, we define this candidate of drug-drug interaction as a neutral candidate.

Since the neutral candidates have not yet been studied in relation extraction task, it was decided to separate them from other candidates by developing a rule-based system in the task of DDI extraction.

1.2.1.2 Improving the Performance of DDI Extraction Methods from Text through Detecting and Discriminating Between Different Clauses

The second major pathway of the present research came naturally when the errors of the developed system were analyzed in DDI extraction challenge 2013.

Errors analysis showed that a large number of errors of the system happened in complex sentences. A complex sentence has one independent clause and at least one dependent clause. Moreover, a clause connector is a word that joins clauses in order to form complex sentences. Coordinators, conjunctive adverbs, and subordinators are three types of connectors.

Since the existing methods of simplifying had poor results, [11], a method was developed to obtain different types of clauses. Likewise, detecting different clauses can improve solving the problem of negation. Negation in independent clauses may not change the status of a DDI interaction. Hence, determining independent and dependent clauses is another effective factor that must be taken into account.

1.2.1.3 Preparing a Corpus For Extracting SNP-Phenotype Association from Text, Annotated With Negation, Modality and Ranked Associations

The third major pathway of the research came naturally in the study of the neutral candidates, the confidence degree of an extracted relationship was found an important factor that has been ignored. This came into mind from the concept of the development of neutral candidates because they had the confidence degree of zero, and negation and modality of markers did not change their status. However, the confidence degree of a relation can determine the strength of a relationship or the intensity of an interaction that can contain useful information.

On the other hand, although there are many biomedical relation extraction methods and corpora, none of them consider the degree of confidence or intensity of the relation. Expanding the available corpora of extraction of biomedical relations with the degree of confidence may not be an appropriate option since in most, because of the variety of sentences, there was not sufficient agreement on the reliable concept and development. Furthermore, detecting the level of confidence of a SNP-phenotype association is important because it can help to identify the strength of the association, which can be used by genetic experts to determine the phenotypic plasticity and the importance of environmental factors.

1.2.1.4 Developing A Method for Extracting Graded SNP-Phenotype Associations from Text through Recognizing Degree of Confidence and Negation

An SNP can be “associated” with the phenotype, when a particular type of variant is frequent within samples obtained from the subjects. The first successful genome-wide

association study dates back to 2005 [12] and it was the start of a worldwide trend which results in finding thousands SNP associations. Figure 1 shows the increasing numbers of papers published in this field from 2001 to 2014, obtained from a PubMed search engine for the query ‘Single Nucleotide Polymorphisms’ (performed in November 2015) [13].

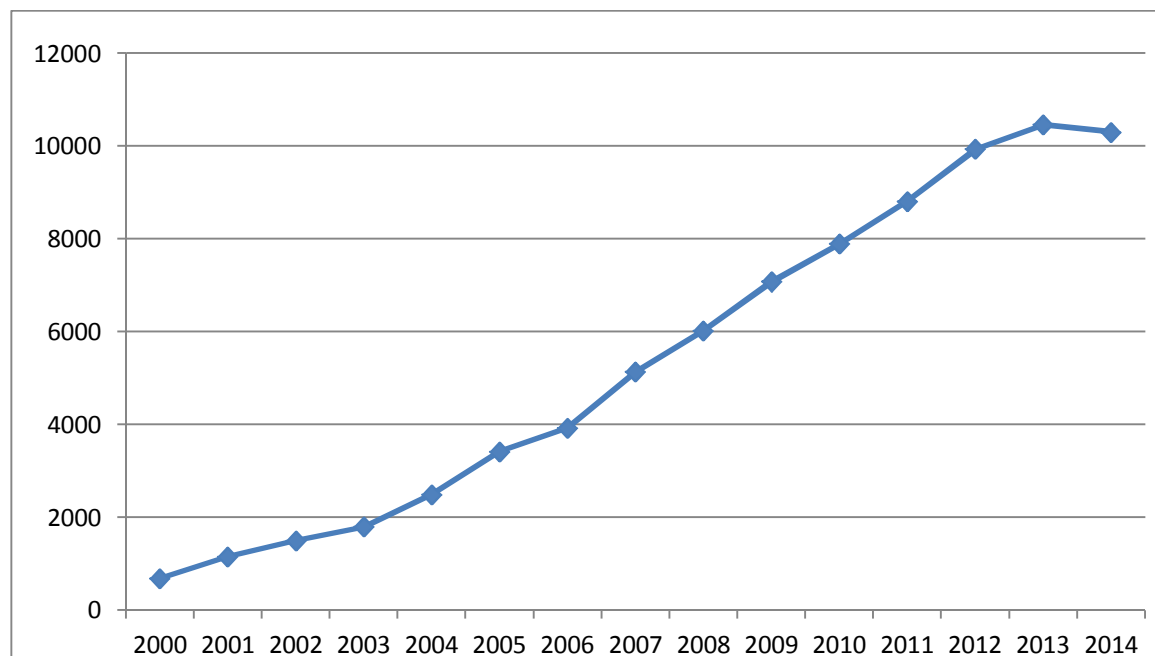


Figure 1: Number of ‘Single Nucleotide Polymorphisms’ publications from 2000 to 2014 in PubMed.

The huge number of identified SNP-phenotype associations implies the necessity of developing an automatic association extraction tool.

Similar to the DDI corpus, the produced SNP-phenotype association corpus was aimed to be annotated with linguistic-based negation scope and cues. With the aid of the produced annotations in the SNPPhenA corpus, a method was also developed to extract the graded relationship that obtain acceptable results, being explained in a paper published in the JAIDM Journal.

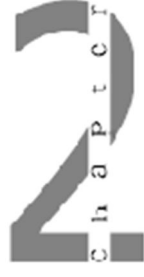
1.2.2 Thesis Structure

Chapter 2 begins with an introduction to the task, i.e. classification of biomedical relation extraction methods from a biological point of view, number of entities, usage of prior knowledge, and other factors. It continues with the introduction of supervised, semi-supervised and unsupervised methods, and kernel methods. They are followed by the employed NLP tasks and the related state of art methods. Finally, our contribution to the scientific community is described. In Chapter 4, we briefly conclude and summarize some (close and far) future works. Part II contains five chapters with our main publications. Part III collects some materials not published in papers that can be used for future works and extension of the research. The appendix also includes the materials related to the

SNPPhenA corpus, including the guideline document, Kappa inter-agreement reliability analyses and tests data, and snapshots of the produced formats of the corpus.

You know you've read a good book when you turn the last page
and feel a little as if you have lost a friend.

Paul Sweeney



Preliminaries

2 Preliminaries

The fundamental concepts and methods which have made possible the research described in this thesis are introduced here. Computer science has undergone a quick evolution from its origins. Nowadays, traditional keyword and indexing search methods can not satisfy researches, so there is a critical necessity for new information retrieval and text mining tools [14]. Therefore, extracting biomedical relations from medical text is a rapidly moving forward task, and the study was devoted to the investigation of the supervised methods and different related linguistic features in the course of the task.

To be more precise, the thesis focuses on the rapports between biomedical relation extraction methods and corpora and related NLP tasks and features. This is why we decided to start this chapter with a section devoted to the basic types of the biomedical relation extraction task and its different categorizations. After that, in Sections 2.2 and 2.3, we will present some previous works in the relation extraction methods and related NLP tasks. After this introductory chapter, we will have the necessary tools to present the contributions obtained in this thesis (Section 3).

2.1 Biomedical Relations

There are many types of biomedical relation extraction tasks that can be categorized from a biomedical point of view and using other criteria such as linguistic factors. From a biomedical point of view, several biological and medical entities and their relations have been investigated and are available in the literature [15].

We start the review of the studied biomedical entities with Protein-protein interactions (PPI), which is probably the most popular biomedical relation investigated in the field.

Protein-protein interactions (PPIs) are physical contacts established between two or more proteins. Determining a protein-protein interaction is essential for the investigation of intracellular signaling pathways, modeling of protein complex structures, and understanding various biochemical processes. Probably, extracting protein-protein interactions from text is the most popular biomedical relation extraction problem. Some of the related researches can be found at [16] and [17] and [18]. The following sentence shows an example of two possible protein-protein interactions.

- a) *sNRP1* is secreted by cells as a 90-kDa protein that binds *VEGF(165)*, but not *VEGF(121)*. It inhibits (125)I-*VEGF(165)* binding to endothelial and tumor cells and *VEGF(165)*-induced tyrosine phosphorylation of KDR in endothelial cells.

Finding associations between specific diseases and their relevant genes or proteins is an important biomedical relation task in bioinformatics. A sample of this relation is shown in Figure 2. Some of these types of research have been carried out by [19] and [20].

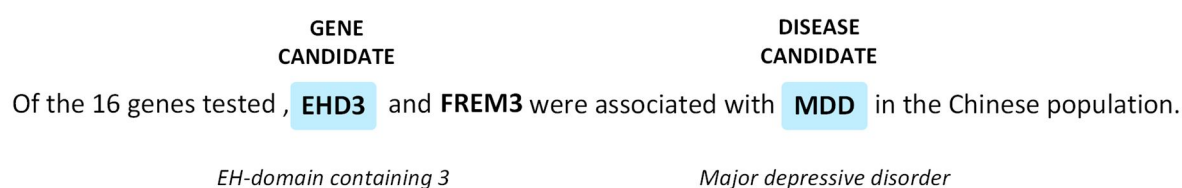


Figure 2: A sample of a sentence with a gene-disease relation

Disease-treatment association extraction is other type of relation extraction that tries to extract different cures for diseases from scientific articles [21] (Figure 3).

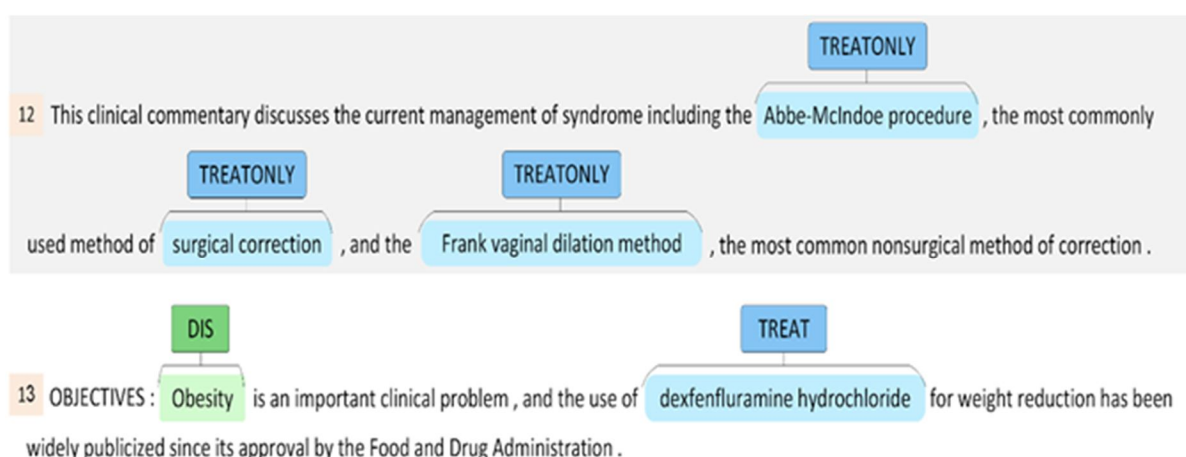


Figure 3: Two samples of sentences with disease-treatment association [22]

Mutation and disease association extraction is the other biomedical relation extraction task recently developed and studied by some researchers [23] [24]. For example, the below sentence mentions the association between mutations (*Q56P*, *P124S*) and *cardio-facio-cutaneous (CFC) syndrome* disease [25]:

- b) Toward this end, the *Q56P* and *P124S* mutations closely mimic *T55P* and *P124L*, respectively—two germline variants observed in patients with *cardio-facio-cutaneous (CFC) syndrome* (15, 16) that confer aberrant MEK activation.

MiRNA-gene [26] and Drug-virus mutation [27] are two other types of biomedical relation studied in the field.

For the purpose of having an estimation of different biomedical relation extractions tasks, 30 related papers published between 2005 and 2013 were randomly selected. According to the current survey, the most studied biomedical relation is protein-protein interaction with 10 papers out of 30 studied papers.

In the rest of this section, other types of categorizations concerning relation extraction task-based on the scale of documents, training knowledge, number of entities and other factors will be explained.

Number of entities exists in the relation and their structure is one of the important factors to be seriously taken into account. Accordingly, the three major categories include:

- **Binary** relation extraction in case of relations with two entities
- **Complex** relation extraction in other cases of relations with more than two entities, such as the research carried out by [1]
- **Hierarchical** relation extraction from text is other type of relation extraction that aims to identify hierarchical relations. One of the methods that investigated ontology-based hierarchical relation extraction system was developed by [28].

Scope of the search text is another basis for categorization. Accordingly, the four major forms of relation extraction tasks are sentence level, paragraph, abstract, and document level relation extraction.

There are other types of categorization such as directional and unidirectional relation and other classifications based on the resolution of the extracted relation including:

A. **Unnamed relations**: this provides the associated biomedical terms but does not specify the actual relation.

B. **Relation class**: this does not specify the relation either but indicates which predefined classes the relation may fall.

C. **Named-relation**: this is the actual relation among terms.

Based on the usage of the prior knowledge, two major categories of relation extractions can be identified: relation extraction with prior knowledge, i.e. domain model, explicit semantic analyses and without prior knowledge, i.e. co-occurrence and kernel methods.

2.2 Related Corpora

As mentioned in the first chapter, this thesis investigates the task of relation extraction in two types of biomedical relations: Drug-Drug interactions and associations between SNP and Phenotype (SNPPhenA). In this section, we provide some backgrounds regarding the related corpora produced previously and the corpora we used during the thesis.

As mentioned earlier, the Drug-Drug Interaction is a vital incident in the medical science and therefore, extracting DDIs from the text is an important task. Accordingly, the first Drug-Drug Interaction extraction corpus [29] was developed by Segura et al. based on a set of 579 xml files describing DDIs, randomly collected from the DrugBank database [30].

	Pairs	Negative	Positive	Effect	Mechanism	Advice	Int
Test	26005	22217	3788	1535	1257	818	178
Train	5265	4381	884	298	278	214	94

Table 1: Statistics of the training and test datasets of the DDI-DrugBank 2013 corpus

The DrugDDI 2013 corpus was developed for the DDI extraction 2013 SemEval task and includes part of the DDI 2011 corpus. Concretely, new documents were annotated from the

DrugBank database and were used for the test dataset (DDI-DrugBank Test 2013 corpus), while 572 documents from the previous corpus were used as training dataset (DDI-DrugBank Train 2013 corpus). Therefore, the DDI-DrugBank 2013 corpus contains a total of 730 documents. A dataset of 233 Medline abstracts (DDI-Medline 2013) was also annotated for the 2013 shared task [31]. Moreover, the DDIs in the DDI corpus 2013 were classified into four types: mechanism, effect, advice and int (Table 1).

```
<sentence id="DDI-DrugBank.d297.s4" text="Concurrent therapy with ORENCIA and TNF
antagonists is not recommended.">
  <entity charOffset="24-30" id="DDI-DrugBank.d297.s4.e0" text="ORENCIA" type="brand"/>
  <entity charOffset="36-50" id="DDI-DrugBank.d297.s4.e1" text="TNF antagonists"
type="group"/>
  <pair ddi="true" e1="DDI-DrugBank.d297.s4.e0" e2="DDI-DrugBank.d297.s4.e1" id="DDI-
DrugBank.d297.s4.p0" type="advise" />
  <negationtags>
    Concurrent therapy with ORENCIA and TNF antagonists is <xcope><cue> not </cue>
recommended </xcope> .
  </negationtags>
</sentence>
```

Figure 4: The unified XML format of a sentence in the DrugBank-DDI 2013 corpus

Comparison of the obtained results for DDI challenge 2011 and 2013 showed a significant improvement in F-score for 2013 teams. According to Segura et al., increasing the size of the corpus and optimizing the quality of annotations contributed to this improvement [31]. It is important to mention that the DDI corpora (2011 and 2013) have been used for the annotation with negation scope and cues. In addition, they have been employed for our experiments on the DDI extraction task.

The rest of this section is dedicated to introducing a number of the few corpora developed with the aim of extracting mutation/polymorphism and disease associations. It is worth mentioning that our contributions in the SNP-phenotype association are based on our produced SNPPhenA corpus. However, to identify the advantages and limitations of previous efforts, we mention three related important corpora: *BRONCO* [25], *Variome* [32] and *EMU* [33]. *BRONCO* contains more than four hundred variants and their associations with genes, diseases, drugs and cell lines in the context of cancer, all extracted from 108 full-text articles. *Variome* covers 12 types of the relations annotated in 10 full-text articles. While, *BRONCO* includes more documents, both corpora annotate several types of relations, such as mutation-disease association, as binary relations on a full-text level. Furthermore, *EMU* corpus contains only gene and disease-related (only breast cancer and prostate cancer) information of each variant.

It is important to mention, although the mentioned corpora are valuable efforts to extract the mutation related information from text, they suffer from lack of annotations for linguistic features.

2.3 Biomedical Relation Extraction Methods

We start this section with a subsection devoted to the basic types of biomedical relation extraction tasks based on the amount of labeled training data. Some of the most innovative and important kernel methods that are most successful and advanced in the field will be presented in Sections 2.3.2.

2.3.1 Amount of the Labeled Training Data

Depending on the amount of the labeled training data and unlabeled training values [34], relation extraction problems can be categorized in three groups:

- **Non-supervised** algorithms without any labeled training data such as co-occurrence method
- **Semi-supervised** methods are methods which work with a small amount of labeled data and a large amount of unlabeled data as training data such as work done by [35]). Four main developed semi-supervised relation extraction systems are: Rationale, DIPRE, Snowball and KnowItAll& TextRunner:
 - KnowItAll [36] is an autonomous domain-independent system that extracts facts from the Web. It also has a set of entity classes to be extracted, such as “city”, “scientist”, “movie”, etc.
 - TextRunner [37] has a self-supervised learner which automatically labels “+/samples” and also has a single-pass extractor which has a single pass over corpus that detects relations in each sentence.
 - Snowball is a similarity based system that has been developed by Agichtein et al [38].
 - DIPRE (Dual Iterative Pattern Relation Expansion) [39] is a technique which exploits the duality between sets of patterns and relations to grow the target relation starting from a small sample.
- **Supervised** relation extraction methods: are those categories of relation extraction methods, i.e. kernel based methods that work with completely labeled training data. A conceptual comparison between supervised and semi-supervised classification method can be seen in Figure 5.

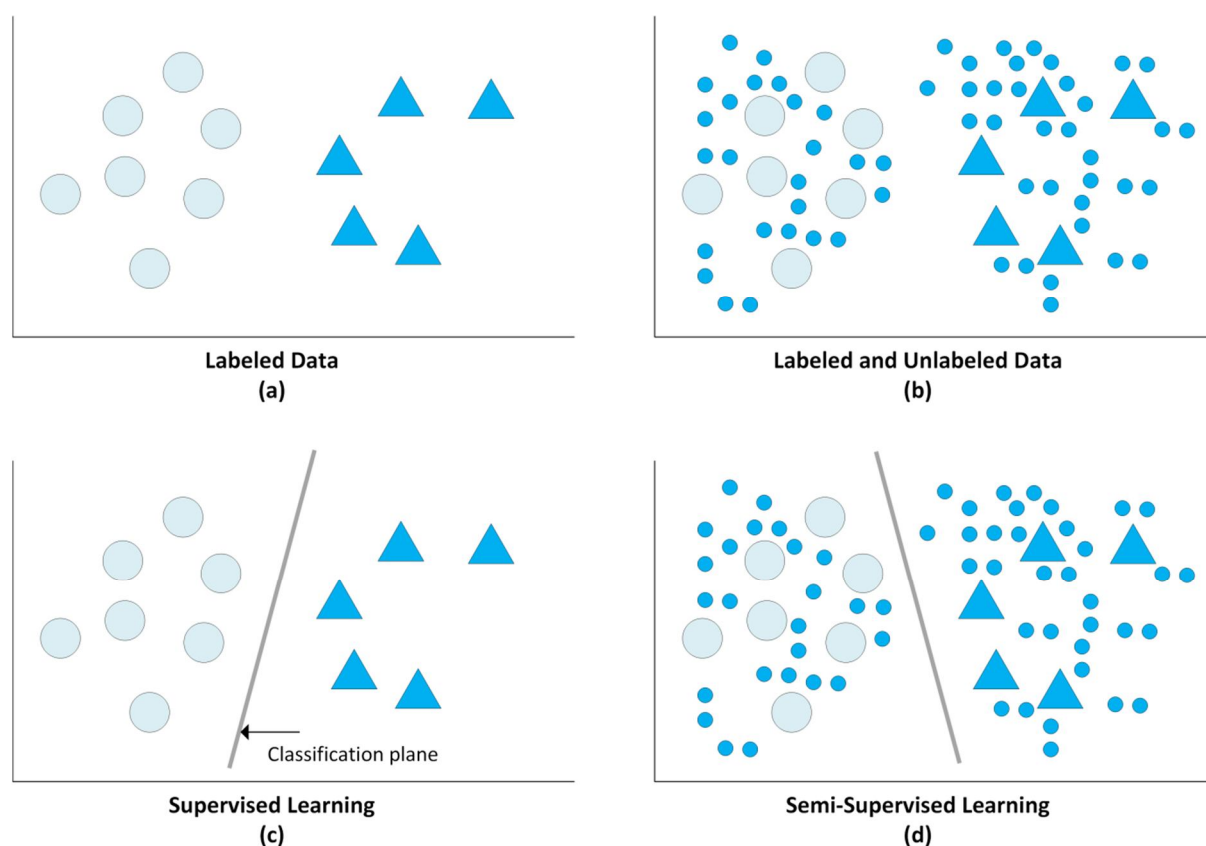


Figure 5 : A conceptual comparison between supervised and semi-supervised classification methods

2.3.2 Kernel Based Methods

Many methods have been developed for supervised relation extraction, but kernel methods are the most popular and successful methods [40]. Generally, in many cases, data cannot be easily expressed via features. For example, in most NLP problems, feature based representations produce inherently local representations of objects, for it is computationally infeasible to generate features involving long-range dependencies. Kernel methods are an attractive alternative to feature-based methods and are most popular as the main classifier for extracting semantic relations from text documents. The major categories of kernel methods have been provided below:

2.3.2.1 Sequence Kernels

Sequence kernels are the primary invented category of relation extraction kernel methods that consider the input text as a sequence of tokens. As an important sample, global context kernel is a sequence kernel with a feature set based on the words occurring in the sentence, fore-between, between, and between-after relative to the extracted pair of entities [41]. Consequently, three term frequency vectors are produced by mean of a bag-of-words model. The global context kernel is then computed as the sum of common words in the three vectors

(Figure 6). Another noted sequence kernel is local context kernel that uses surface (capitalization, punctuation, and numerals) and shallow linguistic (POS-tag, lemma) features generated from the tokens that are left and right of entities of the candidate relation, and the size of the window can be adjusted [41].

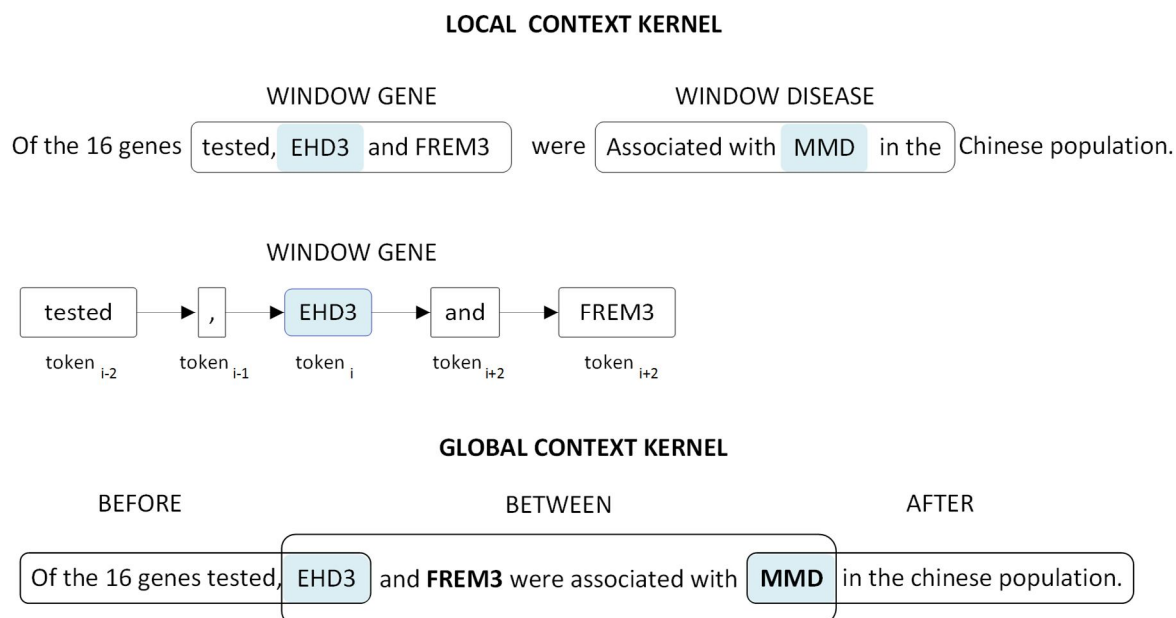


Figure 6: Two samples of sequence kernels (local context kernel and global context kernel) for a gen-disease relation

2.3.2.2 Tree Kernels

Tree kernels are the other category of kernels based on a parse tree and produced by natural language parsers. The key idea of a tree kernel is computing the number of the common substructures between the two trees T_1 and T_2 without explicitly considering the whole fragment space.

There are different types of tree kernels:

- A Subtree kernel considers all common subtrees in the syntax tree representation of two desired sentences (Figure 7).
- Another well-known tree kernel is subset tree kernel, which considers all descendants, including leaves included in the substructures (Figure 8).
- The spectrum tree kernel is another tree kernel counting all common vertex-walks, and sequences of edge-connected syntax tree nodes of a specified length in two sequences [42].
- Unlexicalized Partial Tree Kernel (uPTK) was firstly proposed in [43] and experimented with semantic role labeling (SRL). The results showed no improvement for such task, but it is well known that in SRL, lexical information is essential.

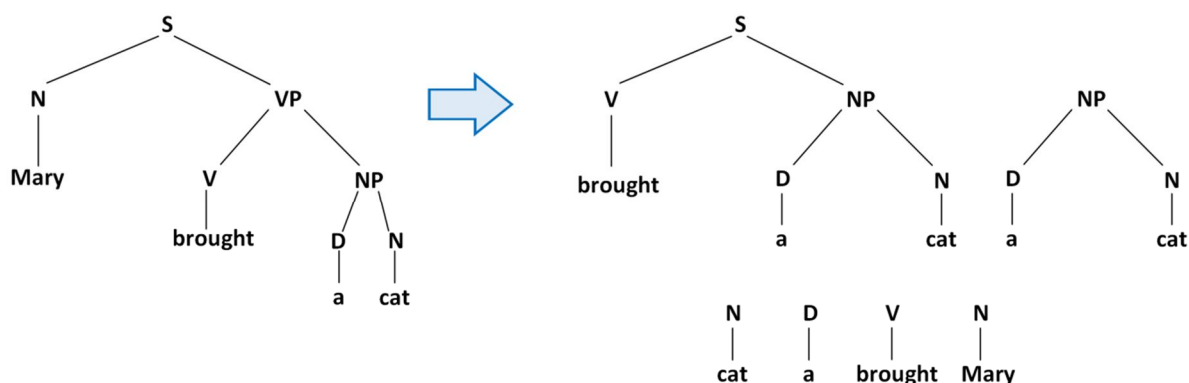


Figure 7: Subtrees of constituent parse tree of a sample sentence

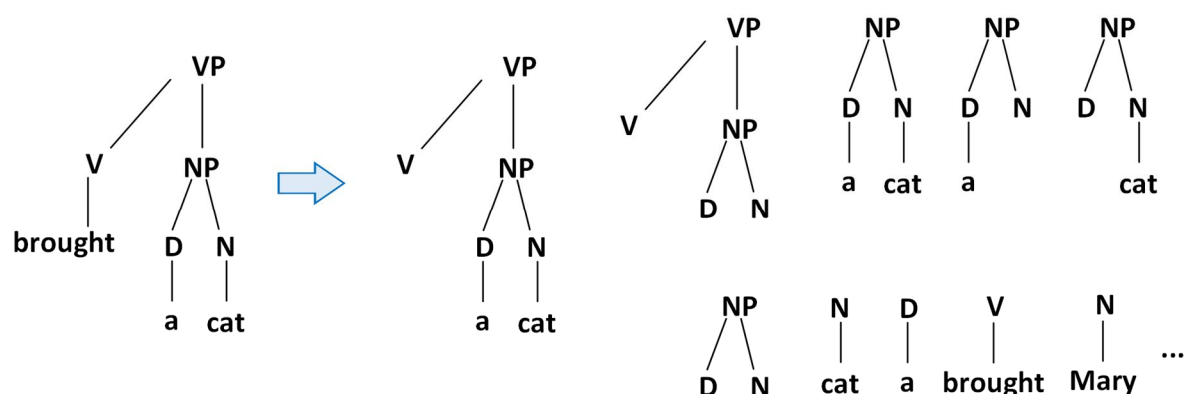


Figure 8: List of subset trees of a constituent parse tree

2.3.2.3 Graph Kernels

The third category of kernels is graph kernels based on graph parsing. All-paths graph kernel is a graph kernel counting weighted shared paths of all possible lengths [44]. Paths are produced from both the dependency parse graph and the surface word sequence of the sentence (Figure 9 and Figure 10).

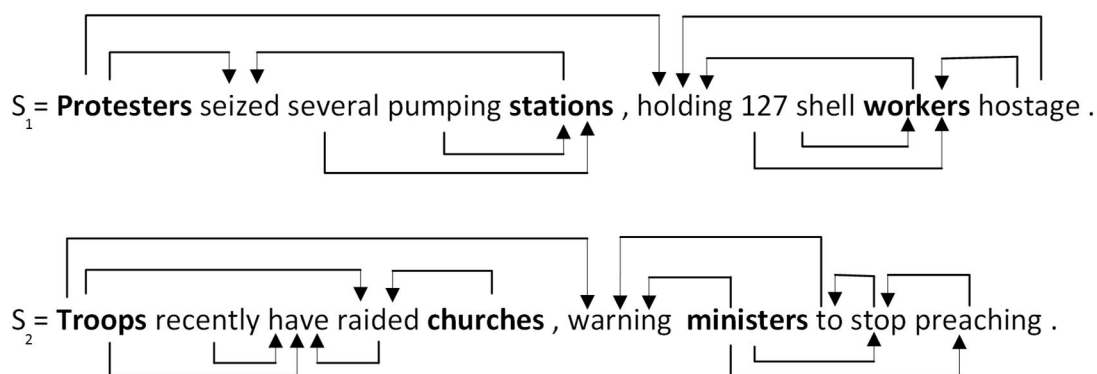


Figure 9: Sentences as sample of directional dependency graphs

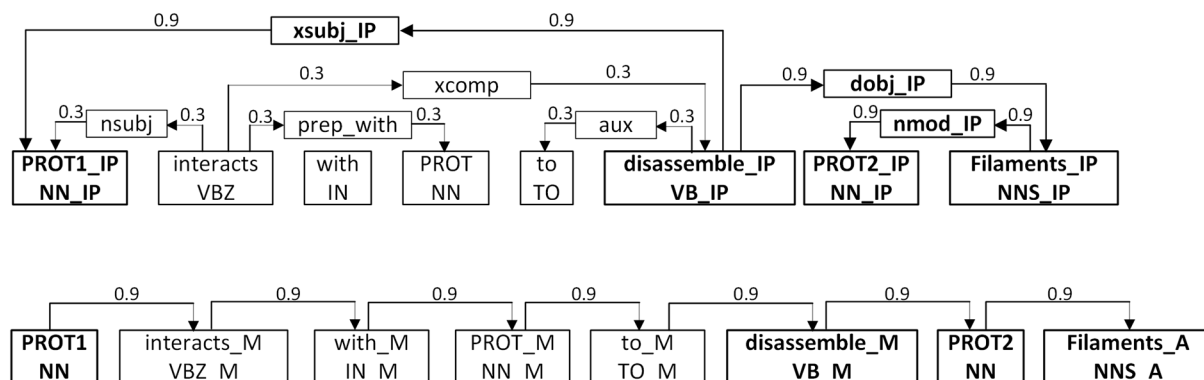


Figure 10: Graph representation generated from an example sentence with the candidate interaction pair is marked as PROT1 and PROT2

2.3.2.4 Other Types of Kernels

Some other forms of kernels have been developed through heuristic or combinational methods. Some new kernels proposed in [45], include predicate, walk, dependency, and hybrid kernels. Each of these kernels work on dependency trees extended with shallow linguistic. The walk kernel exhibited the best performance.

In addition, the general sparse subsequence kernel for the relation extraction [46], calculates the total number of weighted subsequences of a given length between two strings.

In [47], the authors define a Smith–Waterman distance function between two string sequences. Then, they define a local alignment kernel as the sum of Smith–Waterman distance scores on all possible alignments between the strings.

2.4 Related NLP Tasks

In this section, we introduce the basics and state-of-the-art NLP tasks of some linguistically driven operations and phenomena, which are considered to be investigated in the course of biomedical relation extractions in this thesis. They are negation operation, clause related subtasks and level of confidence and modalities. Ultimately, the aim of this research is to improve the performance of the biomedical relation extraction tasks through considering and employing the mentioned operations and feature sets.

2.4.1 Negation Detection Methods

One of the essential tasks in biomedical text mining is identifying negations. Linguists define negation as a morphosyntactic operation [2]. Through this operation, a lexical item either denies or inverts the meaning of another item or construction. The importance of negation in biomedical text mining is revealed when we consider the fact that negation is very common in texts leading to lack of precision in automatic information retrieval systems. As found by Chapman, 95% to 99% of reports in a radiological reports database include negations [48]. As a result, detecting negative assertions in most biomedical text mining

tasks is essential, where, in general, the aim is to derive factual knowledge from textual data.

The most known corpus annotated with negation are BioScope, linguistically based, and Genia, event-oriented [15]. In Bioscope, the main idea is based in the detection of a set of negation cue, like “no” or “not”. After this, the scope of the cue is calculated based on its syntactic context [49]. In Genia, biological concepts (relations and events) have been annotated for negation, but no linguistic cues have been annotated for them. In fact, the main objective of the BioScope corpus is to investigate this language phenomenon in a general, task-independent and linguistically-oriented manner. A more detailed comparison between these two corpora can be found in [50]. A number of examples of the sentences annotated with the BioScope guideline [49] are presented below. They demonstrate the annotation strategy, when the sentence is in active voice and the subject contains the negation cue (a), a sentence is in passive voice (b) and sentence is elliptic (c):

- a) Surprisingly, however, [{neither} of these proteins bound in vitro to EBS1 or EBS2].
- b) [A small amount of adenopathy <cannot be> completely {excluded}].
- c) This decrease was seen in patients who responded to the therapy as well as in those who did [{not}].

The SFU Review Corpus [51] consists of 400 documents annotated with negative and speculative keywords and their scope. A number of changes in their adaptation moved the negation cues from inside of the negation scope tags to the outside, and allowed the sentences to have only the negation cue, not the negation scope. Similarly, ConanDoyle-neg [52] is manually annotated with negation cues and their scope with several differences with the Bioscope corpus.

One of the researches that took negation into account in the relation extraction task was conducted by Faisal Chowdhury et al. [10]. They developed a list of features, such as the nearest verb to the pair entities in the parse tree and few negation cues, feeding the SVM classifier. They reported some improvement, but they did not specified how much the negation identification step enhanced the performance.

Furthermore, a survey that took negation into account was conducted by Pyysalo et al. [44]. They compared five PPI (protein-protein interaction) corpora in terms of several factors, including annotation of negative interactions and certainty level of interaction. According to them, BioInfer was the only corpus that had negative annotation.

2.4.2 Clause Dependency Detection in Relation Extraction

According to linguistics, a dependent clause refers to a group of words including a subject and a verb, which does not express a complete thought and usually extends the main clause. An independent clause or main clause, however, is one that can be seen as a complete sentence by itself, expressing a complete thought. Consequently, a complex

sentence consists of one independent clause together with one or more dependent clauses. Moreover, the term clause connector refers to a word used to join or to connect clauses to compose complex sentences. As an example, a complex sentence with a dependent clause in parenthesis and two negation cues is presented below:

- (Although the clinical significance is not known,) it is not recommended that *cefditoren pivoxil* be taken concomitantly with H₂ receptor antagonists

Identifying the role of clause dependency in complex sentences in DDI and SNP-Phenotype association detection is another linguistically driven subject investigated in this thesis. One of the few researches that has considered clauses, in relation extraction task, was conducted in [53], which attempted to select the best clauses and, consequently, developed a simplification algorithm. They reported some improvements regarding the different types of simplification and clause selection rules they used. In this research, it was attempted to extract new features based on the text or subtree features in the kernel-based relation extraction methods. This process is conducted by detecting their component's existence token or subtree in a dependent or independent clause, as well as the type of the clause itself, through checking several clause connectors.

A subtask employed in relation extraction tasks is sentence and clause simplification to overcome complexity of the sentences. Text simplification modifies, enhances, classifies or otherwise processes an existing text in a way that the grammar and structure of the prose are simplified to a great extent, while the original meaning and information remain the same [54]. ISIMP [55] is a system that simplifies the text so that its mining tools, including relation extraction tasks, can be improved. Along the same line, another research [11] applied some simplification techniques to simplify complex sentences by splitting clauses. They used some rules and patterns to split clauses and then adopted some simplification rules to generate new simple sentences. According to their conclusion, difficulty of resolving nested clauses is the major source of errors.

2.4.3 Level of Confidence and Neutral Candidate Detection in Relation Extraction

Identifying the intensity of a biomedical relation is an important and valuable task that allows us to extract deeper and more useful information about the desired relation. Estimation of the degree of confidence in information retrieval task is a difficult action that has not been seriously investigated for the extraction of relations from text.

Most of the current relation extraction problems and the produced corpora are based on binary relations, which decide whether a binary relation, between the two entities, exists in the sentence. Similarly, in the DrugDDI corpus [29] and almost all of other relation extraction corpora, the implemented systems did not consider the neutral relation candidates and accordingly, they were not annotated in the corpora.

Although detecting positive interaction is the main target of the DrugDDI corpus, there is a difference between a negative interaction candidate which has been stated by the authors

(distinguished candidate) and a negative interaction candidate which has not (neutral candidate); whilst, both are considered as negative in DrugDDI corpus. In other words, the neutral interaction candidate is one with no remark by the author, while the biased interaction candidate is exactly the opposite (with remarks sentence by the author). For instance in the sentence

- Studies in healthy volunteers have shown that *Acarbose* has no effect on either the pharmacokinetics or pharmacodynamics of digoxin, *nifedipine*, *propranolol*, or ranitidine.

There is no remark by the author about the interaction between *propranolol* and *ranitidine*. We define this pair of drugs as a neutral candidate.

One among the few studies on detection of neutral candidates in the course of relation extraction has been conducted by [56], introducing two iteration-based systems of DIPRE and Snowball that regard the confidence level of the relation. In both systems, when the confidence level is zero, there is a neutral candidate. In addition, Frunza and Inkpen conducted another similar research considering neutral candidates [21]. They categorized and extracted the semantic relationships between disease and treatments from biomedical sentences. However, no significant improvement was reported through using neutral class in the work.

On the other hand, although several researches have been conducted within the linguistics community on the use of hedging and confidence level in scientific text such as [57] and [58], there is little of direct relevance to the task of classifying from an NLP/ML perspective. According to linguistics, modality indicates the degree to which an observation is possible, probable, likely, certain, permitted, or prohibited.

One of the very few direct studies is [59], where the speculation identification problem is introduced using examples in the biomedical domain. They address the question of whether there is sufficient agreement among humans regarding what constitutes a speculative assertion to make the task viable from a computational perspective. Although they attempted to separate two shades of speculation: strong and weak, they failed to gather sufficient agreement for such a reliable distinction. However, they concluded that having a reliable distinction between speculative and non-speculative sentences was feasible and reliable automated methods might be also developed.

Nevertheless, to the best of our knowledge, no research has been conducted regarding the degree of confidence of the relations in the biomedical relation extraction task.

The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom.

Isaac Asimov

Chapter

Enhancing Automatic Extraction of Biomedical Relations using Different Linguistic Features Extracted from Text

3 Enhancing Automatic Extraction of Biomedical Relations Using Different Linguistic Features Extracted From Text

This chapter is devoted to present the main contributions shaped by this thesis. We describe our effort during this period of time as an exploration journey, which is more tangible. During the expedition, we visited different places, some stunning places, in the sense that they formed some of our new outcomes; some were sensational, where provided new routes to explore, and some places were left for the future; or too outlandish and even dangerous at this time which could lead us to a point of no return. As a curious adventurer, we always attempted to pick up valuable gifts from each one of these rooms. However, even the difficult steps form good experience, assisting me to evolve as a researcher. Truthfully, although we expected that the effort collected in this thesis will be worth the trouble, it is also true that we decided to conclude the story with some sedition—see Section 4.2 for details—. Furthermore, now there are some exciting directions to follow, we will continue our voyage in a short time with more work that, hopefully, will produce some new results.

After exploring the field, the first direction that we followed was to study and figure out the strengths of methods presented by the participants of the DDI Challenge 2011 [29].

In the rest of this section, we will enumerate the objectives of the thesis and explain our related contributions asserting the mentioned objective.

3.1 Improving the Performance of Methods of the DDI Extraction from Text through Detecting Linguistic-Based Negation

In this part, we mention our contributions in this thesis that is related with our first objective: to extract DDIs from text employing negation related features.

We start with the annotation of the DDI corpus with the linguistic-based negation and then the implemented neutral DDI feature extractor will be mentioned.

3.1.1 Annotating the Drug-DDI Corpus with Negation

Marking the DrugDDI corpus with linguistic based negation is the first contribution in this thesis. As mentioned before, two negation detection methods have been developed and employed to annotate the used corpora: a linguistic-based (Bioscope) approach and an event-oriented (Genia) approach. The adaptations of the Bioscopes guidelines, briefly mentioned earlier, alongside to its capability to be extracted automatically and needing less manual working, have proved potentially capable of feeding the study. However, they may be hard to synthesize with kernel methods in tasks like ours.

In addition, as mentioned earlier, although the DrugDDI corpus is a valuable resource to perform comparable experiments in order to investigate relation extraction methods, one restriction of this corpus is lack of negation annotation.

Some analyses were conducted on the DrugDDI corpus, showing that a significant number of sentences in the corpus have at least one negator [60]. Consequently, it can be concluded that identifying negative statements is an essential task to obtain accurate knowledge from textual data.

In the present thesis, we annotated the DrugDDI corpus (2011 and 2013 versions) with negation scope and cues automatically [9] and manually. The DrugDDI corpus (2013) included two parts: DrugBank and Medline parts, which were annotated automatically with a rule-based method with the BioScope guidelines and then checked manually using the BRAT annotation tool. The NegDDI-DrugBank and NegDDI-MEDLINE corpora are the final products of the whole process (Figure 11).

```
<sentence id="DDI-DrugBank.d297.s4" text="Concurrent therapy with ORENCIA and TNF antagonists is not recommended.">
<entity charOffset="24-30" id="DDI-DrugBank.d297.s4.e0" text="ORENCIA" type="brand"/>
<entity charOffset="36-50" id="DDI-DrugBank.d297.s4.e1" text="TNF antagonists" type="group"/>
<pair ddi="true" e1="DDI-DrugBank.d297.s4.e0" e2="DDI-DrugBank.d297.s4.e1" id="DDI-DrugBank.d297.s4.p0" type="advise" />

<negationtags> <xscope> Concurrent therapy with ORENCIA and TNF antagonists is <cue> not </cue>
recommended </xscope> .</negationtags>

</sentence>
```

Figure 11: A sample of sentence annotated with negation scope and cue in the NegDDI corpus

The analyses of the NEGDDI corpus demonstrate that the negated statements consist of approximately 21% of its sentences. Furthermore, our analyses show, “not” and “no” are by far the most frequent cues in the corpora. However, more changes during the manual checking process have been performed with the cue “not”, forming 27.41% of changes. Most of the errors with the other cues are associated with problems detecting certain patterns of passive voice sentences. It is worth mentioning that although, we conducted a manual checking for the annotations, the experiments mentioned in 3.1.2 and 3.2.1 briefly show that the usage of manually checked annotations do not considerably affect the performance of the proposed DDI extraction method. Therefore, it can be concluded that the automated negation detection method indicate satisfactory results.

The mentioned contribution explained and presented in our papers was published in SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural, 2013) and LREC (International Conference on Language Resources and Evaluation, 2014) conferences.

In the paper presented under the title of “Extracting Drug-Drug Interaction from Text Using Negation Features”, we explain the detailed process of automatically and manually annotation of the DDI 2011 corpus with negation scope and cues according to the Bioscope guidelines.

Additionally, in the paper with the title “Exploring Negation Annotations in the DrugDDI Corpus”, a number of analyses concerning the correlations between negations and DDI annotations are presented. Furthermore, the annotation with negation scope and cue are extended to the DDI corpus 2013. The annotation process includes the automatic and manual checking conducted by the BRAT annotating tool.

It is worth mentioning that the analyses presented in these papers led us to implement the comprehensive enhancement method published in our paper in the PLoS journal under the title “Enhancing Extraction of Drug-Drug Interaction from Literature Using Neutral Candidates, Negation, and Clause Dependency”.

3.1.2 Proposing the Neutral Candidates Features and Linguistic-based Negation

The neutral candidates are a subclass of non-positive class of candidates, not mentioned by the authors in the text. These candidates are important, since their status is not be inverted, if they are located in the negation scope. Other contribution of the thesis is introducing different types of neutral relation candidates and implementing a feature extractor method. The paper presented in the PLoS journal has clarified this work with more details. We identified three types of neutral candidates and developed a rule-based system to detect three types of text patterns.

Taking neutral candidates into account is critical from another perspective, since not doing may induce conflicts in the corpus later. In this situation, the author did not make any remarks concerning the interaction or association between the two biological entities and it is possible that in the future, other researchers could find an interaction, which would lead the corpus to face conflicts.

The significant contribution of neutral candidates and effectiveness of related features has been confirmed in the two studied corpus. The important role of neutral candidates in the SNP-Phenotype Association has been mentioned in our paper presented in the Journal of Biomedical Semantics.

Moreover, it is important to mention that the proposed neutral-related rules can be used with extremely slight changes in other biomedical relation extraction tasks, particularly symmetric relations such as protein-protein interaction.

The semantic analyses of the sentences of the corpus assist us in identifying three types of neutral candidates in sentences, which can be extended in other relation extraction tasks. In addition, we implemented a rule-based system to detect different types of neutral candidates. Our experiments indicate that the features aimed at detecting the neutral-related candidates are the most effective category among the three categories occurred in three linguistic patterns.

Moreover, we extracted six Boolean features to detect relative position of biomedical entities with negation scope. Additionally, the negation cue was employed as other negation related feature.

We conducted a number of experiments using the proposed features. The experiments indicate that the best result was achieved by the enhanced local context kernel method with 68.4% F-measure, which is 2.7%, more than the first system in the DDI extraction (2011) challenge. Moreover, the experiments indicate the best combination of invented feature sets (the two mentioned feature sets and the clause dependency feature set that will be explained in 3.2.1) for the proposed local context kernel is neutral candidate with negation cue and scope features, producing a slightly more improvement than the entire list of the invented features.

As mentioned earlier, to the best of our knowledge, the only research that considers negation in the DDI extraction has been conducted by Chowdhury et al. However, they did not report the improvement obtained for usage of negation. Therefore, we compared our results with other participants of the challenges and also verified the significance of the obtained improvement in comparison to the original methods.

3.2 Improving the Performance of DDI Extraction Methods from Text through Detecting and Discriminating Between Different Clauses

In the second part, we will mention our other contributions in this thesis, creating the second objective, which is pay off to the DDI extraction from text through considering complex sentences. We start with mentioning the proposed method enhancing the utilized kernel methods through considering the clause dependency related features. Then, we state our other contribution of the thesis, which is devoted to the combination of three and sequence kernels.

3.2.1 Enhancing the DDI Supervised Extraction Methods Using Clause Dependency Features

Another contribution of the thesis is dedicated to overcome the complex sentences. The complex and negated sentences are two major sources of inaccuracy in biomedical relation extraction.

We proposed a method to distinguish the components of the studied kernel methods by detecting their position in dependent or independent clauses and their types of related clause connectors. The experiments indicate that the ratio of negation cues is higher in complex sentences in comparison with simple ones. Additionally, the results show that by employing the proposed features combined with a bag of words kernel, the performance of the used kernel methods improves. Furthermore, experiments demonstrate that the enhanced local context kernel outperforms other methods. The proposed method can be used as an alternative approach for sentence simplification techniques in biomedical area, which is an error-prone task.

The paper published in the PLoS journal (2016) shows the comprehensive approach by employing the three linguistics-based features, i.e. negation, clause dependency, and neutral candidates. Although the features pertinent to clause dependency were introduced briefly in a previous paper, we proposed more features with additional details in this paper to improve the performance of the DDI extraction task. Furthermore, the overall performance of the method as well as the contribution of each feature set in the performance of the system is presented for both DrugBank and Medline parts.

Our experiments demonstrate considering the clause dependency and type of clauses beside to negation related features improve the performance of the DDI extraction methods. The obtained improvements of the features in conjunction with neutral related features mentioned in 3.1.2 in comparison with the original kernel methods were verified through identifying the DDIs in the test parts of the DDI corpus as well as a statistical sign test. The sign test demonstrates that the achieved improvements are significant.

3.2.2 Proposing a New DDI Extraction Method through Combining Tree and Sequence Kernels

Combining the tree and sequence kernels through a BOW (bag of words) method is the other contribution of the thesis. The proposed method shows better performance in comparison with each of the subtrees, subset trees and global and local context kernels separately. Our method and experiments show that the combination of different types of kernels can improve the overall performance of the methods.

More details of the contribution have been explained in our paper presented in the SEMEVAL Conference (International Workshop on Semantic Evaluation, 2013) under the title of “NIL UCM: Extracting Drug-Drug Interactions from Text Through Combination of Sequence and Tree Kernels”. In this paper, we employ Subtree and SubSetTree, local context kernel, global context kernel, and some conjunction features. The proposed conjunction features include POSLEMMA (POS+Lemma) and POSSTEM (POS+ Stem), the first verb before Drug1 and the first verb after Drug2, and their stems and lemmas.

It is worth mentioning, we have proposed two DDI extraction approaches in the paper; the first approach with all categories of the DDI sentence types and the second approach that initially extract positive and negative DDIs and then a second classifier was used to classify

the positive extracted DDIs. It is observed that the results of the detection of the DDI are better with the two-step approach: 0.656 against 0.588 on F1. The implemented system with the two-stage classification was ranked the 3rd in the DDI extraction challenge (2013).

3.3 Preparing a Corpus For Extracting SNP-Phenotype Association from Text, Annotated With Negation, Modality and Ranked Associations

Here, we mention our contributions in this thesis regarding the third objective, which is devoted to producing the SNPPhenA corpus. We start with the main steps of producing the corpus including collecting documents, recognizing the entities and annotation of negation, SNP-phenotype associations, three levels of association and their level of confidence. The next related contribution is preparing the website for the corpus, making its usage easier.

3.3.1 Producing the Ranked SNP-Phenotype (SNPPhenA) Association Extraction Corpus

The second biomedical relation extraction corpus prepared and employed during the thesis is the SNPPhenA corpus. The corpus was prepared to extract ranked SNP-Phenotype association from text. It is the first relation extraction corpus annotated with degree of confidence, showing the strength of associations. The process of producing the corpus includes collecting abstracts, recognizing named entity, and annotating the ranked association, negation scope and cues as well as modality markers. In addition, the confident level of positive association was annotated in three categories: Strong, moderate and weak degree of confidence. Most frequently phenotypes, SNPs, and some basic statistics concerning the produced corpus are presented. The corpus is generated in two formats: xml and standoff BRAT formats. Figure 12 and Figure 13 present an example of an annotated sentence in xml and brat standoff formats. Furthermore, the Figure 14 shows a visualization of the annotated sentence using brat. Moreover, the inter-annotator agreement is analyzed, and the Kappa coefficient is calculated for SNP-phenotype associations and the degree of confidence of associations showing the reliability of the corpus. The Kappa inter-annotator agreement between the two annotators was calculated 0.79 for annotating the associations and 0.80 for annotating the confidence degree of associations, approving the reliability of the corpus.

Moreover, the analyses show 16.8% of the sentences have at least one negation cue and 76.3% of the samples are distinguished (i.e. they are positive and negative association candidates). Additionally, 63.8% of the candidate sentences have at least one clause connector, while 36.2% do not have one. It is necessary to mention that the prepared SNPPhenA corpus is the first ranked biomedical relation extraction annotated reliably with the degree of confidence of associations.


```

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE abs SYSTEM "SNPPPhenA.dtd">
<abstract TEXT="UNLABELLED: Background: It has been suggested that the serine/threonine kinase 15 (STK15)
... Stratified analysis by cancer type revealed that the STK rs2273535 polymorphism may contribute to the
risk of breast cancer (AA vs. TT:OR=1.21, 95%CI=1.01-1.44, Pheterogeneity=0.002), colorectal cancer (AA vs.
UNASSIGNED: OR=1.24, 95%CI=1.05-1.47 , Pheterogeneity=0.124), and esophageal cancer (AA vs. UNASSIGNED:
OR=1.19, 95%CI=1.02-1.39, Pheterogeneity=0.148) ... "ABSTRACTID="1130">
  <sentence END="1281" START="920" ID="1130_0">
    <snp TEXT="rs2273535" END="986" START="977" ID="0"/>
    <phenotype END="1281" START="943" ID="0" text="cancer"/>
    <phenotype END="1281" START="1030" ID="3" text="breast cancer"/>
    <phenotype END="1281" START="1105" ID="1" text="colorectal cancer"/>
    <phenotype END="1281" START="1196" ID="2" text="esophageal cancer"/>
    <modality_marker END="963" START="955" text="revealed"/>
    <modality_marker END="1003" START="1000" text="may"/>
    <pair CONFIDENCE="weak" ASSOCIATION="positive" SNPID="0" PHENOTYPEID="0" PAIRID="0"/>
    <pair CONFIDENCE="weak" ASSOCIATION="positive" SNPID="0" PHENOTYPEID="1" PAIRID="1"/>
    <pair CONFIDENCE="weak" ASSOCIATION="positive" SNPID="0" PHENOTYPEID="2" PAIRID="2"/>
    <pair CONFIDENCE="weak" ASSOCIATION="positive" SNPID="0" PHENOTYPEID="3" PAIRID="3"/>
  </sentence>
</abstract>

```

Figure 12: The unified XML format of a sentence in the SNPPPhenA corpus

T1	SNP	57 66	rs2273535
T2	Phenotype	23 29	cancer
T3	Phenotype	110 123	breast cancer
T4	Phenotype	185 202	colorectal cancer
T5	Phenotype	276 293	esophageal cancer
T6	Modality_Marker	35 43	revealed
T7	Modality_Marker	80 83	may
R1	weak_confidence_association	Arg1:T1 Arg2:T2	
R2	weak_confidence_association	Arg1:T1 Arg2:T4	
R3	weak_confidence_association	Arg1:T1 Arg2:T5	
R4	weak_confidence_association	Arg1:T1 Arg2:T3	

Figure 13: The annotation of the sample sentences in brat format (*.ann)

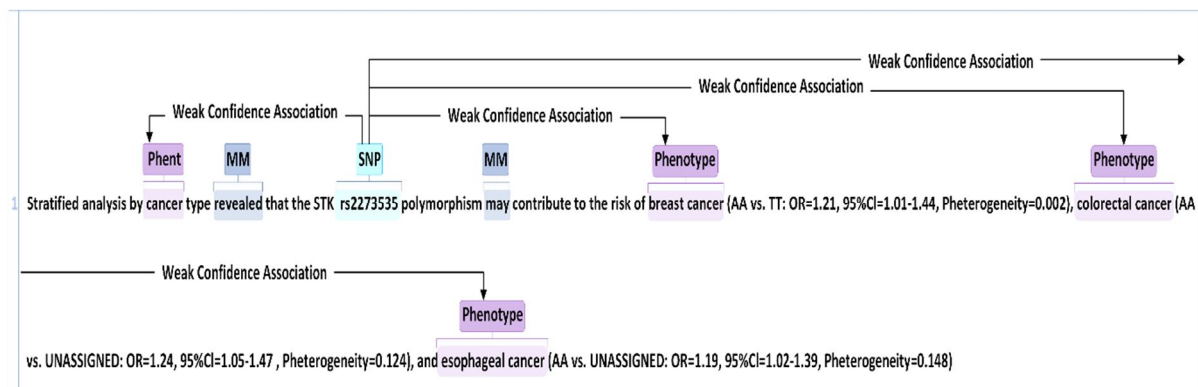


Figure 14: A sample of a sentence in the produced SNPPPhenA corpus drawn by brat

More details of the contribution have been presented in our paper presented in the Journal of Biomedical Semantics (2017) under the title “SNPPhenA: A Corpus for Extracting Ranked Associations of Single-nucleotide Polymorphisms and Phenotypes from Literature”.

In this paper, some statistics regarding linguistic features and annotation tags are presented. In addition, the paper presents the document of guidelines for the corpus and some analyses pertinent to the reliability of the annotation.

The process of producing the corpus is mentioned in the paper and most frequently phenotypes, SNPs, and some basic statistics concerning the corpus produced are presented. In addition, the initial experiments are conducted with the corpus to extract the associations and the confidence level of associations with two popular kernel methods.

3.3.2 Developing a Website for the SNPPhenA Corpus

Additionally, for better and more convenience usage of the corpus, a website for the corpus was developed, which is available online at NIL Server. The website contained a brief introduction of the corpus and inter-annotator agreement analyses. Furthermore, the guidelines document; xml files of the corpus, dtd, ann and txt files were uploaded to the website.

The details of the mentioned contribution and preparing the website are explained in the paper presented in in the Journal of Biomedical Semantics (2017).

[Home](#) : [Behrouz Bokharaeian](#)

SNPPhenA: A corpus for extracting ranked associations of SNP and phenotypes from literature

Submitted by behrouz on Mon. 10/ 17 / 2016 – 14:07

The SNPPhenA corpus

The SNPPhenA corpus consists of medical and biological texts annotated for snp-phenotype associations, negation, modality markers and degree of confidence of associations. This was done to allow a comparison between the development of systems for association extraction as well as the degree of confidence and strength of associations. The corpus is publicly available for research purposes.

The annotation guidelines: [pdf](#)

Annotation principles are also discussed in the following paper:

Corpus download

Information provided in the <http://www.gopubmed.org/> search engine was used to collect genome-wide association abstracts. GoPubMed is a webserver that allows users to explore PubMed search results with Gene Ontology . Here is DTD for the xml files containing the annotations: [DTD](#)

Abstracts of the SNPPhenA corpus: xml v1.0

The full corpus in XML and BRAT formats is available in one file: [zip](#)

An online association extraction system that utilizes the SNPPhenA corpus is available [here](#).

Inter-agreement analysis

In order to evaluate the quality of the corpus and the reliability of the annotations, inter-annotator agreement score was measured for the task of classifying candidate sentences into positive, negative and neutral classes, and also for task of determining the confidence level of the association. two annotators independently have tagged the corpus.

Figure 15 : A screenshot of the produced website corpus for the SNPPhenA

3.4 Developing a Method for Extracting Graded SNP-Phenotype Associations from Text through Degree of Confidence and Negation

Finally, we mention our contributions in this thesis dedicated to our last objective, which is developing a method to extract graded SNP-phenotype associations from text. Initially, we explain the important criteria that must be verified to ensure that the negation-neutral based method can work effectively. The next contribution allowing us to reach the objective is the proposed association extraction method that initially decides on the existence of association and then identifies the degree of confidence of association. Additionally, as the other contribution, we developed a web-based program that first recognizes entities and secondly identifies ranked SNP-Phenotype associations.

3.4.1 Suggesting the Criteria for Reliability of the SNP-Phenotype Association Extraction Method

Another contribution of the thesis is suggesting a criterion that must be verified to ensure that the proposed negation and neutral-based method can be employed in other corpora.

To examine whether the proposed method is applicable to other corpora or not, the verification criteria must be analyzed, i.e. complexity of the sentences and uniform polarity of the sentences.

The paper published in the Journal of AI and Data Mining (JAIDM) under the title “Automatic Extraction of Ranked SNP-Phenotype Associations from Literature through Detecting Neutral Candidates, Negation and Modality Markers” explains the criteria with more precise details.

Moreover, some statistics related to the complexity of the sentences and the innate polarities of the sentences are presented in the paper and the paper published in in the Journal of Biomedical Semantics.

Furthermore, our analyses of the DDI corpus demonstrate the high number of dependent clauses of the sentences and non-uniform polarity of the key verbs, leading us to the fact why the proposed method mentioned in 3.4.2 has poor performance in comparison to the result obtained from the SNPPhenA corpus.

3.4.2 Proposing a New Method for Extraction of SNP-Phenotype Association and Degree of Confidence of Association through Linguistic-Based Negation

Implementing a supervised method to extract SNP-Phenotypes associations through detecting neutral candidates and negation scope and cues is other contribution of the thesis. The paper presented in the Journal of AI and Data Mining (JAIDM) shows the details of the method.

The relation extraction method relied on linguistic-based negation detection and neutral candidates. The experiments showed that negation cues and scope as well as detecting neutral candidates can be employed to implement a superior relation extraction method outperforming the kernel-based counterparts due to a uniform innate polarity of sentences and the small number of complex sentences in the corpus.

Furthermore, we implemented a novel modality-based supervised method (MMS) to identify the level of confidence of the extracted association. The proposed method employs a classifier trained by the modal markers, the mentioned p-value, some other linguistics features and the confidence level of the association annotated in the corpus.

To evaluate the performance of our proposed association extraction method, two other schemes are also tested, namely local context and sub-tree kernel methods. We have used

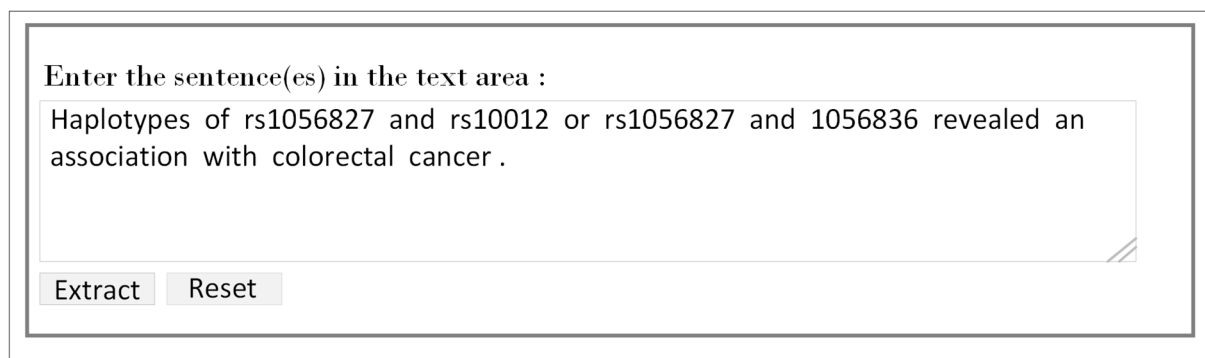
the two enhanced kernel-based methods as the benchmarks, since there is no available method to extract the ranked SNP-Phenotype association from text working particularly with negation and neutral candidates.

Our experiment shows that the proposed method outperforms the mentioned schemes. Moreover, the proposed MMS method outperformed the utilized benchmark method in identifying the degree of confidence of associations. We used the BOW as a benchmark for this purpose, since to the best of our knowledge; there is no method for identifying the level of confidence of associations in the biomedical relation extraction task.

The best f-measure was achieved in candidate expressions related to associations with a weak degree of confidence and the worst result was obtained in the medium degree of confidence. However, the experiments demonstrate that both methods have weak f-score, recall and precision in the category of the middle level of confidence.

3.4.3 Developing an Online Web Based Portal for Extracting SNP-Phenotypes from Text

Other contribution of the thesis is a web-based version of the ranked SNP-phenotype association extraction method. The application detects negation cues and scope and the container clause. Moreover, through employing a trained SVM classifier, the type of the candidate is detected. Additionally, the web-based program classifies the level of confidence of association in three mentioned degrees (Figure 16 and Figure 17). The details of the algorithm and the used technologies are available in our paper presented in the Journal of Biomedical Semantics.

The screenshot shows a web interface for extracting SNP-phenotype associations. It features a text input area with a placeholder instruction "Enter the sentence(es) in the text area :". Below the input area, there is a sample sentence: "Haplotypes of rs1056827 and rs10012 or rs1056827 and 1056836 revealed an association with colorectal cancer .". At the bottom of the interface, there are two buttons: "Extract" and "Reset".

Enter the sentence(es) in the text area :

Haplotypes of rs1056827 and rs10012 or rs1056827 and 1056836 revealed an association with colorectal cancer .

Extract Reset

Figure 16 : A screenshot of the web based ranked association extractor

Sentence: Haplotypes of A1450G and rs10012 or rs1056827 and rs1056836 revealed an association with colorectal cancer.

Phenotype: colorectal cancer

SNP: A1450G

Negation cue:

Status of SNP-Phenotype Association: true

Degree of confidence of Association: Medium

Phenotype: colorectal cancer

SNP: rs10012

Negation cue:

Status of SNP-Phenotype Association: true

Degree of confidence of Association: Medium

Phenotype: colorectal cancer

SNP: rs1056827

Negation cue:

Status of SNP-Phenotype Association: true

Degree of confidence of Association: Medium

Phenotype: colorectal cancer

SNP: rs1056836

Negation cue:

Status of SNP-Phenotype Association: true

Degree of confidence of Association: Medium

Figure 17 : A screenshot of the result of the web based ranked association extractor

4

Chapter

Conclusions and Future Works

Two things are infinite: the universe and human stupidity;
and I'm not sure about the universe.

Albert Einstein

4 Conclusions and Future Works

The publications summarized in the previous chapter (and will be presented in the next part of this thesis) to our knowledge, are a beneficial effort in the field of extracting biomedical relations from text, and more specifically, Drug-Drug interaction and SNP-Phenotype association extraction area. This hard attempt not only helped us to grow as a researcher, but also enabled me to grow as an individual. I hope to have established my future within the scientific community, whose first step is, in fact, becoming a reality while writing these concluding lines of my thesis.

However, this period of my life has taught me that constancy together with the hope of being following the right path, have their reward. Real evidence is this work that has now definitely formed.

This chapter will conclude with a discussion regarding the potential ways of continuing our research. In the first place, concerning the near future, we present some partial results and ideas that we are investigating by now, all of them corresponding to the field of DDI and SNP-Phenotype association extraction from text.

At the end, I will present our proposals for the future, when I expect to explore new horizons broadened to us at this time.

4.1 Conclusions

As we mentioned earlier, the thesis pursued four main objectives shaped by the mentioned contributions. In this section, we explain the conclusions reached during the process of achieving our objectives.

In the rest of this section, we mention the conclusions reached in terms of our main objectives mentioned with further details in the first chapter.

4.1.1 Improving the Performance of Methods of the DDI Extraction from Text through Detecting Linguistic-Based Negation

Our obtained results concerning the DDI task showed that the linguistically-oriented scope-based negation annotating identifying the negation cue and scope, may not have always enough information to overcome the act of negation in the DDI task. Therefore, one must take account of other sorts of bases and factors including identifying the neutral candidates and dependency of clauses.

According to the obtained results, the neutral candidate feature set is the most useful one among the three different invented feature sets, generating better results with the combination of the other two invented feature sets.

In addition, the experiments and analyses demonstrated that, medical texts with symmetric relations such as Drug-Drug Interaction are more difficult regarding the consideration of negation in comparison with biological texts with asymmetric associations. Complex structure of sentences mentioning the DDIs, large average number of the DDI candidates in each sentence and uninform polarity of the key verbs are the main factors that we relied to be considered. Although, developing more accurate methods for identifying the polarity of the key verbs needs more investigation, it leads to better performance of the DDI extraction methods.

4.1.2 Improving the Performance of the DDI Extraction Methods from Text through Detecting and Discriminating between Different Clauses

In addition, as analyses of the DDI corpus show, sentences with negation cue have more clause connectors compared to sentences without negation cue. Therefore, taking clause connectors and dependent clauses into account is important in resolving the negation action.

The experiments show that the used sequence kernels benefit more from the clause related features in comparison to tree kernels. Consequently, it can be concluded that the tree kernels consider clauses more than sequence kernels.

Although combination of clause dependency related features with other kernel methods demonstrates significant improvement, we can reach the conclusion that extracting more features related to different types of clauses. In other words, concessive clause in addition to an effective feature selection method can have better performance.

While, the current results are promising, one of the challenging discussions is whether all kernel methods benefit from the clause, negation and neutral related features. As results of the conducted experiments, we maintain that it is probable that more advanced kernels deriving more informative features from different presentations of the sentence, may not be beneficiary from the proposed features.

Furthermore, as a result of the conducted experiments to determine the contribution of the different invented feature sets, in most of the experiments, the complete list of features provided the best results. However, in few of them, the list did not have the best performance, in comparison to the other possible combinations of datasets of features; therefore, a suitable feature selection method can improve the results more.

4.1.3 Preparing a Corpus For Extracting SNP-Phenotype Association from Text, Annotated With Negation, Modality and Ranked Associations

In addition to the conclusions reached on the DDI extraction task and related annotations, we have arrived conclusion through experiments conducted on the SNPPhenA Corpus. As opposed to the previous biomedical relation extraction corpora containing true and false types of relations, the annotated associations in the corpus were divided into three classes: positive, negative and neutral candidates. Identifying neutral candidates is critical for the

negation process, since the status of those candidates and their corresponding level of confidence did not change, when they were located in the scope of negation terms, while the status of distinguished association candidates did change in such cases. Similarly, the level of confidence, certainty or uncertainty of a neutral candidate did not change if it was located in the scope of a speculation or modality term. Therefore, determining the effect of negation as well as modality terms requires identification in neutral candidates.

It should be mentioned that the SNPPhenA corpus must be considered an initial step in extracting graded associations from the literature so that it can be used by other researchers as a resource to evaluate and compare the association extraction methods trying to identify the degree of confidence of associations.

However, annotation of other modality and clause related features and identification of statistical features and values useful in the task of the degree of confidence identification are among the works that can improve the credibility and authority of the corpus.

4.1.4 Developing a Method to Extract Graded SNP-Phenotype Associations from Text through Recognizing Degree of Confidence and Negation

In this thesis, we proposed a ranked SNP-phenotype association extraction method based on the degree of confidence of association. The results demonstrate the superior performance of the proposed method. Additionally, the results show the neutral candidates are the important category of candidates that can be utilized to implement better relation extraction methods. Furthermore, the achieved results show the importance of the confidence level of the association as a linguistic-based factor that can be used in addition to the existing methods to obtain more useful information.

Identifying the important criteria to be verified in order to perceive the effectiveness of the proposed method is the other contribution of the thesis; however, other types of sentences and factors might remain unrecognized. It may need more analyses and experiments with other biomedical corpora in different fields.

Consequently, it is expected for asymmetric relations such as Gene-Disease and Disease-Treatment associations, to take more advantage from the proposed method and introducing the neutral candidates, rather than symmetric relations. The reason is that for symmetric relations such as PPI and DDI, every binary combination of entities in a sentence is a candidate relation. Therefore, they have significantly more neutral examples in comparison with the asymmetric relations.

Generally, it can be concluded, identification of confidence level of association in biomedical domain is a difficult task. Lack of uniform usage of modality and confidence-related linguistic-rules by the authors, nonexistence of the united interpretation of the statistical tests and usage of different statistically significant tests between the researchers are among the factors making the task difficult. Thus, a precisely ranked SNP-Phenotype association extraction method based on the degree of confidence of associations must include these factors.

4.2 Future Works

We start this section by the future directions that can be pursued concerning the DDI extraction task and discuss a new idea and open rooms for SNP-Phenotype association task.

4.2.1 DDI Extraction

First, encouraging future work over the DDI extraction task can be the expansion of the definition of the Drug-drug interaction extraction from a binary relation to a ranked relation through considering the corresponding confident level and other linguistic features; similar to the work we did with the SNPPhenA Corpus to some extent. Expansion of the confident level concept to a membership function for a fuzzy DDI relation instead of a crisp DDI relation will enable us to compare and combine the extracted results from different sentences. In other words, dissimilar results for a specific DDI candidate extracted from different sentences with different confident levels can be compared and combined with each other. Comparing and combining the different results for a specific candidate will help to identify different types of the errors including systematic or human mistakes, which can lead to boosting the overall performance of the system. The achievement is not possible with a crisp DDI relation that only detects the existence or lack of existence of an interaction. Speculation and deduction cues include modal verbs of possibility such as “may” and related adjective and adverbs such as likely in addition to the proposed rule-based system to identify neutral confidents that can be used to calculate the membership function (the confident level).

In this regard, although during the thesis some experiments for using a few basic simplification methods were conducted on DDI corpus to overcome the complex sentences, no significant improvement was achieved. However, it is believed that a remarkable future work is the combination of simplification and a pronoun resolution specified for drugs leading to better performance.

4.2.2 SNP-Phenotype Association Extraction

In the rest of this section, we present some new future works and directions that can be followed in the continuation of the work we conducted for the new proposed ranked SNP-Phenotype association extraction task.

It is important to remember that the SNPPhenA Corpus and the proposed association extraction method are an initial step to extract graded associations from literature, which could lead to an idea for complete fuzzy association extraction task that can be employed to construct better and more realistic biomedical ontologies automatically. More generally, although all existing relation extraction corpora and methods utilize crisp relations, they could be replaced with a better mathematical model called probabilistic fuzzy relations (PFR) which is newer mathematical model than usual fuzzy relations. Membership function

and probability function are two functions used in a probabilistic fuzzy method indicating the level of strength of association between SNP and phenotype and confidence level of sentence, respectively. Thus, in future work, implementation of probabilistic fuzzy relations could be further investigated. Using PFRs, modals in sentences could be considered, while combining different confidence levels. Presumably, it will lead to improved results.

Moreover, employing other linguistics-based or non-linguistic-based factors that could be utilized to determine the credibility of the reported association is an important future work. Assessing the statistical confidence level of the used case control tests mentioned in the text, as well as using genotyping techniques (such as MLPA or RFLP) in addition to more accurate epistemic modal analysis methods, are among the factors that could be employed to identify the overall degree of confidence and credibility of the reported associations.

Moreover, although the proposed sentence-level ranked SNP-Phenotype association extraction method shows promising results, the estimated level of the confidence of association can be used in addition to other factors such as confidence level of the abstract and the paper itself to define the overall confidence and credibility of the extracted association. For example, number of citations and credibility of the publisher can be among the factors determining the confidence and credibility of the abstract and the paper.

.

Bibliography

- [1] Dayou Zhong Deyu Zhou and Yulan He, "Biomedical Relation Extraction: From Binary to Complex," *Computational and Mathematical Methods in Medicine*, vol. 2014, no. 2, pp. 139-155, 2014.
- [2] Eugene E. Loos, Susan Anderson, Day, Jr. Dwight H., Paul C. Jordan, and J. Douglas Wingate, *Glossary of linguistic terms*. Camp Wisdom Road Dallas: SIL International , 2004.
- [3] Muriel Harris and Katherine E Rowan, "Explaining grammatical concepts," *Journal of Basic Writing*, vol. 8, no. 2, pp. 21-41, 1989.
- [4] R. E. Asher and J. M. Y Simpson, *The Encyclopedia of language and linguistics*, 2nd ed.: Pergamon Press, 2006.
- [5] D. Blakemore, "Evidence and modality," in *The Encyclopedia of language and linguistics*, 2nd ed.: Pergamon Press, 2006, pp. 1183–1186.
- [6] Greene Shepherd, Philip Mohorn, and Kristina Yacoub, "Adverse drug reaction deaths reported in United States vital statistics, 1999--2006," *Annals of Pharmacotherapy*, vol. 46, no. 2, pp. 169-175, 2012.
- [7] Gabor T Marth et al., "A general approach to single-nucleotide polymorphism discovery," *Nature genetics*, vol. 23, no. 4, pp. 452-456, 1999.
- [8] M. Krallinger, "Importance of negations and experimental qualifiers in biomedical literature," in *NeSp-NLP'10 Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, 2010.
- [9] Miguel Ballesteros, Jesús Herrera , Virginia Franci, and Pablo Gervás, "Inferring the Scope of Negation in Biomedical Documents," in *In proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics*, New Delhi, 2012.
- [10] Md. Faisal, Mahbub Chowdhury, Alberto Lavelli, and Fondazione Bruno Kessler, "Exploiting the Scope of Negations and Heterogeneous Features for Relation Extraction: A Case Study for Drug-Drug Interaction Extraction," in *HLT-NAACL13*, 2013, pp. 765-771.
- [11] Isabel Segura-Bedmar, Paloma Martinez, and Cesar de Pablo-Sanchez, "A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents," *BMC Bioinformatics*, vol. 12, no. Suppl 2, p. S1, 2011. [Online]. <http://www.biomedcentral.com/1471-2105/12/S2/S1>
- [12] Robert J Klein et al., "Complement factor H polymorphism in age-related macular degeneration," *Science*, vol. 308, no. 5720, pp. 385-389, 2005.

- [13] Barbara E Stranger, Eli A Stahl, and Towfique Raj, "Progress and promise of genome-wide association studies for human complex trait genetics," *Genetics*, vol. 187, no. 2, pp. 367-383, 2011.
- [14] Pierre Zweigenbaum , Dina Demner-Fushman , and Hong Yu , "Frontiers of biomedical text mining: current progress," *Briefings in Bioinformatics*, vol. 8, no. 5, pp. 358-375, October 2007.
- [15] Rabiah Kadir and Behrouz Bokharaeian, "Overview of Biomedical Relations Extraction," *Journal of Industrial and Intelligent Information*, vol. 1, no. 3, september 2013.
- [16] Christian Blaschke, Miguel A Andrade, Christos A Ouzounis, and Alfonso Valencia, "Automatic extraction of biological information from scientific text: protein-protein interactions.," in *Ismb*, vol. 7, 1999, pp. 60-67.
- [17] James Thomas, David Milward, Christos Ouzounis, Stephen Pulman, and Mark Carroll, "Automatic Extraction of Protein Interactions from Scientific," in *Pacific symposium on biocomputing*, vol. 5, 2000, pp. 538-549.
- [18] Toshihide Ono, Haretsugu Hishigaki, Akira Tanigami, and Toshihisa Takagi, "Automated extraction of information on protein--protein interactions from the biological literature," *Bioinformatics*, vol. 17, no. 2, pp. 155-161, 2001.
- [19] Hong-Woo Chun et al., "Extraction of gene-disease relations from Medline using domain dictionaries and machine learning.," in *Pacific Symposium on Biocomputing*, vol. 11, 2006, pp. 4-15.
- [20] Arzucan Ozgur, Thuy Vu, Gunes Erkan, and Dragomir R Radev, "Identifying gene-disease associations using centrality on a literature mined gene-interaction network," *Bioinformatics*, vol. 24, no. 13, pp. i277--i285, 2008.
- [21] Oana Frunza and Diana Inkpen, "Extraction of disease-treatment semantic relations from biomedical sentences," in *BioNLP '10 Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, Uppsala, Sweden, 2010, pp. 91-98.
- [22] Barbara Rosario and Marti A. Hearst, "Classifying Semantic Relations in Bioscience Text," in *proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelona, July 2004.
- [23] Karin M Verspoor, Go Eun Heo, Keun Young Kang, and Min Song, "Establishing a baseline for literature mining human genetic variants and their relationships to disease cohorts," *BMC Medical Informatics and Decision Making*, vol. 16, no. 1, p. 37, 2016.
- [24] ASM Ashique Mahmood, Tsung-Jung Wu, Raja Mazumder, and K Vijay-Shanker, "DiMeX: A Text Mining System for Mutation-Disease Association Extraction," *PloS one*, vol. 11, no. 4, p. e0152725, 2016.

- [25] Kyubum Lee et al., "BRONCO: Biomedical entity Relation ONcology COrpus for extracting gene-variant-disease-drug relations," *Database*, vol. 2016, p. baw043, 2016.
- [26] Gang Li , Karen E. Ross, and Cecilia N. Arighi, "miRTex: A Text Mining System for miRNA-Gene Relation Extraction," *Plos Computational Biology*, 2015.
- [27] Quoc-Chinh Bui, Breannndan Nuallain, Charles A Boucher, and Peter MA Sloot, "Extracting causal relations on HIV drug resistance from literature," *BMC bioinformatics*, vol. 11, no. 1, p. 101, 2010.
- [28] Ting Wang, Kalina Bontcheva, Yaoyong Li, and Hamish Cunningham, "D2. 1.2/Ontology-Based Information Extraction (OBIE) v. 2," *EU-IST Project IST-2003-506826 SEKT SEKT: Semantically Enabled Knowledge Technologies*, 2005.
- [29] I. Segura-Bedmar, P. Martinez, and D. Sanchez-Cisneros, "The 1st DDIEExtraction-2011 Challenge Task: Extraction of Drug-Drug Interactions from Biomedical Texts," *CEUR-WS*, vol. 761, pp. 1-9, 2011.
- [30] D.S. Wishart et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic acids research*, 2007.
- [31] I. Segura-Bedmar, P. Martinez, and M. Herrero-Zazo, "SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIEExtraction 2013)," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, USA, 2013, pp. 341-350.
- [32] Karin Verspoor et al., "Annotating the biomedical literature for the human variome," *Database*, vol. 2013, p. bat019, 2013.
- [33] Emily Doughty, Attila Kertesz-Farkas, Olivier Bodenreider, and Gary Thompson, "Toward an automatic method for extracting cancer- and other," *bioinformatics*, vol. 27, 2011.
- [34] Bach Nguyen and Sameer Badaskar, "A review of relation extraction," *Literature review for Language and Statistics*, vol. II, 2007.
- [35] Ang Sun, Ralph Grishman, and Satoshi Sekine, "Semi-supervised Relation Extraction with Large-scale Word Clustering," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Stroudsburg, PA, USA, 2011, pp. 521-529. [Online]. <http://dl.acm.org/citation.cfm?id=2002472.2002539>
- [36] Oren Etzioni et al., "Web-scale information extraction in knowitall:(preliminary results)," in *Proceedings of the 13th international conference on World Wide Web*, 2004, pp. 100-110.
- [37] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni, "Open information extraction for the web," in *IJCAI*, vol. 7, 2007, pp. 2670-2676.

- [38] Eugene Agichtein and Luis Gravano, "Snowball: extracting relations from large plain-text collections," in *Proceedings of the fifth ACM conference on Digital libraries*, San Antonio, Texas, 2000, pp. 85-94.
- [39] Sergey Brin, "Extracting patterns and relations from the world wide web," in *The World Wide Web and Databases.*: Springer, 1999, pp. 172-183.
- [40] Jiexun Li, Zhu Zhang, Xin Li, and Hsinchun Chen, "Kernel-based learning for biomedical relation extraction," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 5, pp. 756-769, March 2008.
- [41] Claudio Giuliano, Alberto Lavelli, and Lorenza Romano, "Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature," in *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL '06)*, Trento, Italy, #apr# 2006, pp. 401-408. [Online]. <http://acl.ldc.upenn.edu/E/E06/E06-1051.pdf>
- [42] S. V. N. Vishwanathan and Alexander J. Smola, "Fast Kernels for String and Tree Matching," in *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 15*, 2003, pp. 569-576.
- [43] Aliaksei Severyn and Alessandro Moschitti, "Large-scale support vector learning with structural kernels," in *Machine Learning and Knowledge Discovery in Databases.*: Springer, 2010, pp. 229-244.
- [44] S. Pyysalo, A. Airola, J. Heimonen, J. Bjorne, and F. and Salakoski, T. Ginter, "Comparative analysis of five protein-protein interaction corpora," *BMC bioinformatics*, vol. 9, no. Suppl 3, p. S6, 2008.
- [45] Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii, "Corpus annotation for mining biomedical events from literature," *BMC Bioinformatics*, vol. 9, no. 1, p. 10, 2008. [Online]. <http://www.biomedcentral.com/1471-2105/9/10>
- [46] Raymond J Mooney and Razvan C Bunescu, "Subsequence kernels for relation extraction," in *Advances in neural information processing systems*, 2005, pp. 171-178.
- [47] Sophia Katrenko and Pieter Adriaans, "A local alignment kernel in the context of NLP," in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, 2008, pp. 417-424.
- [48] W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan, *Evaluation of Negation Phrases in Narrative Clinical Reports*, 2002.
- [49] G. Szarvas, V. Vincze, R. Farkas, and J. Csirik, "The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, 2008, pp. 38-45.
- [50] Veronika Vincze, Gyorgy Szarvas, Gyorgy Mora, Tomoko Ohta, and Richard Farkas, "Linguistic

- scope-based and biological event-based speculation and negation annotations in the BioScope and Genia Event corpora," *Journal of Biomedical Semantics*, vol. 2, no. Suppl 5, p. S8, 2011. [Online]. <http://www.jbiomedsem.com/content/2/S5/S8>
- [51] N. Konstantinova, S. de Sousa, N.P. Cruz, M. Taboada, and R. Mitkov, "A review corpus annotated for negation, speculation and their scope," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012.
- [52] Roser Morante and Eduardo Blanco, "'SEM 2012 Shared Task: Resolving the Scope and Focus of Negation," in
- [53] Makoto Miwa, Rune Saetre, Yusuke Miyao, and Jun ichi Tsujii, "Entity-focused Sentence Simplification for Relation Extraction," in *Proceedings of the 23rd International Conference on Computational Linguistics*, Stroudsburg, PA, USA, 2010, pp. 788-796. [Online]. <http://dl.acm.org/citation.cfm?id=1873781.1873870>
- [54] Advait Siddharthan, "A survey of research on text simplification," *the International Journal of Applied Linguistics*, pp. 259-98, 2014.
- [55] Yifan Peng, C.O. Tudor, M. Torii, C.H. Wu, and K. Vijay-Shanker, "iSimp: A sentence simplification system for biomedical text," in *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, Oct 2012, pp. 1-6.
- [56] Ryan McDonald, "Extracting relations from unstructured text," *Rapport technique, Department of Computer and Information Science-University of Pennsylvania*, 2005.
- [57] Chek Kim Loi and Jason Miin-Hwa Lim, "Hedging in Academic Writing - A Pedagogically-Motivated Qualitative Study," in *Procedia - Social and Behavioral Sciences*, 2015, pp. 600-607.
- [58] Halil Kilicoglu and Sabine Bergler, "Recognizing speculative language in biomedical research articles: a linguistically motivated perspective," *BMC Bioinformatics*, vol. 9, no. 11, November 2008.
- [59] Marc Light, Xin Ying Qiu, and Padmini Srinivasan, "The Language Of Bioscience: Facts Speculations And Statements In Between," in *Workshop On Linking Biological Literature Ontologies And Databases*, 2004, pp. 17-24.
- [60] Behrouz Bokharaeian, Alberto Diaz, Mariana Neves, and Virginia Francisco, "Exploring Negation Annotations in the DrugDDI Corpus," in *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, 2014, pp. 84-91.
- [61] (2016, July) Nature Education. [Online]. <http://www.nature.com/scitable/definition/phenotype-phenotypes-35>
- [62] Jorge Carrillo de Albornoz, Laura Plaza, Alberto Díaz, and Miguel Ballesteros, "UCM-I: A Rule-based Syntactic Approach for Resolving the Scope of Negation," in *In proceedings of the *SEM*

2012 Shared Task: Resolving the Scope and Focus of Negation., 2012.

- [63] Deyu Zhou, Dayou Zhong, and Yulan He, "Biomedical Relation Extraction: From Binary to Complex," *Computational and Mathematical Methods in Medicine*, vol. 2014, p. 19, August 2014.

Part II

Publications

I'm sure the universe is full of intelligent life. It's just been too intelligent to come here.

Arthur C. Clarke

Chapter

Extracting Drug-Drug Interaction from Text Using Negation Features

Extracting Drug-Drug Interaction from Text Using Negation Features

Estudio del efecto de la negación en la detección de interacciones entre fármacos

Behrouz Bokharaeian, Alberto Díaz

NIL Group
Universidad Complutense de Madrid
Madrid, Spain
behroubo@ucm.es, albertodiaz@fdi.ucm.es

Miguel Ballesteros

Natural Language Processing Group
Universitat Pompeu Fabra
Barcelona, Spain
miguel.ballesteros@upf.edu

Resumen: La extracción de relaciones entre entidades es una tarea muy importante dentro del procesamiento de textos biomédicos. Se han desarrollado muchos algoritmos para este propósito aunque sólo unos pocos han estudiado el tema de las interacciones entre fármacos. En este trabajo se ha estudiado el efecto de la negación para esta tarea. En primer lugar, se describe cómo se ha extendido el corpus DrugDDI con anotaciones sobre negaciones y, en segundo lugar, se muestran una serie de experimentos en los que se muestra que tener en cuenta el efecto de la negación puede mejorar la detección de interacciones entre fármacos cuando se combina con otros métodos de extracción de relaciones.

Palabras clave: Interacciones entre fármacos, negación, funciones kernel, máquinas de vectores de soporte, funciones kernel.

Abstract: Extracting biomedical relations from text is an important task in BioMedical NLP. There are several systems developed for this purpose but the ones on Drug-Drug interactions are still a few. In this paper we want to show the effectiveness of negation features for this task. We firstly describe how we extended the DrugDDI corpus by annotating it with the scope of negation, and secondly we report a set of experiments in which we show that negation features provide benefits for the detection of drug-drug interactions in combination with some simple relation extraction methods.

Keywords: Drug-Drug interaction, Negation, Support vector machines, kernel-based methods

1. Introduction

A drug-drug interaction (DDI) occurs when one drug affects the level or activity of another drug, this may happen, for instance, in the case of drug concentrations. This interaction can result on decreasing its effectiveness or possibly altering its side effects that may even be the cause of health problems to patients (Stockley, 2007).

There is a great amount of DDI databases and this is why health care experts have difficulties to be kept up-to-date of everything published on drug-drug interactions. This fact means that the development of tools for automatically extracting DDIs from biomedical resources is essential for improving and updating the drug knowledge and databases.

There are also many systems on the extraction of biomedical relations from text;

however the research on studying the effect of negation in biomedical relation extraction is still limited. On the other hand, negation is very common in clinical texts and it is one of the main causes of making errors in automated indexing systems (Chapman et al., 2001); the medical personnel is mostly trained to include negations in their reports. Particularly when we are detecting the interaction between drugs, the presence of negations can produce false positives classifications, for instance, the sentence *Co-administration of multiple doses of 10 mg of lenalidomide had no effect on the single dose pharmacokinetics of R- and S- warfarin* a DDI between *lenalidomide* and *warfarin* could be detected as a practicable fact if negation is not taken into account. We therefore believe that detecting the words that

are affected by negations may be an essential part in most biomedical text mining tasks that try to obtain automatically the accurate knowledge from textual data.

In order to avoid errors derived of using automatic negation detection algorithms such as NegEx (Amini et al., 2011), we annotated a DDI corpus - previously developed with the scope of negations. The corpus is called DrugDDI corpus (Segura-Bedmar et al., 2011b), and it was developed for the Workshop on Drug-Drug Interaction Extraction (Segura-Bedmar et al., 2011a) that took place in 2011 in Huelva, Spain. The DrugDDI corpus contains 579 documents extracted from the DrugBank database. We analyzed the corpus and we annotated the sentences within with the scope of negation in order to find the effect of negation features in the detection of DDIs. We annotated it using the same guidelines of the BioScope corpus (Vincze et al., 2008), that is, we annotated cues and scopes affected by negation statements into sentences in a linear format.

For detecting the DDIs we used a fast version of a support vector machine (henceforth, SVM) classifier with a linear kernel based on a bag of words (henceforth, BOW) representation obtained from the extracted features. We carried out some experiments with different kernels (global context, subtree, shortest path), with and without negation information. The results presented in this paper show that negation features can improve the performance of relation extraction methods.

The rest of the paper is structured as follows. In Section 2 we discuss previous related work about biomedical relation extraction and relevant information about the DrugDDI corpus. In Section 3 we explain how we annotated the corpus with the scope of negation. In Section 4 we explain how we used the obtained information from negation tags to improve the DDI detection task. In Section 5 we discuss the results obtained. Finally, in Section 6, we show our conclusions and suggestions for future work.

2. Related work

In this Section we describe the DrugDDI corpus and we present some related work on kernel-based relation extraction.

2.1. DrugDDI corpus

There are some annotated corpora that were developed with the intention of studying biomedical relation extraction, such as, Aimed (Bunesu et al., 2005), LLL (Nedellec et al., 2005), BioCreAtIvE-PPI (Krallinger et al., 2008) on protein-protein interactions (PPI) and DrugDDI (Segura-Bedmar et al., 2011b), on drug-drug interactions. In particular, the DrugDDI corpus is the first annotated corpus on the phenomenon of interactions among drugs and it is the one that we used for our experiments. It was designed with the intention of encouraging the NLP community to conduct further research on this type of interactions. The DrugBank database (Wishart et al., 2008) was used as source to develop this corpus. This database contains unstructured textual information on drugs and their interactions.

The DrugDDI corpus is available in two different formats: (i) the first one contains the information provided by MMTX (Aronson, 2001) and the unified format adapted from PPI corpora format proposed in (Pyysalo et al., 2008). The unified XML format (see Figure 1) does not contain any linguistic information; it only provides the plain text sentences, their drugs and their interactions. Each entity (drug) includes reference (*origId*) to the sentence identifier in the MMTX format corpus. For each sentence contained in the unified format, the annotations correspond to all the drugs entities and the possible DDI candidate pair that represents the interaction. Each DDI candidate pair is represented as a *pair ID* node in which the identifiers of the interacting drugs are registered on its *e1* and *e2* attributes. If the pair is a DDI, the *interaction* attribute is set to *true*, otherwise this attribute is set to *false*. Table 1 shows related statistics of the DrugDDI corpus (Segura-Bedmar et al., 2011b).

2.2. Biomedical Relation Extraction

Nowadays, there are many systems developed for extracting biomedical relations from text that can be categorized in (i) feature based and (ii) kernel-based approaches. Feature-based approaches transform the context of entities into a set of features; this set is used to train a data-driven algorithm. On the other hand, kernel-based approaches are

```

--<sentence id="DrugDDI.d346.s0" origId="s0" text="Uricosuric Agents: Aspirin may decrease the effects of probenecid,
sulfinpyrazone, and phenylbutazone."-->
  <entity id="DrugDDI.d346.s0.e0" origId="s0.p0" charOffset="0-17" type="drug" text="Uricosuric Agents"/>
  <entity id="DrugDDI.d346.s0.e1" origId="s0.p2" charOffset="19-26" type="drug" text="Aspirin"/>
  <entity id="DrugDDI.d346.s0.e2" origId="s0.p6" charOffset="55-65" type="drug" text="probenecid"/>
  <pair id="DrugDDI.d346.s0.p0" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e1" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p1" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e2" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p4" e1="DrugDDI.d346.s0.e1" e2="DrugDDI.d346.s0.e2" interaction="true"/>
</sentence>

```

Figure 1: The unified XML format in the DrugDDI corpus.

	No.
Documents	579
Sentences	5,806
Drugs	8,260
Sentences with at least two drug	3,775
Sentences with at least one DDI	2,044
Sentences with no DDI	3,762
Candidate drug pairs	30,757
Positive interactions	3,160
Negative interactions	27,597

Table 1: Basic statistics for the DrugDDI corpus.

based on similarity functions. This idea provides the option of checking structured representations, such as parse trees and computing the similarity between different representations directly. Combining kernel based and feature based approaches were investigated by Thomas et al. (2011), they developed a voting system (based on majority) that benefits from the outcomes of several methods.

So far, the sequence and tree kernels are the ones that have shown a superior performance for the detection of biomedical relations from text (Bunescu et al., 2005). In particular, global context kernel, subtree and shortest path kernels are three important kernel methods that were applied successfully for biomedical relation extraction task. For instance, Giuliano et al. (2005) applied by considering three different patterns and they calculated the similarity between two sentences by computing common n-grams of two different patterns.

The shortest path kernel (Bunescu y Mooney, 2005) uses the shortest path between two entities (or drugs) in a phrase structure tree. The subtree kernel (Moscitti, 2006) counted the number of common subtrees in whole parse trees by comparing two different sentences. Moreover, a comparative survey about different kernel-based approaches and their performances can be found in (Frunza y Inkpen, 2010).

More recent research on tree kernels were

carried out by Guodong et al. (2010). They introduced a "context-sensitive" convolution tree kernel, which specifies both "context-free" and "context-sensitive" sub-trees by traversing the paths of their ancestor nodes as their contexts to capture structural information in the tree structure. Another motivating work was reported by Chen et al. (2011), that presented a protein-protein interaction pair extractor, it consists on a SVM classifier that exploits a linear kernel with a complete set of features.

Finally, Simões et al. (2013) introduced an approach for Relation Extraction (RE) based on labeled graph kernels, they proposed an implementation of a random walk kernel (Neuhaus y Bunke, 2006) that mainly explores two characteristics: (i) the words between the candidate entities and (ii) the combined information from different sources.

3. Annotating the DrugDDI corpus with negations

We followed the Bioscope guidelines in order to annotate the corpus (Vincze et al., 2008). The main idea is based in the detection of a set of negation cues, like 'no' or 'not'. After this, the scope of the cue is calculated based on its syntactic context. There are several systems that annotate the scope of negation, in our approach we used the one published by Ballesteros et al. (Ballesteros et al., 2012), which is publicly available,¹ rule-based system that works on biomedical literature (Bioscope) and the input is just the sentence without any required annotation, which serves very well for our purposes.

We used as input all the sentences of the DrugDDI corpus, containing 5,806 sentences and 579 files. The output was therefore a set of sentences annotated with the scope of negation. After applying the system, we observed that there were a set of 1,340 sentences containing negations in the corpus,

¹<http://minerva.fdi.ucm.es:8888/ScopeTagger/>

which conforms 23% of the corpus.

Taking into account that the negation scope detection system is fully automatic, we manually checked the outcome correcting the annotations when needed. In order to do so, we divided the annotated corpus in 3 different sets that were assigned to 3 different evaluators. The evaluators checked all the sentences contained in each set and corrected the sentences that contained annotation errors. After this revision, a different evaluator revised all the annotations produced by the other 3 evaluators. Finally, we got the whole set of 1,340 sentences (correctly) annotated with the scope of negation.

The algorithm produced errors -according to the evaluators- in 350 sentences from the 1,340, including false positives matches (there were 16 cases). Which means that 74% of sentences was annotated correctly in an automatic way, when considering a full scope match. The main errors produced by the algorithm were related with the processing of passive voice sentences, commas, and copulative keywords (and, or). In particular the problem of passive voice sentences was related with the pattern *It + to be + not + past participle*, which seems that it was not captured by the system, at least in all cases. The false positives were related with the cue *failure*, which is not a negation when it is a noun modified by an adjective, for instance, *renal failure* or *heart failure*. In the DrugDDI corpus these words appear always as nouns, and therefore all of the performed annotations were incorrect.

The following paragraph shows some examples and corrections made by the evaluators:

- Scope closed in an incorrect way containing words from two different clauses such as: Example: *It is [not] clear whether this represents an interaction with TIKOSYN or the presence of more severe structural heart disease in patients on digoxin;*. The scope should be closed in or.
- Scope closed in an incorrect way in copulative coordinated sentences: Example: *The following medications have been administered in clinical trials with Simulect? with [no] increase in adverse reactions: ATG/ALG, azathioprine, corticosteroids, cyclosporine, mycophe-*

nolate mofetil, and muromonab-CD3.

- Scope opened incorrectly in coordinated copulative sentences: Example: *In an in vitro study, cytochrome P450 isozymes 1A2, 2A6, 2C9, 2C19, 2D6, 2E1, [and 3A4 were{not} inhibited by exposure to cevimeline].*
- Some passive voice sentences were not detected. In particular, as it is already mentioned, sentences with the format 'It (this and that) + finite form of to be + not + past participle'. Example: *[Concomitant use of bromocriptine mesylate with other ergot alkaloids is{not} recommended].*

We also carried out some analysis concerning the number of different cues in the corpus and the number of different errors observed. Table 2 shows that *not* and *no* are by far the most frequent cues in the corpus. It can be observed that the most problematic cue is *neither ... nor*.

Cue	No.	MODFs	Rate
Not	855	266	31.1%
No	439	58	13.2%
without	47	8	17.0%
Neither ... nor	14	12	85.7%
Absence	10	5	50%
Lack	8	1	12.5%
cannot	7	4	57.1%

Table 2: Statistics of negations cues in the corpus and modifications for each cue in the manual checking process.

We finally explored the sentences that are not automatically annotated but they indeed show a negative statement in order to find false negatives. We looked into several negations cues that are not detected by the system such as *unaffected*, *unchanged* or *nonsignificant*. We detected and corrected 75 different sentences that belong to this problem.

Here we show some examples of false negatives:

- *[The pharmacokinetics of naltrexone and its major metabolite 6-beta-naltrexol were {unaffected} following co-administration with Acamprosate].*
- *[Mean T max and mean plasma elimination half-life of albendazole sulfoxide were {unchanged}].*

- *Monoamine Oxidase Inhibition: Linezolid is a reversible, [nonselective] inhibitor of monoamine oxidase.*

Therefore, the corpus finally contains 1,399 sentences annotated with the scope of negation, of which 932 correspond to sentences in which there are at least two drugs mentioned. It is worth mentioning that there are 1,731 sentences with 2 or more drug mentions but no DDI, and 2,044 with 2 or more drugs and at least one interaction.

Finally, the extension of the DrugDDI corpus consists of adding a new tag in the annotation of each sentence with the scope of negations. Figure 3 shows an example. The produced corpus is available for public use.²

4. DDI detection

In this Section, we explain in detail the experiments we carried out by using negation features. First, we illustrate in detail the methods we used without negation features, and then we present our proposed combined negation method, see figure 4. All the experiments were carried out by using the Stanford parser³ for tokenization and constituent parsing (Cer et al., 2010), and the SVMs provided by Weka as training engine.

4.1. DDI detection without negation features

The DDI extraction method consists of four different processes: (1) initial preprocessing, (2) feature extraction, (3) Bag of Words computation and (4) classification. The preprocessing step (1) consists of removing some stop words and tokens, for instance removing question marks at the beginning of the sentence. We also carried out a normalization task for some tokens due to the usage of different encoding and processing methods, mainly HTML tags. In the feature extraction step (2), we extracted three different feature sets corresponding to different used relation extraction methods. The feature extraction step for global context kernel consists of extracting *fore-between*, *between*, and *between-after* tokens that we mentioned in Section 2. The feature extraction step for shortest path kernel method included constituent parsing

of the sentence and then extracting shortest path between two drugs in the generated parse tree. And for the subtree kernel we also extracted all subtrees from the mentioned constituent parse tree. After extracting features, we applied the BOW method (3) to generate new feature sets that the SVM classifier uses. The aim of this step is producing a new representation of the instances which is used in the classification step. And finally in the classification step (4), we applied the Weka SVM classifier (Platt, 1998) (SMO), with a linear composition of features produced by the BOW method to detect the interactions among drugs. The Inner product of new features was used as kernel function between two new representations.

4.2. DDI detection with negation features

In this section, we explain our proposed method that merges negation features with the features mentioned in Section 4.1. We divided the corpus in instances affected by negation and instances without negation statements. The last ones were classified in the same way as in Section 4.1, while for the instances with negations we added negation features to the representation. The positive instances were classified in the same way as previous approaches but the sentences containing negations were categorized using negation features in addition to the other previous features. As in previous subsection, the combined method for instances containing negations consists of 4 steps. After a simple preprocessing step we carried out a feature extraction process. In this step, we generated six negation features in addition to three feature sets corresponding to global context kernel (GCK), Shortest Path and Subtree kernel methods. Negation feature consists of tokens inside the negation scope, left side tokens outside of the negation scope and right side tokens, and the negation cue tokens, negation cue, and position of open and closed negation scope. For instance in the sentence shown in Figure 3: tokens inside brackets create middle scope features, right side tokens construct right features and tokens in the left side of the negation scope form left scope features. As in the previous subsection, we used a BOW method to convert negation string features to word features. Finally, the new feature set is used to classify the drug-drug interactions by

²<http://nil.fdi.ucm.es/sites/default/files/NegDrugDDI.zip>

³<http://nlp.stanford.edu/software/lex-parser.shtml>

```

<sentence origId="s0" id="DrugDDI.d291.s0" text="Zidovudine: There is no significant
pharmacokinetic interaction between ZDV and zalcitabine which has been confirmed
clinically.">
  <entity .... />
  <pair .... />
  <negationtags>Zidovudine: There is <xcope><cue>no</cue> significant pharmacokinetic
  ... clinically</xcope>.</negationtags>
</sentence>

```

Figure 2: A sentence annotated with the Scope of Negation in our version of the DrugDDI corpus.

making use of the Weka SVM.

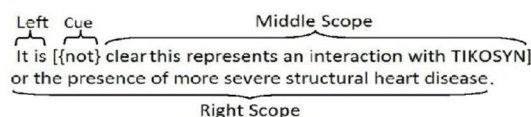


Figure 3: Left, middle and right side scope and negation cue in a negative sentence.

In summation, our approach is a feature based method that uses a bag of word kernel utilizing basic features to compute simple basic kernels and negation features. We applied a fast implementation of the support vector machine provided in Weka, which uses sequential minimal optimization. By carrying out some experiments we also limited the size of the words in each feature bag in the BOW approach to 1000 words per feature class.

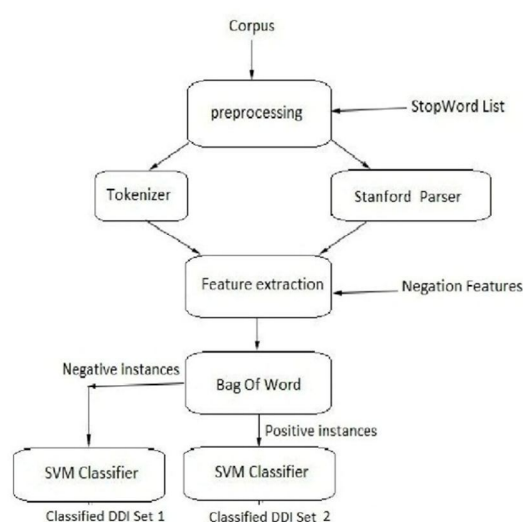


Figure 4: The different processes followed by our proposed approach.

5. Evaluation

5.1. Evaluation Setup

In order to demonstrate the improvements provided by using negation features, our experiments consisted of a 10-fold cross validation over the training part of the DrugDDI corpus. Therefore, our results are not directly comparable to the ones provided in the DDI challenge (Segura-Bedmar et al., 2011a). The training DrugDDI corpus contains 437 documents extracted from Drug-Bank database. It consists of 4267 sentences with average of 9.8 sentences per document and 25,209 instances with 2,402 interactions between different drugs.

Our measurement metrics included true positive, false positive, false negative, total number of positive instances, Precision, Recall and F-1 score.

5.2. Results

Table 3 shows the outcomes of the experiments by computing the metrics and by training over the DDI corpus. The table shows results for Global context kernel (GCK), Subtree and shortest path kernel (SubtreeK) and corresponding combined negation methods (GCKNS = GCK with negation features; SubtreeKNS = Subtree kernel with negation features). The first three rows of the table show the performance of the three basic kernels and the last three ones (with NS postfix) show the outcomes for the combined version that includes negation features. The best result was obtained with GCKNS, and the worst result was obtained by the shortest path tree approach. Moreover, the best improvement was obtained by the GCK approach; it improves 3.8 percentual points of the F score.

As we can see in the table, there is an improved performance when we applied the negation features for classification. This fact

Method	TP	FP	FN	Total	P	R	F1
GCK:	902	1094	1500	2402	0.452	0.376	0.410
SubtreeK:	818	1105	1584	2402	0.425	0.341	0.378
ShortestPathTK:	795	1066	1607	2402	0.427	0.331	0.373
GCKNS:	987	1021	1415	2402	0.492	0.411	0.448
SubtreeKNS:	919	1280	1483	2402	0.418	0.383	0.399
ShortestPathTKNS:	936	1240	1466	2402	0.430	0.390	0.409

Table 3: 10- cross validation results for the methods that do not use negation features and the methods that use negation features.

demonstrates our hypothesis and the emphasizes the purpose of the present work.

6. Conclusions and Future Work

Due to the huge amount of drug related information in bio-medical documents and the importance of detecting dangerous drug-drug interactions in medical treatments, we believe that implementing automatic Drug-Drug interaction extraction methods from text is critical. The DrugDDI corpus is the first annotated corpus for Drug-Drug interaction tasks used in the DDI Extraction 2011 challenge.

In this paper, after reviewing related work on biomedical relation extraction, we first explained the process of annotating the DrugDDI corpus with negation tags; and then we explored the performance of combining negation features with three simple relation extraction methods. Our results show the superior performance of the combined method utilizing negation features over the three basic experimented relation extraction methods.

However, the experiments also show that the application of negation features can indeed improve the relation extraction performance but the obtained improvement clearly depends on the number and rate of positive and negative relations, rate of negative cues in the corpus, and other relation extraction features. It is also true that combining negation features with a huge number of other features may not improve the performance and even may hurt the final result, and this is why we used a limited number of features. It is therefore obvious that corpora having more sentences with negation cues can benefit more from using negation features.

For further work, we plan to use a different type of annotation such as negation events

instead of scopes, and also handling hedge cues and speculative statements in conjunction with negations.

References

- Amini, I., M. Sanderson, D. Martinez y X. Li. 2011. Search for clinical records: Rmit at trec 2011 medical track. En *Proceedings of the 20th Text REtrieval Conference (TREC 2011)*.
- Aronson, A. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. En *Proceedings of the AMIA Symposium*, páginas 17–27. URL <http://metamap.nlm.nih.gov/>.
- Ballesteros, M., V. Francisco, A. Diaz, J. Herrera y P. Gervas. 2012. Inferring the scope of negation in biomedical documents. En *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012)*.
- Bunescu, R., R. Ge, R. Kate, E. Marcotte, R. Mooney, A. Ramani y Y. Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155.
- Bunescu R. y R. Mooney. 2005. A shortest path dependency kernel for relation extraction. En *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, páginas 724–731.
- Cer, D., M. de Marneffe, D. Jurafsky y C. D. Manning. 2010. Parsing to Stanford Dependencies: Trade-offs between speed and accuracy. En *Proceedings of the 7th In-*

- ternational Conference on Language Resources and Evaluation (LREC 2010)*.
- Chapman, W., W. Bridewell, P. Hanbury, G. F. Cooper y B. G. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301-310.
- Chen, Y., F. Liu y B. Manderick. 2011. Extract Protein-Protein Interactions From the Literature Using Support Vector Machines with Feature Selection. *Biomedical Engineering, Trends, Researchs and Technologies*.
- Frunza, O. y D. Inkpen. 2010. Extraction of disease-treatment semantic relations from biomedical sentences. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, páginas 91–98.
- Giuliano, C., A. Lavelli y L. Romano. 2005. Exploiting shallow linguistic information for relation extraction from biomedical literature. En *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, páginas 5–7.
- Guodong, Z., Q. Longhua y F. Jianxi. 2010. Tree kernel-based semantic relation extraction with rich syntactic and semantic information. *International Journal on Information Sciences*, 180(8):1313–1325.
- Krallinger, M., A. Valencia y L. Hirschman. 2008. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biology*, 9(Suppl 2):S8.
- Moschitti, A. 2006. Making Tree Kernels Practical for Natural Language Learning. En *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.
- Nedellec, C. 2004. Machine Learning for Information Extraction in Genomics - State of the Art and Perspectives. *Text Mining and its Applications*, Springer Verlag.
- Neuhaus, M. y H. Bunke. 2006. A Random Walk Kernel Derived from Graph Edit Distance. *Lecture Notes in Computer Science*, 4109(5):191-199.
- Platt, J. 1998. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Advances in kernel methods - Support vector learning*.
- Pyysalo, S., A. Airola, J. Heimonen, J. Bjorne, F. Ginter y T. Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC bioinformatics*, 9(Suppl 3):S6.
- Segura-Bedmar, I., P. Martínez y D. Sánchez-Cisneros. 2011. En *Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*. CEUR Workshop Proceedings, Vol. 761.
- Segura-Bedmar, I., P. Martínez y C. de Pablo Sánchez. 2011. Using a shallow linguistic kernel for drug-drug interaction extraction. *Journal of Biomedical Informatics*, 44(5):789–804.
- Simões, G., D. Matos y H. Galhardas. 2013. A Labeled Graph Kernel for Relationship Extraction. *CoRR*, abs/1302.4874.
- Stockley, I. H. 2007. *Stockley's Drug Interaction*. Pharmaceutical Press.
- Thomas, P., M. Neves, I. Solt, D. Tikk y U. Leser. 2011. Relation extraction for drug-drug interactions using ensemble learning. En *Proceedings of the First Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*, pp:11–17.
- Vincze, V., G. Szarvas, R. Farkas, G. Mora y J. Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. *BMC Bioinformatics*, 9(Suppl 11):S9.
- Wishart, D. R., C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, 36(Suppl 1):D901-D906.

Anybody who has been seriously engaged in scientific work of any kind realizes that over the entrance to the gates of the temple of science are written the words: 'Ye must have faith.'

Max Planck

6

Chapter

Exploring Negation Annotations in the DrugDDI Corpus

Exploring Negation Annotations in the DrugDDI Corpus

Behrouz Bokharaeian*, Alberto Diaz*, Mariana Neves[†], Virginia Francisco*

*NIL Group, Universidad Complutense de Madrid

Madrid, Spain

behroubo@ucm.es, albertodiaz@fdi.ucm.es, virginia@fdi.ucm.es

[†]Hasso-Plattner-Institute at the University of Potsdam, Germany

marianalaraneves@gmail.com

Abstract

Detecting drug-drug interactions (DDI) is an important research field in pharmacology and medicine and several publications report every year the negative effect of combining drugs and chemical treatments. The DrugDDI corpus is a collection of documents derived from the DrugBank database and contains manual annotations for interactions between drugs. We have investigated the negated statements in this corpus and found that they consist of approximately 21% of its sentences. Previous works have shown that considering features related to negation can improve results for the DDI task. The main goal of this paper is to describe the process for annotating the DDI-DrugBank corpus with negation cues and scopes, to show the correlations between these and the DDI annotations and to demonstrate that negations can be used as features for a DDI detection system. Basic experiments have been carried out to show the benefits when considering negations in the DDI task. We believe that the extended corpus can be a significant progress in training and testing algorithms for DDI extraction.

Keywords: Negation detection, Drug-Drug Interaction, Relation extraction

1. Introduction

A drug-drug interaction (DDI) usually occurs when one drug changes the level of activity of another drug. According to FDA's reports and acknowledged surveys (Gurwitz et al., 2000), over 2 million serious Adverse Drug Reactions (ADRs) occur in the United States every year, including the register of one hundred thousand deaths (Lazarou et al., 1998). Moreover, 3.5% of these deaths are due to drug-drug interaction (Martin, 1990). Detecting and identifying interactions between drugs is a crucial field of research given the high risks of most drug-drug interactions and the importance of patient safety and health care cost control. Many academic researchers and pharmaceutical companies have developed databases where DDI are recorded, but most of the research and valuable information is still only found in unstructured text documents, such as scientific publications and technical reports.

Information extraction is an important task in natural language processing (NLP) and has also been used in many applications in the biomedical domain, ranging from simple binary relationships to complex and hierarchical relation extraction (McDonald et al., 2005). Recent research on biomedical information extraction has focused on biological entities and relationships, since many annotated corpora are available for this purpose, which are valuable resources for repeatable automatic training and evaluation of NLP techniques. For instance, several corpora have been annotated for protein-protein or gene-protein interactions, such as Aimed (Bunescu et al., 2005), LLL (Nedellec, 2005), IEPA (Ding et al., 2002) or BioCreAtivE-PPI (Krallinger et al., 2008)).

A DrugDDI corpus was initially developed by (Segura-Bedmar et al., 2011a) based on a set of 579 xml files describing DDIs which was randomly collected from the DrugBank database (Wishart et al., 2007). The UMLS MetaMap (MMTx) tool (Aronson, 2001) was used to anal-

yse the corpus and was manually annotated with the help of pharmacist experts (DDI 2011 corpus). With the aim of encouraging researchers to explore new methods for extracting drug-drug interactions, the first DDI Extraction challenge task¹ was held in 2011 with the participation of ten teams (Segura-Bedmar et al., 2011b). The best results were an F-measure of 65.74%, a precision of 65.04% and a recall of 71.92% in detecting and classifying DDIs (Thomas et al., 2011). A second challenge was held on 2013 as part of SemEval: the DDI Extraction2013². A new corpus was developed which included the corpus used in 2011 (DDI-DrugBank 2013) as well as Medline abstracts. Participating teams developed solutions based on supervised and sentence-level relation extraction methods and the best F1 score obtained was 80%. According to Segura and her colleagues, increasing the size of the corpus and optimizing the quality of annotations have contributed to this improvement (Segura-Bedmar et al., 2013).

The DDI-DrugBank 2013 corpus is a useful resource for performing comparable experiments and for investigating relation extraction methods. However, one limitation of this corpus is the lack of negation annotation. For instance, in the sentence below, an interaction between *itraconazole* and *S-ketamine* drugs could be identified if negation is ignored.

Ticlopidine treatment increased the mean area under the plasma concentration-time curve extrapolated to infinity (AUC(0-)) of oral ketamine by 2.4-fold, whereas itraconazole treatment did not increase the exposure to S-ketamine. Negation is frequently used in clinical and biomedical documents and it is an important cause of low precision in automated indexing systems (Chapman et al., 2002). For instance, Chapman has observed that 95% to 99% of the searched reports would state *no signs of fracture* or similar

¹<http://labda.inf.uc3m.es/DDIExtraction2011/>

²<http://www.cs.york.ac.uk/semeval-2013/task9/>

expressions in a certain radiology report database (Chapman et al., 2002). As a result, identifying negative statements is an important task to obtain accurate knowledge from textual data.

In previous work (Bokharaeian et al., 2013), the DDI 2011 corpus was annotated with negations (NegDrugDDI corpus) and a basic experiment showed that improvements in drug-drug interactions can be obtained when considering annotations for negations. Additionally, the best-scoring team in the DrugDDI 2013 challenge also use negation information in their system (Chowdhury et al., 2013).

In this paper, we describe the annotation of the DDI-DrugBank 2013 corpus with negation cues and scopes, the hereafter called NegDDI-DrugBank 2013, following the BioScope guidelines (Szarvas et al., 2008). We also present the correlations between the negation annotations and the position of the drugs in a sentence. Finally, we have performed some experiments with the TEES event extraction tool (Bjorne and Salakoski, 2013) to confirm the positive effect of the negation annotations for the DDI task.

In Section 2, we present related work on previous corpora annotated with negation, while Section 3 describes corpora annotated with drug-drug interactions. In Section 4, the annotation process and the obtained results are described. Section 5 presents the correlations between DDI and negation that have been found in the extended corpus while Section 6 shows the experiments carried out to confirm the effects of negation annotations for the DDI task. Finally, Section 7 presents discussions and suggestions for future works.

2. Corpora annotated with negation

In this section, we review the main corpora annotated with negation, emphasizing the annotation guidelines that were followed and the main differences between them.

2.1. Bioscope

Bioscope³ (Szarvas et al., 2008) is an open access corpus of biomedical documents, manually annotated with negation and speculation. It contains more than 20,000 sentences which are split in three collections: clinical documents (6,383 sentences, 863 with negations), scientific papers (2,670 sentences, 339 with negations) and scientific abstracts (11,871 sentences, 1,597 with negations). All the sentences which assert the non-existence of something are annotated, including sentences which do not contain any biomedical term. Each negated sentence is annotated with information about the negation cue and the scope of negation.

The annotation of Bioscope followed a min-max strategy: the minimal unit that expresses negation is considered the negation cue (min strategy) and the scope is extended to the largest syntactic unit possible (max strategy). The negation cue is always included in the scope. However, it is worth emphasizing that when the scope is opened at the cue and continues to the right of the cue (around 90% of the cases), the scope affected by the cue leaves the subject out. This corresponds to sentences in active voice which

are the most frequent case. Additionally, there are cases in which the scope is opened to the left of the cue. The most frequent ones are the structures in passive voice. As shown in (Szarvas et al., 2008), passive voice is an exception in the way of tagging sentences in Bioscope. In this case, the subject is annotated within the scope, because if the sentence had been written in active voice, it would have been the object of a transitive verb.

2.2. SFU Review Corpus

SFU Review Corpus (Konstantinova et al., 2012) is a freely available corpus annotated with negation and speculation. It consists of 400 documents of movie, book and consumer product reviews. It is annotated with negative and speculative keywords and their scope. The entire corpus was manually annotated by one linguist and reviewed by another one. The guidelines followed during the annotation was an adaptation of Bioscope guidelines, which main changes were:

- Negation cues were not included in the scope.
- Coordination was annotated in a different way.
- Sentences with negation cue and without scope were possible.

2.3. ConanDoyle-neg

ConanDoyle-neg (Morante and Daelemans, 2012) was released in conjunction with the 2012 shared task on NR hosted by The First Joint Conference on Lexical and Computational Semantics (*SEM 2012). It is a corpus of Arthur Conan Doyle's stories manually annotated with negation cues and their scope. The annotation was performed by two annotators using the Salto Tool.

The following is annotated in each sentence which contain negation statements: the negation cue, its scope and the negated event. For example, in the sentence "*After mine I asked no questions*" *no* is identified as the negation cue, *after mine I asked questions* is identified as the scope and *asked* is the negated event.

This corpus annotation was inspired by the guidelines of Bioscope, but with several differences, being the following the most important ones:

- The negated event is annotated.
- Negation cues are not included in the scope.
- Scopes can be discontinuous.
- All arguments of the negated event are included in the scope, including the subject (which in Bioscope corpus was kept out in active sentences).
- Affixal cues are annotated. If the scope of a negation cue is not explicit, the negation cue is marked as such, but the scope is not annotated. If the scope is recoverable from the same sentence, it is added to the scope.

The domain of this corpus is very restrictive so some constructions that are typical in other domains are left out. For instance, constructions that express absence of an entity, which are very frequent in biomedical texts, are not included in this corpus.

³<http://www.inf.u-szeged.hu/rgai/bioscope>

2.4. UAM Spanish Treebank

UAM Spanish Treebank (Moreno Sandoval et al., 2003) is a corpus composed of 1,501 syntactically annotated sentences derived from Spanish newspapers. The syntactic annotation was extended with annotations for negation. Annotation of negation was carried out by two experts in Corpus Linguistics. The annotation guidelines were very similar to those of Bioscope, except for one main difference: all arguments of the negated events are included in the scope, including the subject (which were kept out in active sentences in the Bioscope corpus).

3. Drug-drug interaction annotation

The DDI 2011 corpus was the first annotated corpus dealing with the interaction phenomenon between drugs. The corpus was designed by (Segura-Bedmar et al., 2011a) in order to encourage the NLP community to conduct further research in the field of pharmacology. A set of 579 xml files describing DDIs was randomly collected from the DrugBank database (Wishart et al., 2007). The corpus was analyzed by the UMLS MetaMap tool (MMTx) (Aronson, 2001) and was manually annotated with the help of pharmacist experts.

This corpus is provided in the unified format used for PPI corpora proposed in (Pyysalo et al., 2008) (see Figure 1). Each entity (drug) includes reference (*origId*) to the id phrase in the MMTX format corpus text in which the corresponding drug appears. For each sentence in the corpus, all DDI candidate pairs are generated from the possible combination of different drugs appearing therein. Each DDI candidate pair is represented as a *pair* node in which the ids of the interacting drugs are registered in its *e1* and *e2* attributes. If the pair is a DDI, the *interaction* attribute must be set to *true*, otherwise this attribute must be set to *false*.

The DDI-DrugBank 2013 corpus was developed for the DDI Extraction 2013 SemEval task and includes part of the the DDI 2011 corpus. Concretely, new documents were annotated from the DrugBank database and were used for the test dataset (DDI-DrugBank Test 2013 corpus), while 572 documents from the previous corpus were used as training dataset (DDI-DrugBank Train 2013 corpus). Therefore, the DDI-DrugBank 2013 corpus contains a total of 730 documents. A dataset of 233 MedLine abstracts (DDI-MedLine 2013 corpus) was also annotated for the 2013 shared task, however, in this work we have concentrated on the DrugBank documents.

Table 1 shows basic statistics of the DDI-DrugBank 2013 corpus. It contains 6,648 sentences with 9.1 sentences per document on average. The average number of drug mentions per document was 21.15, and the average number of drug mentions per sentence was 2.4. Finally, among the 31,270 candidate drug pairs, only 4,672 (14.94%) were annotated as positive interactions, (i.e., DDIs), while 26,598 (85.06%) were marked as negative interactions (i.e., non-DDIs). There is a much larger proportion of negative instances than positive ones.

All drug-drug interactions in the DDI-DrugBank 2013 corpus was also annotated with one of the following four interaction types: *advice*, *effect*, *mechanism* and *int*. The *advice* type corresponds to an advice or recommendation

regarding the concomitant use of the two drugs, the *effect* category refers to the effect of DDIs, the *mechanism* type were assigned to DDIs which describe pharmacodynamic or pharmacokinetic mechanism and the default *int* category is used otherwise. More detailed definition of the types can be found at (Segura-Bedmar et al., 2013). With respect to the distribution of categories, as can be seen in Table 2, there is a smaller number of instances for categories *int* and *advice* and *effect* type is the most frequent.

4. Annotating DDI-DrugBank corpus with negation

The aim of this paper is to extend a drug-drug interactions corpus (DDI-DrugBank 2013) with annotations for negation, the NegDDI-DrugBank 2013, as none of the existing corpora meets this requirement. All the sentences in the original corpus were annotated, which conforms 6,648 sentences from 730 files. For the DDI DrugBank 2013 training dataset, annotations from the NegDrugDDI corpus (Bokharaeian et al., 2013) have been transferred to the NegDDI-DrugBank 2013 corpus and then reviewed, given that there were some discrepancies between the documents from the two DDI editions.

For the DDI DrugBank 2013 test dataset, a first annotation was done with the rule based system (Ballesteros et al., 2012), which follows the BioScope guidelines to annotate sentences with negation. The annotation consisted on adding two new tags, the cue and the scope of the negations, as depicted in Figure 3. The pre-annotation automatically obtained was then reviewed by four annotators using the Brat NLP annotation tool⁴. Brat is a web based software tool which was developed for rich annotating which has proven to decrease the annotation time and to increase the quality of the resulting annotations (Stenetorp et al., 2012). A screenshot of the NegDDI-DrugBank 2013 corpus as visualized in the tool is shown in Figure 2. The test dataset was split in four parts, one for each annotator, who have manually corrected the automatically generated annotations, whenever necessary, and have added the missing ones. Subsequently, the more experienced annotator reviewed all the annotations to ensure coherence. According to the annotators, 18 modifications have been done. That is, the algorithm have annotated the majority of the sentences correctly. The extended corpus is available for public use⁵. We have performed an analysis on the number of distinct cues in the entire NegDDI-DrugBank 2013 corpus and the number of different problematic annotation that were observed. This analysis is shown in Table 3. As can be seen in this table, *not* and *no* are by far the most frequent cues in the corpora: 1018 and 498 occurrences. However, more changes have been performed with cue *not*, 27.41% of changes. On the other hand, it can be observed that the most problematic cue is *neither ... nor ...*, with a 85.71% of changes. It is due to the difficult double cue pattern associated to this cue. Most of the errors with the other cues are associated with problems detecting certain patterns of pas-

⁴<http://brat.nlplab.org/>

⁵http://nil.fdi.ucm.es/sites/default/files/NegDDI_DrugBank.zip


```

-<sentence id="DrugDDI.d346.s0" origId="s0" text="Uricosuric Agents: Aspirin may decrease the effects of probenecid,
sulfinpyrazone, and phenylbutazone.">
  <entity id="DrugDDI.d346.s0.e0" origId="s0.p0" charOffset="0-17" type="drug" text="Uricosuric Agents"/>
  <entity id="DrugDDI.d346.s0.e1" origId="s0.p2" charOffset="19-26" type="drug" text="Aspirin"/>
  <entity id="DrugDDI.d346.s0.e2" origId="s0.p6" charOffset="55-65" type="drug" text="probenecid"/>
  <entity id="DrugDDI.d346.s0.e3" origId="s0.p7" charOffset="67-81" type="drug" text="sulfinpyrazone"/>
  <entity id="DrugDDI.d346.s0.e4" origId="s0.p9" charOffset="87-101" type="drug" text="phenylbutazone"/>
  <pair id="DrugDDI.d346.s0.p0" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e1" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p1" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e2" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p2" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e3" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p3" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e4" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p4" e1="DrugDDI.d346.s0.e1" e2="DrugDDI.d346.s0.e2" interaction="true"/>
  <pair id="DrugDDI.d346.s0.p5" e1="DrugDDI.d346.s0.e1" e2="DrugDDI.d346.s0.e3" interaction="true"/>
  <pair id="DrugDDI.d346.s0.p6" e1="DrugDDI.d346.s0.e1" e2="DrugDDI.d346.s0.e4" interaction="true"/>
  <pair id="DrugDDI.d346.s0.p7" e1="DrugDDI.d346.s0.e2" e2="DrugDDI.d346.s0.e3" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p8" e1="DrugDDI.d346.s0.e2" e2="DrugDDI.d346.s0.e4" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p9" e1="DrugDDI.d346.s0.e3" e2="DrugDDI.d346.s0.e4" interaction="false"/>
</sentence>

```

Figure 1: The unified XML format of a sentence in the DrugBank-DDI 2013 corpus.

	Number	Avg. per document
Documents	730	
Sentences	6,648	9.11
Entities	15,441	21.15
Candidate drug pairs	31,270	42.84 (4.70 per sentence)
Positive interactions (DDIs)	4,672	6.40 (14.94%)
Negative interactions (no DDIs)	26,598	36.44 (85.06%)

Table 1: Basic statistics of the DDI-DrugBank 2013 corpus.

Training	pairs	negative DDIs	positive DDIs	effect	mechanism	advice	int
DrugBank	26005	22217	3788	1535	1257	818	178
Test	pairs	negative DDIs	positive DDIs	effect	mechanism	advice	int
DrugBank	5265	4381	884	298	278	214	94

Table 2: Statistics of the training and test datasets of the DDI-DrugBank 2013 corpus.

1	Cholestyramine: Concomitant cholestyramine administration decreased the mean AUC of total ezetimibe approximately 55%.
2	The incremental LDL-C reduction due to adding ezetimibe to cholestyramine may be reduced by this interaction.
3	Fibrates: scope The safety and effectiveness of ezetimibe administered with fibrates have cue not been established.
4	Fibrates may increase cholesterol excretion into the bile, leading to cholelithiasis.
5	In a preclinical study in dogs, ezetimibe increased cholesterol in the gallbladder bile.
6	Co-administration of ZETIA with fibrates scope is cue not recommended until use in patients is studied.
7	Fenofibrate: In a pharmacokinetic study, concomitant fenofibrate administration increased total ezetimibe concentrations approximately 1.5-fold.
8	Gemfibrozil: In a pharmacokinetic study, concomitant gemfibrozil administration increased total ezetimibe concentrations approximately 1.7-fold.
9	HMG-CoA reductase inhibitors: cue No clinically significant pharmacokinetic interactions were seen when ezetimibe was co-administered with atorvastatin, simvastatin, pravastatin, lovastatin, or fluvastatin.
10	Cyclosporine: The total ezetimibe level increased 12-fold in one renal transplant patient receiving multiple medications, including cyclosporine.

Figure 2: Examples of negation cue and scope annotations.

Cue	DDI-DrugBank Train	Changes	DDI-DrugBank Test	Changes	DDI-DrugBank	Total	Rate
not	855	266	163	13	1018	279	27.41%
no	439	58	59	1	498	59	11.85%
without	47	8	9	4	56	12	21.43%
neither ... nor ...	14	12	0	0	14	12	85.71%
absence	10	5	3	0	13	5	38.46%
lack	8	1	0	0	8	1	12.50%
cannot	7	4	3	0	10	4	40.00%
Total	1380	354	237	18	1617	372	23.01%

Table 3: Statistics of the negative cues in the training and test datasets, the changes for each cue during manual checking and the rate of changes, for the NegDDI-DrugBank 2013.

```

<sentence id="DDI-DrugBank.d297.s4" text="Concurrent therapy with ORENCIA and TNF antagonists is not recommended.">
  <entity charOffset="24-30" id="DDI-DrugBank.d297.s4.e0" text="ORENCIA" type="brand"/>
  <entity charOffset="36-50" id="DDI-DrugBank.d297.s4.e1" text="TNF antagonists" type="group"/>
  <pair ddi="true" e1="DDI-DrugBank.d297.s4.e0" e2="DDI-DrugBank.d297.s4.e1" id="DDI-DrugBank.d297.s4.p0" type="advise"/>
  <negationtags><xscope> Concurrent therapy with ORENCIA and TNF antagonists is <cue>not</cue>
    recommended</xscope>.</negationtags>
</sentence>

```

Figure 3: The extended unified XML format of a sentence with negation cue in NegDDI-DrugBank corpus.

sive voice sentences and with the bad processing of commas and copulative keywords.

5. Analysis of correlations between negations and DDI annotations

NegDDI-DrugBank 2013 corpus contains 1,448 sentences with at least one negation scope, which correspond to 21.78% of the sentences (4). This confirms the statement that negation is frequently used in clinical and biomedical documents, and particularly, in pharmacological documents describing drug activity.

Table 5 shows the correlations between the annotations for negation scopes and the position of the two candidate drugs that represent a DDI. The first two columns indicate the position of the drugs, there are 5 possibilities:

- both drugs inside of the negation scope (inside, inside).
- both drugs outside of the negation scope but on the right hand side of the sentence (right, right).
- both drugs outside of the negation scope but on the left hand side of the sentence (left, left).
- one drug inside the negation scope but the other one outside on the right-hand side (inside, right).
- one drug inside the negation scope but the other one outside on the left hand side (inside, left).

For instance, Figure 3 shows a sentence with a negation cue and two drug which are both inside of the negation scope. We can conclude from this data that, in the majority of the cases (around 90%), there is no DDI when a negation scope is present. With respect to the position of the drugs, the best correlation occurs when both drugs are inside the negation scope (93.78%), while the worst correlation occurs when one drug is inside and the other one is outside or on the right hand side (88.43%).

The correlation between DDI type and drug positions compared to negation scope has also been analyzed. As Table 6 confirms, there is a clear correlation between the DDI type and relative candidate drug positions to negation scope. The highest correlation can be seen when both candidate drugs are inside the negation scope and DDI type is *advice* (78.65% of all *advice* type cases with negation cue mention a positive DDI). For instance in the below sentence:

<xscope>It is recommended that the combination of intravenous **dantrolene sodium** and calcium channel blockers, such as **verapamil**, <cue>not</cue>be used together during the management of malignant hyperthermia crisis

until the relevance of these findings to humans is established.</xscope>

The candidate drugs (*dantrolene sodium* and *verapamil*) are both inside of the negation scope and the *advice* type was assigned to the DDI.

The other three DDI types (*effect*, *mechanism* and *int*) have a similar behavior regarding the correlation between DDI type and candidate drug positions. For instance, for all of them, percentages are low (*effect*= 4.49%, *mechanism*=16.8% and 0% for *int*) when two candidate drugs are inside the scope.

Table 7 shows the average of correlations between the DDI type and candidate drug positions. As can be seen there is a significant difference between *advice* type and the other three DDI types. The 53.87% of the sentences with negation that contains a positive DDI correspond to *advice* type and the 1.3% of the sentences with negation that contains a positive DDI correspond to *int* type.

We can conclude that the position of entities regarding the scope of negation is an important factor in determining the effect of negation and the candidate DDIs.

On the other hand, regarding to Drug-Drug Interaction relation, *recommended* and *advised* words have negative polarity. In fact *recommendation* is used to avoid co-administration of two drugs, instead of recommending them, but *effect*, *excretion* and *interact* phrases have positive polarities. Consequently, classifying and extracting positive DDIs should consider these important factors in addition to other syntactic factors that are usually employed.

Our analysis shows that we need semantic and polarity-based processing to efficiently employ negation information in relation extraction task. For instance, the two sentences below are in passive voice and they have similar length and annotations for negations. The first one mentions a drug-drug interaction in an advisory notion, while the second one explains a mechanism for possible drug interaction, but does not mention a DDI. In both sentences, two drug names are inside the negation scope and related verb and adverbs are also inside of scope. These two sentences are good examples that deep and semantic processing are needed to employ negation in detecting positive drug-drug interactions.

- <xscope>Concurrent therapy with ORENCIA and TNF antagonists is <cue>not</cue> recommended.</xscope>
- <xscope>This small decrease in ec of gabapentin by cimetidine is <cue>not</cue> expected </xscope> to be of clinical importance.

	Number	Percentage (%)
Documents	730	
Sentences	6,648	
Sentences with negation	1,448	21.78
Sentences without negation	5200	78.22

Table 4: Basic statistics from the NegDDI-DrugBank 2013 corpus

Drug1position	Drug2position	DDI	Train	Test	Total	Percentage (%)
inside	inside	false	613	730	1343	93.78
inside	inside	true	39	50	89	6.63
left	left	false	141	1191	1332	89.82
left	left	true	27	124	151	11.34
right	right	false	101	819	920	92.56
right	right	true	12	62	74	8.04
inside	left	false	256	921	1177	92.31
inside	left	true	6	92	98	8.33
inside	right	false	52	437	489	88.43
inside	right	true	7	57	64	13.09

Table 5: Correlations between DDI and drug positions compared to negation scope for the NegDDI-DrugBank 2013 corpus. The third column indicates if the candidate DDI associated with the annotation is true or false. The fourth and fifth columns indicate if the correlation appears in the training or in the test dataset. Finally, the last column indicates the total of possible correlations of each type and the corresponding percentage.

Drug1position	Drug2position	Type	Total	Percentage (%)
inside	inside	advise	70	78.65
left	left	advise	50	33.11
right	right	advise	24	32.43
inside	right	advise	37	57.81
inside	left	advise	66	67.34
inside	inside	effect	4	4.49
left	left	effect	56	37.08
right	right	effect	14	18.91
inside	left	effect	26	26.53
inside	right	effect	15	23.43
inside	inside	mechanism	15	16.85
left	left	mechanism	44	29.13
right	right	mechanism	34	45.94
inside	left	mechanism	6	6.12
inside	right	mechanism	10	15.62
inside	inside	int	0	0
left	left	int	1	0.66
right	right	int	2	2.7
inside	left	int	0	0
inside	right	int	2	3.12

Table 6: Correlations between positive DDI and drug position compared to negation scope. The last columns show the total of possible correlations of each type and the corresponding percentage.

Type	Total	Average (%)
advise	247	53.87
effect	115	22.01
mechanism	129	26.81
int	5	1.3

Table 7: Total average correlations between DDI type and candidate drug positions

6. Exploring negation features

In addition to extending the DDI-DrugBank 2013 corpus, we carried out experiments using the version 2.1 of TEES

event extraction software tool⁶ to verify the effects of the negation annotations in a relation extraction task. TEES is a well known machine-learning based tool for extract-

⁶<http://jbjorne.github.io/TEES/>

ing text-bound graphs from natural language text and has shown successful performance in many binary relationship and event extraction tasks (Bjorne and Salakoski, 2013).

TEES supports negation detection using the schema used in the BioNLP Genia event Extraction tasks⁷, where a negation attribute is assigned to the event, but no cue or scope are annotated. When performing our experiments with TEES, we have added the negation cues and scopes as additional entities with the corresponding entity types ("cue" or "xscope").

We have carried out experiments only with the training dataset (NegDDI-DrugBank Train 2013), i.e., training and testing on the 572 documents dataset (90% and 10%, respectively), and using the complete corpus (NegDDI-DrugBank 2013), i.e., the training dataset of the 2013 edition for training (i.e., 572 documents) and testing on the test dataset of the 2013 edition (i.e., 158 documents). For each of these experiments, we considered three situations: the original corpus without any negation annotations, the addition of only the negation cues and also considering the scope annotations. Results are presented in Table 8. Results show that we get different responses for each of the experiments when considering the negation annotations, nonetheless, both of them positive. When using only the training dataset, the number of true positives does not change much, but TEES returns less false positives, i.e., higher precision, without the degradation of the recall, thus, also with an improvement on the F-score. However, contrary to what was expected, the negation annotations had more effect on the recall for the 2013 test dataset, i.e., decrease of false negatives, instead on the precision, thus the additional true positives. A future error analysis on the results returned by TEES will shed some light on this behavior and give use some insights on how to better use negation annotations for drug-drug interactions.

7. Conclusions and Future Work

We have annotated the DDI-DrugBank 2013 corpus with annotations for negation following BioScope guidelines. It consists of 730 files with 6,648 sentences extracted from the DrugBank database. The extended corpus (NegDDI-DrugBank 2013) contains 1,448 sentences with at least one negation scope. This is the 21.78% of the sentences, confirming the tendency of use of negation expressions on biomedical documents.

We have computed correlations between the DDI and negation annotations present in the corpus. We can conclude that all these effective factors should be considered as potential features for a machine learning based method or in combination with a rule based system for extracting positive DDI from sentences with negation.

We plan to continue exploring the effect of features extracted from negation annotations in the DDI task, given the promising results which have been obtained in the preliminary experiments carried out with TEES, which explored only indirectly the potentials of negation cue and scope annotations.

⁷<http://bionlp.dbcls.jp/redmine/projects/bionlp-st-ge-2013/wiki/Wiki>

Additionally, we plan to extend the annotations to the DDI-MedLine 2013 corpus. We expect differences in these annotations due of the language used in scientific publications.

Finally, we have used BioScope guidelines to annotate the NegDDI-DrugBank 2013 corpus but we plan to explore other guidelines as well, such as the one considered in *SEM conference (Morante and Blanco, 2012).

8. Acknowledgements

The authors would like to acknowledge Prof. Ulf Leser (Humboldt-Universität zu Berlin) for the use of software resources. MN would like to acknowledge funding from the HPI Research School.

9. References

- A.R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.
- M. Ballesteros, V. Francisco, A. Díaz, Herrera J., and P. Gervás. 2012. Inferring the scope of negation in biomedical documents. In *13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012)*, New Delhi. Springer, Springer.
- J. Bjorne and T. Salakoski. 2013. Tees 2.1: Automated annotation scheme learning in the bionlp 2013 shared task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 16–25.
- B. Bokharaeian, A. Diaz, and M. Ballesteros. 2013. Extracting drug-drug interaction from text using negation features. *Procesamiento del Lenguaje Natural*, 51:49–56.
- R. Bunesco, R. Ge, R.J. Kate, E.M. Marcotte, R.J. Mooney, A.K. Ramani, and Y.W. Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155.
- W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. 2002. Evaluation of negation phrases in narrative clinical reports.
- Md. Chowdhury, M. Faisal, and A. Lavelli. 2013. 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*, pages 765–771, Atlanta, Georgia, June. Association for Computational Linguistics.
- J. Ding, D. Berleant, D. Nettleton, and E. Wurtele. 2002. Mining MEDLINE: Abstracts, sentences, or phrases? In *Pacific Symposium on Biocomputing*, page 326. World Scientific Publishing Company.
- J.H. Gurwitz, T.S. Field, J. Avorn, D. McCormick, S. Jain, M. Eckler, M. Benser, A.C. Edmondson, and D.W. Bates. 2000. Incidence and preventability of adverse drug events in nursing homes. *The American Journal of Medicine*, 109:87–94.

	true positive	false positive	false negative	precision	recall	F-score
NegDDI-DrugBank Train 2013	225	63	172	78.12	56.67	65.69
NegDDI-DrugBank Tain 2013+Cue	226	54	171	80.71	56.92	66.76
NegDDI-DrugBank Train 2013+Cue+scope	226	53	171	81.00	56.92	66.86
NegDDI-DrugBank 2013	602	173	278	77.67	68.40	72.74
NegDDI-DrugBank 2013+Cue	612	172	271	78.06	69.30	73.42
NegDDI-DrugBank 2013+Cue+Scope	618	175	265	77.93	69.98	73.74

Table 8: Results obtained from TEES for drug-drug interaction predictions using different configurations of the NegDDI-Drugbank corpus.

- N. Konstantinova, S. de Sousa, N.P. Cruz, M.J. Maa, M. Taboada, and R. Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia. 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9(Suppl 2):S4.
- J. Lazarou, B.H. Pomeranz, and P.N. Corey. 1998. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. *JAMA*, 279(15):1200–1205.
- L.E. Martin. 1990. Knowledge extraction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*.
- R. McDonald, F. Pereira, S. Kulick, S. Winters, Y. Jin, and P. White. 2005. Simple algorithms for complex relation extraction with applications to biomedical ie. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05*, pages 491–498, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Roser Morante and Eduardo Blanco. 2012. *sem 2012 shared task: Resolving the scope and focus of negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274, Montréal, Canada, 7-8 June. Association for Computational Linguistics.
- R. Morante and W. Daelemans. 2012. Conandoyle-neg: Annotation of negation in conan doyle stories. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*.
- A. Moreno Sandoval, S. Lopez Ruesga, F. Sanchez, and R. Grishman. 2003. Developing a spanish treebank. *Building and using parsed corpora*, pages 149–163.
- C. Nedellec. 2005. Learning language in logic-genic interaction extraction challenge. In *Proceedings of the ICML05 workshop: Learning Language in Logic (LLL05)*, volume 18, pages 97–99. Citeseer.
- S. Pyysalo, A. Airola, J. Heimonen, J. Bjorne, F. Ginter, and T. Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC bioinformatics*, 9(Suppl 3):S6.
- I. Segura-Bedmar, P. Martínez, and C. de Pablo-Sánchez. 2011a. Using a shallow linguistic kernel for drug-drug interaction extraction. *Journal of Biomedical Informatics*, In Press, Corrected Proof.
- I. Segura-Bedmar, P. Martínez, and D. Sánchez-Cisneros. 2011b. The 1st ddiextraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts. *CEUR-WS*, 761:1–9.
- I. Segura-Bedmar, P. Martinez, and M. Herrero-Zazo. 2013. 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: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, USA. Association for Computational Linguistics.
- P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. Brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, April. Association for Computational Linguistics.
- G. Szarvas, V. Vincze, R. Farkas, and J. Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45. Association for Computational Linguistics.
- P. Thomas, M. Neves, I. Solt, D. Tikk, and U. Leser. 2011. Relation extraction for drug-drug interactions using ensemble learning. In *Proceedings of the First Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*, pages 11–18.
- D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. 2007. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*.

The mediocre teacher tells. The good teacher explains. The superior teacher demonstrates. The great teacher inspires.
William Arthur Ward

Chapter

Extracting Drug-Drug Interactions from Text through Combination of Sequence and Tree Kernels

NIL-UCM: Extracting Drug-Drug interactions from text through combination of sequence and tree kernels

Behrouz Bokharaeian, Alberto Díaz

Natural Interaction Based on Language Group

Universidad Complutense de Madrid

Madrid 28011, Spain

{bokharaeian, albertodiaz}@fdi.ucm.es

Abstract

A drug-drug interaction (DDI) occurs when one drug affects the level or activity of another drug. Semeval 2013 DDI Extraction challenge is going to be held with the aim of identifying the state of the art relation extraction algorithms. In this paper we firstly review some of the existing approaches in relation extraction generally and biomedical relations especially. And secondly we will explain our SVM based approaches that use lexical, morphosyntactic and parse tree features. Our combination of sequence and tree kernels have shown promising performance with a best result of 0.54 F1 macroaverage on the test dataset.

1 Introduction

A drug-drug interaction occurs when one drug affects the level or activity of another drug, for instance, drug concentrations. This interaction can result on reducing its effectiveness or possibly increasing its side effects (Stockley, 2007). There are some helpful DDIs but most of them are dangerous (Aronson, 2007), for example, patients that take *clarithromycin* and *glibenclamide* concurrently may experiment *hypoglycaemia*.

There is a great amount of information about DDI described in papers that health experts have to consult in order to be updated. The development of tools for extracting this type of information from biomedical texts would produce a clear benefit for these professionals reducing the time necessary to review the literature. Semeval 2013 DDI Extraction challenge decided to being held with the aim of identifying the

state of the art algorithms for automatically extracting DDI from biomedical articles. This challenge has two tasks: recognition and classification of drug names and extraction of drug-drug interactions. For the second task, the input corpus contains annotations with the drug names.

A previous Workshop on Drug-Drug Interaction Extraction (Segura-Bedmar et al., 2011) was held in 2011 in Huelva, Spain. The main difference is that the new challenge includes the classification of the drug-drug interactions in four types depending on the information that is described in the sentence making the task much more complicated than before. Additionally the current task involves DDIs from two different corpora with different characteristics (Segura-Bedmar et al., 2013).

We participated in the task of extracting drug-drug interactions with two approaches that exploit a rich set of tree and sequence features. Our implemented methods utilize a SVM classifier with a linear kernel and a rich set of lexical, morphosyntactic and semantic features (e.g. trigger words) extracted from texts. In addition some tree features such as shortest path and subtree features are used.

2 Related work

Due to the importance of detecting biological and medical relations several methods have been applied for extracting biological relation information from text. In (Song et al., 2010) is presented a method for extracting protein-protein interaction (PPI) through combination of an active learning technique and a semi-supervised SVM.

Another motivating work can be found in (Chen et

al., 2011) in which a PPI Pair Extractor was developed that consists of a SVM for binary classification which exploits a linear kernel with a rich set of features based on linguistic analysis, contextual words, interaction words, interaction patterns and specific domain information.

Another PPI extraction method have been developed in (Li et al., 2010). They have applied an ensemble kernel composed of a feature-based kernel and a structure-based kernel. A more recent research on tree kernels has been carried out by (Guodong et al., 2010). They have introduced a context-sensitive convolution tree kernel, which specifies both context-free and context-sensitive sub-trees by taking into account the paths of their ancestor nodes as their contexts to capture structural information in the tree structure. A recent work (Simões et al., 2013) has introduced an approach for Relationship Extraction (RE) based on labeled graph kernels. The proposed kernel is a specification of a random walk kernel that exploits two properties: the words between the candidate entities and the combination of information from distinct sources. A comparative survey regarding different kernel based approaches and their performance can be found in (Frunza and Inkpen, 2008).

Using external knowledge and resources to the target sentence is another research direction in the relation extraction task that Chan and Roth have investigated in (Chan and Roth, 2010). They have reported some improvements by using external sources such as Wikipedia, comparing to basic supervised learning systems. Thomas and his colleagues in (Thomas et al., 2011) have developed a majority voting ensemble of contrasting machine learning methods using different linguistic feature spaces.

A more systematic and high quality investigation about feature selection in kernel based relation expression can be found in (Jiang and Zhai, 2011). They have explored a large space of features for relation extraction and assess the effectiveness of sequences, syntactic parse trees and dependency parse trees as feature subspaces and sentence representation. They conclude that, by means of a set of basic unit features from each subspace, a reasonably good performance can be achieved. But when the three subspaces are combined, the performance can

slightly improve, which shows sequence, syntactic and dependency relations have much overlap for the task of relation extraction.

Although most of the previous researches in biomedical domain has been carried out with respect to protein-protein interaction extraction, and more recently on drug-drug interaction extraction, other types of biomedical relations are being studied: e.g. gene-disease (Airola et al., 2008), disease-treatment (Jung et al., 2012) and drug-disease.

3 Dataset

The dataset for the DDIEExtraction 2013 task contains documents from two sources. It includes MedLine abstracts and documents from the DrugBank database describing drug-drug interactions (Segura-Bedmar et al., 2013). These documents are annotated with drug entities and with information about drug pair interactions: true or false.

In the training corpus the interaction type is also annotated. There are 4 types of interactions: *effect*, *mechanism*, *int*, *advice*.

The challenge corpus is divided into training and evaluation datasets (Table 1). The DrugBank training data consists of 572 documents with 5675 sentences. This subset contains 12929 entities and 26005 drug pair interactions. On the other hand, the MedLine training data consists of 142 abstracts with 1301 sentences, 1836 entities and 1787 pairs.

The distribution of positive and negative examples are similar in both subsets, 12.98% of positives instances on MedLine and 14.57% on DrugBank. With respect to the distribution of categories, the figures show that there is a small number of positive instances for categories *int* and *advice* on the MedLine subset. The *effect* type is the most frequent, outmatching itself on the MedLine subset.

The evaluation corpus contains 158 abstracts with 973 sentences and 5265 drug pair interactions from Drugbank, and 33 abstracts with 326 sentences and 451 drug pair interactions from Medline. It is worth to emphasize that the distribution of positive and negative examples is a bit greater (2.22%) in the DrugBank subset compared to the training data, but is almost doubled with respect to MedLine (12,98% to 21,06%). The categories *advice* and *int* have very few positive instances in the MedLine subset.

Training	pairs	negative DDIs	positive DDIs	effect	mechanism	advice	int
DrugBank	26005	22217	3788	1535	1257	818	178
MedLine	1787	1555	232	152	62	8	10
Test	pairs	negative DDIs	positive DDIs	effect	mechanism	advice	int
DrugBank	5265	4381	884	298	278	214	94
MedLine	451	356	95	62	24	7	2

Table 1: Basic statistics of the training and test datasets.

4 Method

Initially several experiments have been developed to explore the performance of shallow linguistic (SL) and parse tree based methods on a subset of the training corpus. Although the SL kernel achieved considerably good results we have found that the best option was the combination of different kernels using linguistic and tree features.

Our implemented kernel based approach consists of four different processes that have been applied sequentially: preprocessing, feature extraction, feature selection and classification (Figure 1). Our two submitted results were obtained by two different strategies. In the first outcome, all the DDIs and type of interactions were extracted in one step, as a 5-class categorization problem. The second run was carried out in two steps, initially the DDIs were detected and then the positively predicted DDIs were used to determine the type of the interaction. In the next subsection the four different processes are described.

4.1 Preprocessing

In this phase we have carried out two types of text preprocessing steps before training the classifier.

We have removed some stop words in special places in the sentences that clearly were a matter of concern and caused some inaccuracy, for example, removing question marks at the beginning of a sentence. We also carried out a normalization task for some tokens because of usage of different used encodings and processing methods, mainly html tags.

4.2 Feature extraction

Initially 49 feature classes were extracted for each instance that correspond to a drug pair interaction between Drug1 and Drug2:

- Word Features: Include Words of Drug1, words of Drug2, words between Drug1 and Drug2,

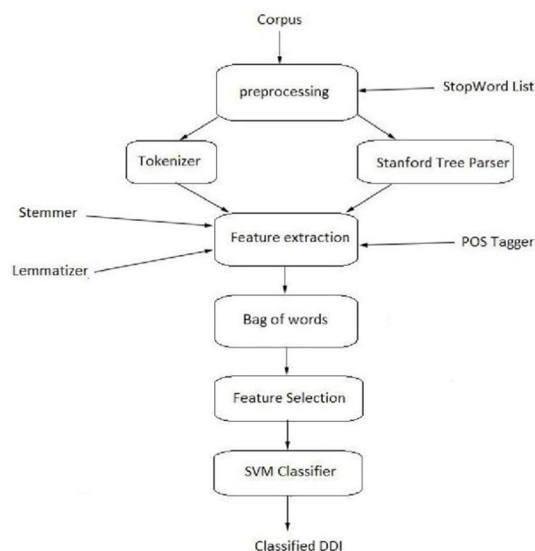


Figure 1: The different processes followed for our two submitted results.

three words before Drug1 and three words after Drug2. Lemmas and stems of all these words. We have used TreeTagger to obtain lemmas and Paice/Husk Stemmer (Paice, 1990) to obtain stems.

- Morphosyntactic Features: Include Part-of-speech (POS) tags of each drug words (Drug1 and Drug2), POS of the previous 3 and next 3 words. We have used TreeTagger.
- Constituency parse tree features: Include shortest path between Drug1 and Drug2 in the constituency parse tree, shortest path between first token in the sentence and Drug1, and shortest path between Drug2 and last token in the sentence in the parse tree, and all subtrees gener-

ated from the constituency parse tree. We have used Stanford parser ¹ for producing tree features.

- Conjunction features: We have produced some new conjunction features by combination of different word features and morphosyntactic features such as POSLEMMA and POSSTEM for all the words before Drug1, words between Drug1 and Drug2 and words after Drug2.
- verbs features: Include verbs between Drug1 and Drug2, first verb before Drug1 and first verb after Drug2. Their stem, lemma and their conjunction features are also included.
- negation features: Only if the sentence contains negation statements. The features extracted include the left side tokens of the negation scope, the right side tokens of the negation scope and the tokens inside the negation scope. We have used NegEx² as negation detection algorithm.

Finally we have deployed a bag of words approach (BoW) for each feature class in order to obtain the final representation for each instance. We have limited the size of the vocabulary in the BoW representation with a different number depending on the data subset. We carried out several experiments to fix these numbers and at the end we have used 1000 words/feature class for MedLine and 6000 words/feature class for DrugBank.

4.3 Feature selection

We have conducted some feature selection experiments to select the best features for improving the results and reducing running time. We have finally used Information Gain ranker to eliminate the less effective features. We have computed the information gain for each feature class as the linear combination of the information gain of each corresponding word. Empirically we have selected the best 42 feature classes.

On the other hand, we have done a preliminary study of the effect of the negation related features. We have found more than 3000 sentences containing negation, most of them corresponds to sentences

associated with negative examples of interactions. However, these features have been eliminated because we have not obtained a clear improvement when we combined them with the other features.

4.4 Classification

First we have performed several experiments with different supervised machine learning approaches such as SVM, Naivebayes, Randomtree, Random forest, Multilayer perceptron in addition to combination of methods. Finally we decided to use a SVM approach, the Weka Sequential Minimal Optimization (SMO) algorithm. We used the inner product of the BoW vectors as similarity function.

We have submitted two approaches:

- approach 1: SVM (Weka SMO) with 5 categories (effect, mechanism, int, advice and null).
- approach 2: We have extracted final results in two stages. In the first step we have used a SVM (Weka SMO) with 2 categories (positive and negative) and then we have used a second SVM classifier with 4 classes on positive extracted DDIs to train and extract the type of interaction in the test dataset.

The classifiers have been applied separately with each data subset, that is, a classifier per approach has been developed using the DrugBank training subset and has been evaluated using the DrugBank test subset, and the same process has been applied with the MedLine training and test subset.

5 Results

In this section we first show the evaluation results with our two approaches. Secondly an error analysis was carried out with a development set extracted from the training corpus.

5.1 Test data results

We have submitted two runs that corresponds with the approaches described in the previous section. Table 2 shows the results obtained with the first approach (one step) and Table 3 shows the results with the second approach (two steps).

It can be observed that the results on detection of DDI are better with the approach 2: 0.656 against 0.588 on F1. This result is a consequence that we

¹<http://nlp.stanford.edu/software/lex-parser.shtml>

²<http://code.google.com/p/negex/>

Run	P	R	F1
NILUCM1 (All)	0.632	0.464	0.535
NILUCM2 (All)	0.547	0.507	0.526
NILUCM1 (Drugbank)	0.651	0.498	0.565
NILUCM2 (Drugbank)	0.558	0.542	0.550
NILUCM1 (Medline)	0.333	0.074	0.121
NILUCM2 (Medline)	0.221	0.073	0.110

Table 4: Macroaverage test set results.

have more information to obtain the detection of the interaction if we use the information from all the different types than if we obtain it joining the results obtained per each category. With respect to detection and classification the results are also better with approach 2 for a similar reason: 0.548 against 0.517 on F1.

With respect to the categories, in the more populated ones the general tendency of better results from approach 2 continues, especially in *effect* type: 0.556 against 0.489. With respect to *advice* and *int*, the recall is better in approach 2 but the improvement in precision is greater in approach 1 giving a better result on F1 to approach 1, especially in *int* type: 0.427 against 0.393.

Table 4 shows the macroaverage results separated by subset data. The best results obtained for approach 1 are due to that this type of average gives equal weight to each category, favouring then the categories with less instances.

Other important insight that can be extracted from this table is that our results are much better for DrugBank dataset with both approaches. These results can be justified due to high similarity between sentences in Drugbank training and test corpus. In fact the Medline corpus commonly has more words unrelated to DDI subjects. In addition to this argument, the smaller number of training pairs in the Medline corpus can be other reason to obtain worst results.

5.2 Error analysis

We have extracted a stratified development corpus from the training corpus in order to perform an error analysis. We have used a 10% of the training corpus. It contains 2779 pairs, of which 397 are DDIs. Table 5 shows the results obtained with the two submitted approaches.

The results with our development corpus shows the same tendency, that is, approach 2 is better than approach 1 on detection of DDI and on microaverage classification. On the other hand, results are higher than those on test corpus because the information contained in the development corpus is more similar to the rest of training corpus than information on the test set.

We have performed an analysis of the errors produced for both approaches in the Detection and Classification of DDI subtask. The errors obtained are 112 false positives (Fp) and 116 false negatives (Fn) associated to approach 1, and 111 false positives (Fp) and 112 false negatives (Fn) to approach 2. Apart from the comments explained in the previous section about the small number of instances on the MedLine subset, we think the main problem is related with the management of long sentences with complex syntax. These sentences are more difficult for our approaches because the complexity of the sentence generates more errors in the tokenizing and parsing processes affecting the representation of the instances both in training and test phases. We show below some false positives and false negatives examples.

- The effects of **ERGOMAR** may be potentiated by **triacytyleandomycin** which inhibits the metabolism of ergotamine. DrugBank. False negative.
- Prior administration of **4-methylpyrazole** (90 mg kg(-1) body weight) was shown to prevent the conversion of **1,3-difluoro-2-propanol** (100 mg kg(-1) body weight) to (-)-erythro-fluorocitrate in vivo and to eliminate the fluoride and citrate elevations seen in 1,3-difluoro-2-propanol-intoxicated animals Med-Line. False negative.
- Drug Interactions with Antacids Administration of 120 mg of **fexofenadine hydrochloride** (2 x 60 mg capsule) within 15 minutes of an aluminum and magnesium containing antacid (Maalox) decreased **fexofenadine** AUC by 41% and cmax by 43%. DrugBank. False positive.
- **Dexamethasone** at 10(-10) M or retinyl acetate

approach 1	Tp	Fp	Fn	total	P	R	F1
Detection of DDI	557	359	422	979	0.608	0.569	0.588
Detection and classification of DDI	490	426	489	979	0.535	0.501	0.517
Score for type mechanism	147	122	155	302	0.546	0.487	0.515
Score for type effect	200	258	160	360	0.437	0.556	0.489
Score for type advice	115	39	106	221	0.747	0.520	0.613
Score for type int	28	7	68	96	0.800	0.292	0.427

Table 2: Test corpus results (approach1).

approach 2	Tp	Fp	Fn	total	P	R	F1
Detection of DDI	631	315	348	979	0.667	0.645	0.656
Detection and classification of DDI	527	419	452	979	0.557	0.538	0.548
Score for type mechanism	146	102	156	302	0.589	0.483	0.531
Score for type effect	210	186	150	360	0.530	0.583	0.556
Score for type advice	139	96	82	221	0.591	0.629	0.610
Score for type int	32	35	64	96	0.478	0.333	0.393

Table 3: Test corpus results (approach2).

approach 1	Tp	Fp	Fn	total	P	R	F1
Detection of DDI:	292	101	105	397	0.743	0.736	0.739
Detection and Classification of DDI:	281	112	116	397	0.715	0.708	0.711
approach 2	Tp	Fp	Fn	total	P	R	F1
Detection of DDI:	296	102	101	397	0.744	0.746	0.745
Detection and Classification of DDI:	285	111	112	397	0.720	0.718	0.719

Table 5: Error analysis with a development corpus.

at about $3 \times 10^{(-9)}$ M inhibits **proliferation** stimulated by EGF. MedLine. False positive.

6 Conclusions

In this paper we have shown our approaches for the Semeval 2013 DDI Extraction challenge. We have explored different combinations of tree and sequence features using the Sequential Minimal Optimization algorithm.

The first approach uses a SVM with 5 categories, and the second one extracts the final results in two steps: detection with all the categories, and classification on the positive instances. The results are better for approach 2 mainly due to the improvement on the detection subtask because the information from all the categories is used.

We think some of our errors come from using a general tool (Stanford parser) to obtain the parse tree

of the sentences. In the future we are going to explore other biomedical parsers and tokenizers.

With respect to the data used, we think the MedLine dataset needs to be greater in order to obtain more significant analysis and results. Our approaches are especially affected by this issue because the small number of positive instances on *advice* and *int* categories implies that the algorithm can not learn to classify new instances accurately. On the other hand, although n-fold cross validation is considered as the best model validation technique, it was time consuming for DDI and need powerful processors.

Another interesting future work is related with the application of simplification techniques in order to solve the problems in the processing of complex long sentences (Buyko et al., 2011).

References

- A. Airola, S. Pyysalo, J. Bjorne, T. Pahikkala, F. Ginter, T. Salakoski. 2008. Allpaths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning *BMC Bioinformatics*, 9(Suppl 11):S2.
- JK. Aronson. 2007. Communicating information about drug interactions. *British Journal of Clinical Pharmacology*, 63(6):637–639.
- E. Buyko, E. Faessler, J. Wermter, U. Hahn 2011. Syntactic Simplification and Semantic Enrichment - Trimming Dependency Graphs for Event Extraction. *Computational Intelligence*, 27(4):610–644.
- Y. Chen, F. Liu, B. Manderick. 2011. Extract Protein-Protein Interactions from the Literature Using Support Vector Machines with Feature Selection. *Biomedical Engineering, Trends, Researchs and Technologies*, 2011.
- YS. Chan and D. Roth. 2010. Exploiting Background Knowledge for Relation Extraction *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics*, pp:152–160.
- O. Frunza and D. Inkpen. 2010. Extraction of disease-treatment semantic relations from biomedical sentences *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pp:91–98.
- Z. Guodong, Q. Longhua, F. Jianxi. 2010. Tree kernel-based semantic relation extraction with rich syntactic and semantic information *International Journal on Information Sciences*, 180(8):1313–1325.
- J. Jiang and C. Zhai. 2011. A systematic exploration of the feature space for relation extraction *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACLHLT07)*, pp:113–120.
- H. Jung, S. Choi, S. Lee, S. Song. 2012. Survey on Kernel-Based Relation Extraction.
- L. Li, J. Ping, D. Huang. 2010. Protein-Protein Interaction Extraction from Biomedical Literatures Based on a Combined Kernel *Journal of Information & Computational Science*, 7(5):1065–1073.
- Chris D. Paice 1990. Another stemmer. *ACM SIGIR Forum*, 24(3):56–61.
- I. Segura-Bedmar, P. Martínez, D. Sánchez-Cisneros. 2011. *Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)* CEUR Workshop Proceedings, Vol. 761.
- I. Segura-Bedmar, P. Martnez, M. Herrero-Zazo. 2013 SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts. *In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.
- G. Simões, D. Matos, H. Galhardas. 2013. A Labeled Graph Kernel for Relationship Extraction. *CoRR*, abs/1302.4874.
- M. Song, H. Yu, W. Han. 2010. Combining active learning and semi-supervised learning techniques to extract protein interaction sentences. *International Workshop on Data Mining in Bioinformatics*.
- I H Stockley. 2007. *Stockley's Drug Interaction*. Pharmaceutical Press.
- P. Thomas, M. Neves, I. Solt, D. Tikk, U. Leser. 2011. Relation extraction for drug- drug interactions using ensemble learning *Proceedings of the First Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*, pp:11–17.

An experiment is a question which science poses to Nature, and a measurement is the recording of Nature's answer.inspires.

MAX PLANCK



Enhancing Extraction of Drug-Drug Interaction from Literature using Neutral Candidates, Negation, and Clause Dependency

RESEARCH ARTICLE

Enhancing Extraction of Drug-Drug Interaction from Literature Using Neutral Candidates, Negation, and Clause Dependency

Behrouz Bokharaeian^{1*}, Alberto Diaz¹, Hamidreza Chitsaz²

1 NIL Group, Complutense University of Madrid, Ciudad Universitaria, Calle Profesor José García Santesmases, 28040 Madrid, Spain, **2** Department of Computer Science, Colorado State University, Fort Collins, Colorado, United States of America

* behroubo@ucm.es



OPEN ACCESS

Citation: Bokharaeian B, Diaz A, Chitsaz H (2016) Enhancing Extraction of Drug-Drug Interaction from Literature Using Neutral Candidates, Negation, and Clause Dependency. PLoS ONE 11 (10): e0163480. doi:10.1371/journal.pone.0163480

Editor: Francisco M Couto, University of Lisbon, PORTUGAL

Received: April 23, 2016

Accepted: September 9, 2016

Published: October 3, 2016

Copyright: © 2016 Bokharaeian et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: We have produced and used two data(sets) in our paper, available on Figshare and DOI(digital object identifier): 1- NegDDI_MEDLINE(.zip) URL <https://figshare.com/s/b657c8ccfa152ed8a426> DOI [10.6084/m9.figshare.3827013](https://doi.org/10.6084/m9.figshare.3827013) <<https://dx.doi.org/10.6084/m9.figshare.3827013>> 2- NegDDI_DrugBank(.zip) URL <https://figshare.com/s/8cf5353cd42aee075fb> DOI [10.6084/m9.figshare.3827010](https://doi.org/10.6084/m9.figshare.3827010) <<https://dx.doi.org/10.6084/m9.figshare.3827010>>.

Funding: The authors received no specific funding for this work.

Abstract

Motivation

Supervised biomedical relation extraction plays an important role in biomedical natural language processing, endeavoring to obtain the relations between biomedical entities. Drug-drug interactions, which are investigated in the present paper, are notably among the critical biomedical relations. Thus far many methods have been developed with the aim of extracting DDI relations. However, unfortunately there has been a scarcity of comprehensive studies on the effects of negation, complex sentences, clause dependency, and neutral candidates in the course of DDI extraction from biomedical articles.

Results

Our study proposes clause dependency features and a number of features for identifying neutral candidates as well as negation cues and scopes. Furthermore, our experiments indicate that the proposed features significantly improve the performance of the relation extraction task combined with other kernel methods. We characterize the contribution of each category of features and finally conclude that neutral candidate features have the most prominent role among all of the three categories.

Introduction

Extracting biomedical relations from texts is a relatively new, but rapidly growing research field in natural language processing. Owing to the increasing number of biomedical research publications and the key role of databases of biomedical relations in biological and medical research, extracting biomedical relations from scientific articles and text resources is of utmost importance. Drug-drug interaction (DDI) is, in particular, a widespread concern in medicine,

Competing Interests: The authors have declared that no competing interests exist.

and thus, extracting this kind of interaction automatically from texts is of high demand in BioNLP. Drug-drug interaction usually occurs when one drug alters the activity level of another drug. According to the reports prepared by the Food and Drug Administration (the FDA) and other acknowledged studies [1], over 2 million life-threatening DDIs occur in the United States every year. Many academic researchers and pharmaceutical companies have developed relational and structural databases, where DDIs are recorded. Nevertheless, most up-to-date and valuable information is still found only in unstructured research text documents, including scientific publications and technical reports.

In this paper, we first introduce the basics of three complementary, linguistically driven feature sets of (i) negation, (ii) clause dependency, and (iii) neutral candidates. The ultimate aim of this research is to enhance the performance of DDI extraction task by considering and employing the above-mentioned three operations and feature sets.

First, it is essential to detect negative assertions in most biomedical text-mining tasks, where the overall purpose is to derive factual knowledge from textual data. According to Loos et al. [2], negation is a morphosyntactic operation in which a lexical item denies or inverts the meaning of another lexical item or construction. Likewise, a negator is a lexical item that expresses negation. Negation is commonly utilized in biomedical articles and is an important origin of low precision in automated information retrieval systems [3]. Generally, two negation detection methods have been developed and employed for annotating the applied corpora: a linguistic-based approach and an event-oriented approach. Two of the known negation annotated corpora are the linguistically focused, scope-based *BioScope* and the event-oriented *Genia* [4].

Second, identifying the role of clause dependency in complex sentences in DDI detection is another linguistically driven subject which is investigated in this research. According to Harris and Rowan [5], a dependent clause is a group of words with a subject and a verb that do not express a complete thought, cannot stand alone, and usually extend the main clause. An independent clause, or main clause, is one that can stand alone as a sentence and express a complete thought. Consequently, a complex sentence has one independent clause and at least one dependent clause. Moreover, a clause connector is a word that joins clauses in order to form complex sentences. Coordinators, conjunctive adverbs, and subordinators are three types of connectors.

Miwa et al. [6] have considered clauses in relation extraction task. They have reported some improvements regarding different types of simplification and clause selection rules which they have applied. By contrast, in this research we extract new features based on the text or subtree features in a kernel-based relation extraction method. Our features detect the existence token or subtree in a dependent or independent clause as well as the type of the clause itself via checking several clause connectors.

Finally, we study the role of neutral DDI candidates in the relation extraction. Most of the current relation extraction problems and the produced corpora are based on binary relations; they decide a binary relation between two entities. Similarly, in the DrugDDI corpus [7], the implemented systems must predict whether or not an interaction between the two drugs has occurred. Although detecting DDI interactions is the main target of the DrugDDI corpus, there is a difference between a negative interaction candidate having been stated by the authors (distinguished candidate) and that which has not (neutral candidate). Both of these candidates are considered negative in DrugDDI corpus. In other words, the neutral interaction candidate is a co-mention of two drugs with no remarks by the author in the sentence or the discussed clause, while the distinguished interaction candidate is exactly the opposite (with remarks by the author). In point of fact, neutral candidates are a particular subclass of non-positive candidates whose lack of interaction cannot be exactly determined by the confident level above zero. For instance, consider the following sentence:

- Studies in healthy volunteers have shown that acarbose has no effect on either the pharmacokinetics or pharmacodynamics of digoxin, nifedipine, propranolol, or ranitidine.

There is no remark by the author about the interaction between *propranolol* and *ranitidine*. Therefore, we define this candidate of drug-drug interaction as a neutral candidate.

One among the few studies on detection of neutral candidates has been conducted by [8], introducing two iteration-based systems of DIPRE and Snowball that take into account the confidence level of the relation. In both systems, when the confidence level is zero, there is a neutral candidate. Moreover, Frunza and Inkpen have carried out another similar research which considers neutral candidates [9]. They categorize and extract the semantic relationships between disease and treatments from biomedical sentences. However, no significant improvement has been reported through using neutral class in the work.

In the present study, we characterize the role and the potential importance of the three above-mentioned categories of features in DDI extraction. We employ the combinations of the extracted features along with the existing well-established kernel methods. For instance, the status of a neutral DDI candidate is not inverted when negation is used, whereas a non-neutral candidate is inverted. In addition, when a negator is added, the overall status of a DDI candidate may or may not be reversed, depending on the type of the clause connector that contains DDI candidate and negator. This issue will be expounded in the methods section.

The rest of the paper is structured as follows. The following section provides the background in some of the kernel-based relation extraction methods, beneficial NLP subtasks, and some of the related data sources. In section 3, we present our approach and the feature extraction process, and section 4 is devoted to presenting the results obtained. The final section concludes the paper and gives some suggestions for future research.

Background

The majority of previous works on biomedical relation extraction, including the DDI detection, have been carried out on the basis of supervised binary relations extraction [8]. In this paper, we summarize kernel-based relation extraction methods as well as some NLP preprocessing enhancements and the related corpora.

2.1 Kernel-based methods

Sequence kernels [10], Tree kernels such as parse tree based [11], and Graph kernels such as graph parsing [12] are among the most important kernel-based methods [13]. Two more recent approaches have been proposed by [14] and [15], being ranked first and second in DrugDDI challenge (2013), respectively. Chowdhury and Lavelli [14] proposed a hybrid kernel through linear combination of a feature-based kernel, a Shallow Linguistic (SL) kernel, and a Path-Enclosed Tree (PET) kernel. Through defining a multiplicative constant, they assigned more (or less) weight to the information obtained by tree structures. Another recent work has been accomplished by [16] who employed a feature-based linear kernel that contains five categories of features, including word pair and dependency graph features. In addition to the previous methods, a number of research have improved the performance of the task through ensemble approaches. For example, Thomas and his colleagues [15] proposed a two-step approach in which the relation candidates are initially extracted, using the ensembles of up to five different classifiers and then are relabeled to one of the four used categories in the task. The other work which has been suggested by He and her colleagues [17] applies a stacked generalization approach to learn the weights which have been exploited to combine graph and tree kernels.

2.2 NLP enhancements

Several related NLP enhancements have improved the performance of the relation extraction algorithms. They are often employed as a preprocessing step which is a pivotal stage in enhancing the overall performance and results. In particular, we summarize the studies on negation, sentence, and clause simplification.

Faisal et al. [18] took negation into account in the relation extraction task. They developed a list of features, such as the nearest verb to the candidate entities in the parse tree and few negation cues, which are fed into an SVM classifier. They reported some improvements, but did not specify how much the negation identification step enhanced the performance.

Another NLP enhancement in the relation extraction is sentence and clause simplification to overcome the complexity of the sentences. Text simplification modifies, enhances, classifies, or otherwise processes an existing text in such a way that the grammar and the structure of the prose are simplified to a great extent, while the original meaning and information remain the same [19]. ISIMP is a system that simplifies the text so that its mining tools, including the relation extraction tasks, can be improved [20]. In the same direction, Segura and her colleagues proposed techniques to simplify complex sentences by splitting the clauses [21]. They applied some rules and patterns to split the clauses and then utilized some simplification rules to generate new simple sentences. However, according to their conclusion, difficulty of resolving nested clauses is the major source of errors. There are other NLP subtasks enhancements that can be employed in the relation extraction task, although they were not applied in our work. To name a few, Velldal et al. [22] proposed speculation detection, and Lappin [23] utilized anaphora resolution.

2.3 Related corpora

DrugDDI corpora. Drug-Drug Interaction corpus was primarily developed by Segura and Mart [7], with 579 XML files describing DDIs which were collected randomly from the DrugBank database [24]. The first DDI Extraction competition was held in 2011 with the aim of encouraging researchers to explore new methods for extracting drug-drug interactions [7]. A second competition was held in 2013 as part of SemEval-2013 (International Workshop on Semantic Evaluation). Furthermore, a new corpus was developed which included the corpus used in 2011 (DDI-DrugBank, 2011) as well as some MEDLINE abstracts. The teams participating in this venue had developed solutions based on supervised and sentence-level relation extraction methods, and the best F-measure achieved was 75% [25].

Corpora annotated with negation. As mentioned earlier, thus far two negation detection methods have been developed and employed for annotating the corpora utilized: a linguistic-based approach and an event-oriented approach. Linguistically-focused BioScope and the event-oriented Genia [4] are two of the known negation annotated corpora.

In BioScope, the scopes aim to recognize the negation position of the key event in the sentence and with each argument of these key events was located under the negation scope as well [26]. In contrast, Genia deals with the modality of events within the events independently. In the Genia event, biological concepts (relations and events) are annotated for negation, but no linguistic cues are annotated for them. In point of fact, the main objective of the BioScope corpus is to investigate this language phenomenon in a general, task-independent, and linguistically-oriented manner. Additionally, in the BioScope, in-sentence negation scope and cues can be recognized automatically [4].

NegDDI-DrugBank corpus. Konstantinova et al. developed two corpora [27] and Morante and Blanco [28] adapted Bioscope's guidelines. These adaptations in addition to the previously mentioned advantages of the bioscope annotations prove them to be a valuable resource.

```
<sentence id="DDI-DrugBank.d297.s4" text="Concurrent therapy with ORENCIA and TNF antagonists is not recommended.">
  <entity charOffset="24-30" id="DDI-DrugBank.d297.s4.e0" text="ORENCIA" type="brand"/>
  <entity charOffset="36-50" id="DDI-DrugBank.d297.s4.e1" text="TNF antagonists" type="group"/>
  <pair ddi="true" e1="DDI-DrugBank.d297.s4.e0" e2="DDI-DrugBank.d297.s4.e1" id="DDI-DrugBank.d297.s4.p0" type="advise"/>
  <negationtags><xscope> Concurrent therapy with ORENCIA and TNF antagonists is <cue>not</cue>
    recommended</xscope>.</negationtags>
</sentence>
```

Fig 1. The extended unified XML format of a sentence with negation cue in NegDDI-DrugBank corpus.

doi:10.1371/journal.pone.0163480.g001

Consequently, we produced NegDDI-DrugBank corpus based on the Bioscope's guidelines. For this purpose, all sentences of DrugDDI (2011) and DrugBank part of the DrugDDI (2013) were utilized and automatically annotated. Bokharaeian et al. [29] explained the annotation process and presented a detailed analysis of the number of distinct negation cues in the NegDDI-DrugBank corpus. The extended corpus is available for public use [30]. A sample of the extended negation annotation can be seen in Fig 1. The negation scope and the cue xml tags are highlighted in the extended part which is transparent in this figure.

Methods

In this section, the feature extraction phase as well as the proposed method for the DDI prediction will be presented. Our features, presented in Table 1, are categorized into three major categories based on the linguistic definition of negation, the position of the drugs discussed in the sentence, and the linguistic-based confident level of an interaction: (i) negation scope and cue-related features, (ii) clause dependency features, and (iii) neutral candidate's features. In all of the presented tables, "NEG" has been used as the abbreviation for the negation scope and cue feature set, and "CLA" and "NEUT" stand for the clause dependency feature set and the neutral candidate feature set, respectively. Moreover, it is worth mentioning that all of the sample sentences in this paper have been obtained from the DrugDDI corpus [25].

Additionally, as previously mentioned, DDI Extraction (2013) datasets also include 233 MEDLINE abstracts in addition to the obtained DrugBank texts. This extension was carried out due to dealing with different types of texts and language styles [25]. While, DDI-DrugBank texts focus on the description of drugs and their interactions, the main topic of DDI-MEDLINE texts does not necessarily focus on DDIs. Consequently, in addition to the annotation of the DrugBank part of the corpus, we annotated the MEDLINE part with negation scope and cue. The annotation process was carried out in a similar way to the above-mentioned DrugBank part. The prepared corpus is available at this address (<https://figshare.com/s/b657c8ccfa152ed8a426>)

3.1 Negation scope and cue features

In negative sentences, the relative position of the entities compared to the negation scope and cue is an important factor that can be extracted directly from the extended corpus. For

Table 1. The list of the extracted features used in the system.

Feature category	Feature name	Type	Definition
Negation Scope and Cue	BothInsideNegSc	Boolean	is set as true when both drugs are inside the negation scope
	BothRightNegSc	Boolean	is set as true when both drugs are on the right side of the negation scope
	BothLeftNegSc	Boolean	is set as true when both drugs are on the left side of the negation scope
	OneLeftOneInsideNegSc	Boolean	is set as true when one drug is on the left side of the negation scope, and the other on the inside
	OneRightOneInsideNegSc	Boolean	is set as true when one drug is on the right side of the negation scope, and the other on the inside
	OneLeftOneRightSc	Boolean	is set as true when one drug is on the right side of the negation scope, and the other on the left
	NegationCue	String	Negation cue
Clause Dependency Detection	AlthoughIS	Boolean	set as true when the sentence has <i>although</i> token
	WhileIS	Boolean	set as true when the sentence has <i>while</i> token
	WhenIS	Boolean	set as true when the sentence has <i>when</i> token
	BeforeIS	Boolean	set as true when the sentence has <i>before</i> token
	NowthatIS	Boolean	set as true when the sentence has <i>now that</i> token
	AssoonasIS	Boolean	set as true when the sentence has <i>as soon as</i> token
	AslongasIS	Boolean	set as true when the sentence has <i>as long as</i> token
	AnywhereIS	Boolean	set as true when the sentence has <i>anywhere</i> token
	UntilIS	Boolean	set as true when the sentence has <i>until</i> token
	OnceIS	Boolean	set as true when the sentence has <i>once</i> token
	TillIS	Boolean	set as true when the sentence has <i>till</i> token
	BecauseIS	Boolean	set as true when the sentence has <i>because</i> token
	ThoughIS	Boolean	set as true when the sentence has <i>though</i> token
	EventthoughIS	Boolean	set as true when the sentence has <i>even though</i> token
	SinceIS	Boolean	set as true when the sentence has <i>since</i> token
	ButIS	Boolean	set as true when the sentence has <i>but</i> token
	UnlessIS	Boolean	set as true when the sentence has <i>unless</i> token
	afterIS	Boolean	set as true when the sentence has <i>after</i> token
	whereasIS	Boolean	set as true when the sentence has <i>where</i> token
	asthoughIS	Boolean	set as true when the sentence has <i>as though</i> token
	sothatIS	Boolean	set as true when the sentence has <i>so that</i> token
	inorderthatIS	Boolean	set as true when the sentence has <i>in order to</i> token
	everywhereIS	Boolean	set as true when the sentence has <i>everywhere</i> token
	evenifIS	Boolean	set as true when the sentence has <i>even if</i> token
	RatherthanIS	Boolean	set as true when the sentence has <i>rather than</i> token
	AslongasIS	Boolean	set as true when the sentence has <i>as long as</i> token
	OnlyifIS	Boolean	set as true when the sentence has <i>only if</i> token
	JustasIS	Boolean	set as true when the sentence has <i>just as</i> token
	F-StructuresDependencies	String	Corresponding to every feature F of the original method which contains only tokens or subtrees, if the token or subtree X located in an independent clause, a string X-IDC added to this new feature, otherwise if the token or subtree X located in a dependent clause, a string X-DC added to this new text feature
Neutral Candidate Detection	NeutralCandRule1	Boolean	(.) [*] d1(/s/)d2(.)
	NeutralCandRule2	Boolean	d2 d1.contains(OtherNs(d2)) (d2.contains(OtherNs(d1)))
	NeutralCandRule3	Boolean	(.) [*] d1((s) (N, e.g. i.e. s DrgNaOth ,)) [*] d2(.) [*]
	NeutralCandRule4	Boolean	(.) [*] d1(s) [*] , (s DrgNaOth , , and , other oral) [*] d2(.) [*]
	NeutralCandRule5	Boolean	(.) [*] (: such as e.g. i.e. s DrgNaOth , and or and/or) [*] d1(s DrgNaOth , and) [*] d2(.) [*]
	NeutralCandRule6	Boolean	(.) [*] (been studied)(.) [*]
	NeutralCandRule7	Boolean	(.) [*] been investigated (.) [*] & (.) [*] (although)(.) [*]
	NeutralCandRule8	Boolean	(.) [*] (been established)(.) [*]
	NeutralCandRule9	Boolean	(.) [*] (studies)(.) [*] (performed)(.) [*] & (.) [*] (studies)(.) [*] (conducted)(.) [*]
	NeutralCandRule10	Boolean	[(:) [*]][no experience][(:) [*]]

doi:10.1371/journal.pone.0163480.t001

instance, consider the negated sentence in Fig 2 [31]. As can be seen in this figure, the scope of negation is highlighted in green.

Population pharmacokinetic analyses revealed that MTX, NSAIDs, corticosteroids, and TNF blocking agents did not influence abatacept clearance.

Fig 2. A sample of a negated sentence with some DDI candidates.

doi:10.1371/journal.pone.0163480.g002

In the sentence, *MTX* and *NSAIDs*, which have been highlighted in the image, are two drug names that are located outside the negation scope, and consequently, their interaction status is not inverted by negation. However, *abatacept* and *MTX* interaction status is inverted by negation due to the position of *abatacept* located in the negation scope. Regarding the position of drug names inside or outside the negation scope, there are 6 different possibilities used as the six features:

1. BothInsideNegSc: A Boolean feature which is set true when both drugs are inside the negation scope and is set false in all other situations.
2. BothRightNegSc: A Boolean feature which is set true when both drugs are on the left side of the negation scope and is set false in all other situations.
3. BothLeftSNegSc: A Boolean feature which is set true when both drugs are on the right side of the negation scope and is set false in all other situations.
4. OneLeftOneInsideNegSc: A Boolean feature which is set true when one drug is on the left side of the negation scope and the other drug is inside it. The Boolean feature is set false in all other situations.
5. OneRightOneInsideNegSc: A Boolean feature which is set true when one drug is on the right side of the negation scope and the other drug is inside it. The Boolean feature is set false in all other situations.
6. OneLeftOneRightSc: A Boolean feature which is set true when one drug is on the right side of the negation scope and the other drug is on its left side. The Boolean feature is set false in all other situations.

In addition to these six features, the negation cue is utilized as a text feature.

3.2 Clause dependency features

Previous studies generally indicate that complex and compound sentences, which are very common in the biomedical literature, produce more errors than simple sentences with one clause [21]. Thus, distinguishing between independent and dependent clauses is a critical matter. The analyses demonstrate that more than 27% of the sentences in the test part of NegDDI--DrugBank and 19% of the sentences in the training part of NegDDI--DrugBank have at least one dependent clause. Since a large number of sentences have more than one clause in complex structures, taking clause dependency features into account is important. However, there are different types of dependent clauses that can alter the overall meaning of a sentence in different ways. For instance, *concessive* clause is a clause which begins with “although” or “even though” and expresses an idea that suggests the opposite of the main part of the sentence, like in the sentence shown in Fig 3. The sentence also has one negation cue and scope which has been highlighted in green, and the two drug candidates are highlighted in blue. The clause connector is highlighted in red.

The main clause (“*Co-administration of TIKOSYN with verapamil resulted in increases in dofetilide peak plasma levels by 42%.*”) conveys a meaning opposite to that of the dependent

```

- <sentence id="DDI-DrugBank.d558.s7" text="Co-administration of TIKOSYN with verapamil resulted in increases in dofetilide peak plasma levels of 42%, although overall exposure to dofetilide was not significantly increased.">
  <entity id="DDI-DrugBank.d558.s7.e0" text="TIKOSYN" type="brand" charOffset="21-27"/>
  <entity id="DDI-DrugBank.d558.s7.e1" text="verapamil" type="drug" charOffset="34-42"/>
  <entity id="DDI-DrugBank.d558.s7.e2" text="dofetilide" type="drug" charOffset="69-78"/>
  <entity id="DDI-DrugBank.d558.s7.e3" text="dofetilide" type="drug" charOffset="136-145"/>
  <pair id="DDI-DrugBank.d558.s7.p0" type="mechanism" e2="DDI-DrugBank.d558.s7.e1" e1="DDI-DrugBank.d558.s7.e0" ddi="true"/>
  <pair id="DDI-DrugBank.d558.s7.p1" e2="DDI-DrugBank.d558.s7.e2" e1="DDI-DrugBank.d558.s7.e0" ddi="false"/>
  <pair id="DDI-DrugBank.d558.s7.p2" e2="DDI-DrugBank.d558.s7.e3" e1="DDI-DrugBank.d558.s7.e0" ddi="false"/>
  <pair id="DDI-DrugBank.d558.s7.p3" e2="DDI-DrugBank.d558.s7.e2" e1="DDI-DrugBank.d558.s7.e1" ddi="false"/>
  <pair id="DDI-DrugBank.d558.s7.p4" e2="DDI-DrugBank.d558.s7.e3" e1="DDI-DrugBank.d558.s7.e1" ddi="false"/>
  <pair id="DDI-DrugBank.d558.s7.p5" e2="DDI-DrugBank.d558.s7.e3" e1="DDI-DrugBank.d558.s7.e2" ddi="false"/>
  <negationtags>Co-administration of TIKOSYN with verapamil resulted in increases in dofetilide peak plasma levels of 42%, although overall exposure to dofetilide was <xscope><cue>not</cue> significantly increased</xscope>.</negationtags>
</sentence>

```

Fig 3. A sample of a negated sentence with a concessive clause.

doi:10.1371/journal.pone.0163480.g003

clause (“Overall exposure to dofetilide did not significantly increase.”). As another example, a graphical view of a parse tree for a complex sentence with a highlighted dependent clause and two highlighted negation cues is presented in Fig 4 [31]. Although it appears that the first clause conveys the same idea as the main clause expresses, the dependent clauses carry less important information than do the main clauses from a linguistic point of view. This point has been neglected in most previous methods, particularly in the sequence kernels.

The next most frequent type of clause in the corpus is adverbial clauses of time that indicate the time of a DDI prevalent in the pharmacological literature. The analysis carried out in the corpus shows that the most frequent adverbial clause connectors are “when”, “while”, and “before”. They collectively constitute approximately half of the total clause connectors.

Considering the different types of dependent clauses, two categories of features were extracted. The first group consists of 28 Boolean features corresponding to 28 clause connectors. The complete list of those connectors as well as their corresponding features is presented in Table 1. The second group of features is based on the substructures (token or subtree) utilized in the applied method, which locates whether the substructure is inside the main clause or not. Three new text features, similar to the features used in the Global context kernel [32], were extracted with IDC prefix for independent clause tokens and DC for dependent clause tokens. Similarly, to improve the subtree kernel, we defined new subtrees. In short, the subtree inside a dependent or independent clause comes with DC or IDC prefix beside the root name, respectively.

3.3 Neutral candidate features

As it was previously explained, the distinction between distinguished and neutral interaction candidates is critical. A neutral candidate is one with no remark by the author in the sentence, while a distinguished candidate is exactly the opposite (with remarks by the author). In point of fact, neutral candidates are a particular subclass of non-positive candidates that are detectable by meticulously defined features. However, a distinguished candidate can belong to the positive or negative class of DDI's. In the sentence in Fig 5, the two mentioned interaction candidates are shown. The status of the interaction candidate between each of the discussed drugs at the end of the sentence, i.e. *Propranolol*, *Ranitidine*, etc. is neutral because there are no remarks by the author about their interaction with each other. However, the status of the relation between *Acarbose* and *Ranitidine*, *Propranolol*, and the other mentioned drugs is

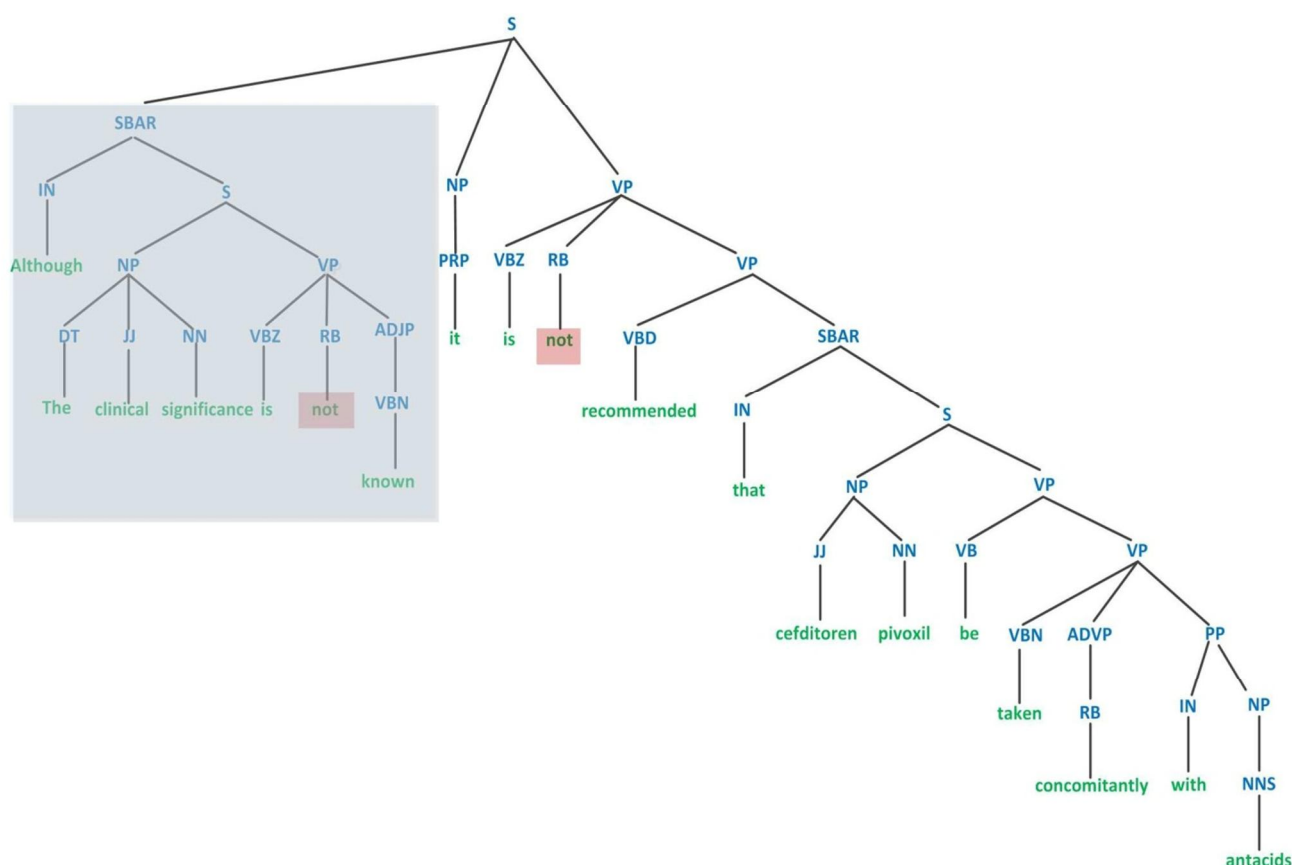


Fig 4. A constituency parse tree of a sentence with a concessive dependent clause highlighted in blue and two negation cues.

doi:10.1371/journal.pone.0163480.g004

distinguished since the author explicitly explains the lack of interaction (“... *Acarbose* has no effect on either the...”).

From the negation action perspective, a negation cue inverts the distinguished candidate, but does not invert the status of a neutral interaction candidate. For instance, in the sentence in Fig 5, the negation has inverted the status of the distinguished candidate *Acarbose* and *Ranitidine* from positive into negative. However, it has not changed the status of the neutral candidates *Propranolol* and *Ranitidine*, and thus, the interaction has remained negative.

In more precise terms, a DDI candidate is called neutral if it has the following two properties:

1. The interaction or lack of interaction between two drugs cannot be extracted from the sentence (or container clause) with confidence level more than zero.
2. The status of the interaction or lack of interaction between two drugs does not change from positive to negative or vice versa if the sentence (or container clause) is negated and drug names are located in the scope of the negation.

It is worth mentioning that being a neutral candidate can be defined in different linguistic scopes such as a clause, sentence, or a paragraph. In the present paper, a neutral candidate is defined in the scope of the container clause and sentence. Accordingly, 10 Boolean features have been defined concerning linguistically different patterns to detect neutral candidates in

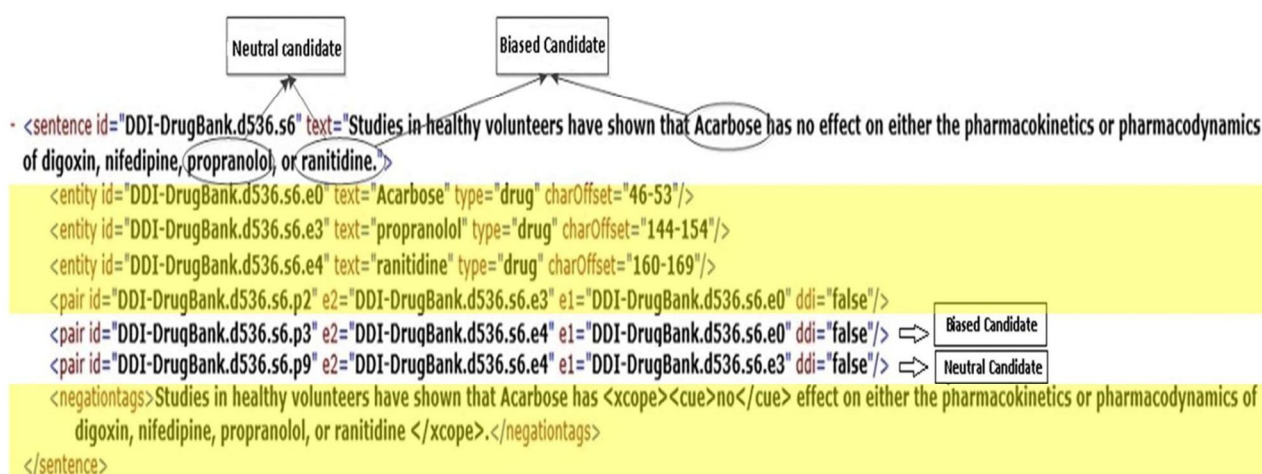


Fig 5. A sample sentence with negation from NegDDI-DrugBank with neutral and distinguished false DDIs.

doi:10.1371/journal.pone.0163480.g005

the clause and sentences (Table 1). A rule-based system was implemented, using regular expression language beside to some defined functions to extract the features below. In the table, java regular expression patterns have been utilized to mention the rules [33]. In addition to the used patterns, some predefined variables and functions were used in the written rules such as “DrgNaOth” constant has been used as non-DDI candidate drug names. Moreover, “OtherNs (Drug)” is a function, which returns other generic or brand name of the Drug. Below shows the corresponding feature names alongside to the implemented rules:

- NeutralCandRule1-2: Two Boolean features are set true when the second drug name is a sample, a commercial brand or other common name of the first drug or both drugs belong to the same pharmacological class. For instance, in the sentence given in Fig 6, *Purinethol* is the brand name for *mercaptopurine*, and similarly *Imuran* for *azathioprine*:

The first Boolean (NeutralCandRule1) feature identifies textual patterns, where a “/” and a “(” separate the two drug names, and the second Boolean (NeutralCandRule2) feature detects textual patterns in which one of drug names contain another drug name or its synonyms. In both cases, the interaction status between the two recognized drug names is a valueless concept, hence a neutral candidate.

- NeutralCandRule3-5: Three Boolean features are set true when an interaction between the two desired drugs with a third drug (or drugs) has been investigated; however, the interaction between the two drugs discussed has not been inspected. For instance, in the sentence presented in Fig 7, the interaction between *doxorubicin* and *bleomycin* (highlighted in red) has not been studied.

In patients receiving **mercaptopurine (Purinethol)** or **azathioprine (Imuran)**, the concomitant administration of 300–600 mg of allopurinol per day will require a reduction in dose to approximately one-third to one-fourth of the usual dose of mercaptopurine or azathioprine.

Fig 6. A sample of a sentence having two neutral DDI candidates.

doi:10.1371/journal.pone.0163480.g006

However, in a well-controlled study of patients with lymphoma on combination therapy, **Allopurinol** did not increase the marrow toxicity of patients treated with cyclophosphamide, **doxorubicin**, **bleomycin**, procarbazine and/or mechlorethamine.

Fig 7. A sample of a sentence including a neutral and a distinguished DDI candidate.

doi:10.1371/journal.pone.0163480.g007

The first feature detects those drug candidates that have the same part of speech and grammatical roles (Object or Subject), and they are separated by “,” or “;” or an “additive transition” word. The second feature detects the case, where both drug names are samples of the same drug category (NeutralCandRule3). In this case, both drug names are mentioned after an introduction additive transition word, and they are also separated by “,” or an “additional additive transition” word.

The idea behind this category of features is that the interaction between two drug names, that have exactly the same “part of speech” and “grammatical role”, cannot be determined by the confident level more than zero. Therefore, the two drugs form a neutral candidate. Although these two features are the only patterns we could detect through analyzing textual language patterns, other similar features could possibly be extracted based on the similar “part of speech” and “grammatical roles” idea.

- NeutralCandRule6-10: Five Boolean features are defined for detecting those DDI candidates that are located in a clause (or sentence) with no additional information to the DDI, i.e. the lack of any investigation. We call these clauses non-informative clauses throughout this paper. Both dependent and independent clauses can be non-informative. Moreover, although non-informative clauses can have negation cue or do not have, the negated clauses have more neutral DDI candidates in comparison with non-negated clauses. For instance, in the following example, the sentence is non-informative, and the interaction between the drugs cannot be determined by the confident level greater than zero; consequently, the identified DDI candidates are neutral:

➤ “Pharmacokinetic interaction trials with cetirizine in adults were conducted by *pseudoephedrine*, *antipyrine*, *ketoconazole*, *erythromycin* and *azithromycin*.”

Taking neutral candidates into account is critical from another perspective, since not doing so may induce conflicts in the corpus later. For instance, in sentence presented in Fig 2, no investigation has actually been conducted into the possible interactions between *Propranolol* and *Ranitidine*, while such an interaction is considered as a negative DDI candidate in DrugDDI corpus. In this situation, the author did not make any remarks about the interaction between the two drugs, and it is possible that in the future, other researchers could find an interaction which would lead the corpus to face conflicts.

Ultimately, it is worth noting that the significant contribution of neutral candidates and features has been reconfirmed in our other research with other corpus [34]. Moreover, it is important to mention that the proposed neutral-related rules can be used with very slight change in other biomedical relation extraction tasks, especially symmetric relations such as protein-protein interaction. The first subcategory of neutral detection rules identify superficial patterns that can be applied to other biomedical domains. However, more patterns can be employed for identifying equivalent names of an entity in addition to the proposed patterns. The second subcategory of neutral features detects candidates that are located in non-informative sentences

which may provide the background information or mention the objectives of the research which are common in biomedical articles. Finally, the third category detects candidates that occur more frequently in symmetric biomedical relations in which every the combination of entities can be a relation candidate.

3.4 Drug-drug interaction prediction

Finally, the proposed method and different components of the system are discussed. The implemented framework is depicted in Fig 8. As the flowchart shows, the sentence, drug names, and negation scopes and cues extracted from the NegDDI corpus are employed as inputs for the three improved methods. Each of the three proposed methods consists of linear combination of the novel proposed features and the substructures of the kernel method (e.g. all tokens for global context kernel and subtrees for subtree kernel). During the experiments, the training parts of the DrugBank and MEDLINE of the corpus was utilized to train the classifiers, and the test part was used to test the system.

A support vector machine with SMO implementation [35] was applied, which performed likewise with *libSVM*, when the best setting of parameters was employed. *Weka* API was utilized as the implementation platform. The tokenization of the text features was executed without stemming process. Furthermore, in all of the above-mentioned methods, all the entities were considered as blind, replacing all the drug names in the generated features with two general terms, i.e. *DrugName* (for the two drugs whose interaction is being investigated) and *OtherDrugNames* (for the other drugs). Tokenization was carried out by the Stanford *BioNLP-Tokenizer* [36] which was adapted with pharmaceutical text, while the Stanford parser was used for constituent parsing. In addition, *TreeTagger* [37] was employed for Lemmatizing and POS tagging which were applied by the winning team in the DDI extraction challenge in 2011.

Results

We first present our results of the comparison between the augmented method and the original method as well as the contribution of different features. Following that, the results of a statistical sign test for characterizing the significance of the obtained improvements will be presented. F-measure is selected as a single performance measure.

It is important to mention that the two datasets of the DDI corpus was utilized due to dealing with different types of texts and language styles [25]. DrugBank texts focus on the description of drugs and their interactions, while the MEDLINE text would not emphasis on DDIs. Table 2 demonstrates some of the basic statistics of the two used datasets.

4.1 Overall comparison of methods

The results of experiments that are similar to the SemEval DDI are presented in this section. In these results the training parts of the NegDDI-DrugBank and NegDDI-MEDLINE of the corpus was used to train the system, and the test parts were utilized to test the system.

Table 3 demonstrates the results for our improved global context (GC), subtree (ST), and local context (LC) kernel methods with NCT features in comparison with the standard methods. Four categories as well as the overall result are presented in that table as well: (i) those candidate sentences in the test part that have negation cue, but do not have any clause connectors, (ii) those with negation cue that have clause connectors, (iii) those without negation cue and with clause connector, and (iv) those without negation cue and clause connectors. The number of tested DDI candidates for each categories of sentence is presented in column four of the table.

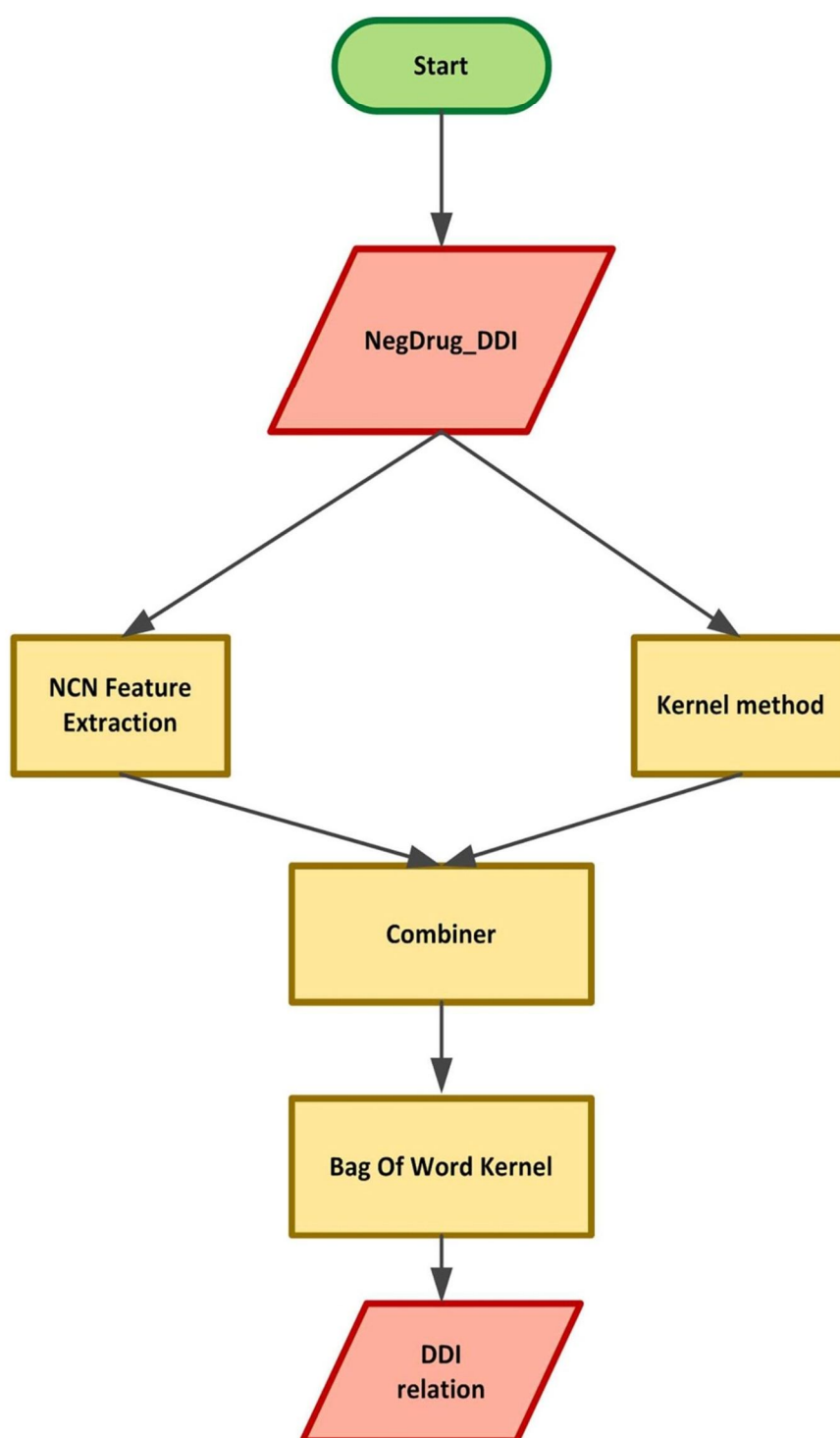


Fig 8. Basic components of the implemented framework.

doi:10.1371/journal.pone.0163480.g008

Table 2. Basic statistics of the two utilized datasets of the DDI corpus.

	MEDLINE			DrugBank		
	Test	Train	Total	Test	Train	Total
Documents	33	142	175	158	572	730
Sentences	326	1301	2308	973	5675	6648
Drug Names	426	1836	2308	2512	12,929	15,441
True DDI candidates	95	232	327	884	3788	4672
False DDI candidates	356	1555	1911	4381	22,217	26,598
Candidates with clause connectors	126	478	604	2067	9215	11,282
Number of Tokens	14,358	61,525	75,883	244,658	1,163,072	1,407,730
DDI Candidates with negation	43	316	359	1367	4558	5925
Total number of DDI candidates	482	2033	1787	5265	31,432	36,697

doi:10.1371/journal.pone.0163480.t002

The best result for the test part of the DrugBank part was achieved by the enhanced local context kernel method (LC+NCT), with 68.4% F-measure which is 2.7% more than the first system in DDI Extraction (2011) challenge (DrugBank part) with an F-measure of 65.7% that was implemented by the University of Trento, Italy.

In the global context and the local context kernel methods, the sentences without negation cues and clause connectors demonstrate the best improvement with an average of +8.1% (6.5% for MEDLINE part) increases in the F-measure. Moreover, in the subtree kernel method, the sentences without negation cues but with clause connectors indicate the best improvement with an average of +15.4% (+5.2% for MEDLINE part) increases in the F-measure.

We conclude that by using the proposed NCT features, not only the sentences with negation cues and clause connectors, but also the other sorts of sentences, including the sentences without negation cues and clause connectors benefit. As elaborated on in section 4.2, the main reason for this finding is neutral candidate features.

4.2 Contribution of each feature set

Table 4 shows that the proposed global context kernel with NCT features has the best performance in sentences that lack negation cues and clause connectors. The best improvement is gained by combining the neutral candidates and clause dependency features in the global

Table 3. F1-measure results for Global Context (GC), SubTree (ST), and Local Context (LC) kernel methods with and without the NCT augmenting features.

	Category		Test Size	GC (%)			ST (%)			LC (%)		
				-	+NCT	↑	-	+NCT	↑	-	+NCT	↑
DrugBank	+Negation	-Connector	971	56.5	62.2	5.7	61.0	68.5	7.5	62.6	65.2	2.6
		+Connector	396	51.7	58.4	6.7	63.2	63.4	0.2	58.0	63.1	4.9
	-Negation	+Connector	1,005	62.3	66.2	3.9	58.6	73.8	15.4	64.8	69.5	4.8
		-Connector	2,893	64.8	72.9	8.1	36.3	39.1	2.8	63.9	69.9	5.8
	Total		5,265	61.7	68.3	6.5	47.1	53.0	5.9	63.4	68.4	4.9
MEDLINE	+Negation	-Connector	198	31.1	39.2	8.1	18.7	22.7	4	39.4	50.6	11.2
		+Connector	161	28.3	38.2	9.9	17.9	23.1	5.2	40.1	50.1	10
	-Negation	+Connector	443	34.6	38.4	3.8	18.9	19.8	0.9	44.2	48.7	4.5
		-Connector	1436	34.1	40.6	6.5	18.2	18.4	0.2	41.7	44.8	3.1
	Total		2238	33.9	38.8	5.9	18.4	21.5	3.1	42.3	48.4	6.1

doi:10.1371/journal.pone.0163480.t003

Table 4. F1-measure results for the global context kernel with combination of different feature sets: Negation scope and cue (N), Clause dependency (C), and neuTral candidate (T).

	Category		Global Context (%)							
			-	+N	+C	+T	+NC	+CT	+NT	+NCT
DrugBank	+Negation	-Connector	56.6	54.9	58.6	66.2	57.8	67.2	59.8	62.1
		+Connector	51.7	52.2	52.9	59.7	52.3	59.8	58.2	58.0
	-Negation	-Connector	64.7	64.8	64.8	71.8	64.8	71.9	71.9	72.9
		+Connector	62.3	62.3	65.3	65.3	63.7	66.4	65.7	65.9
	Total		61.7	61.3	62.9	68.6	62.4	69.0	67.5	68.3
MEDLINE	+Negation	-Connector	31.1	32.6	34.2	37.5	37.2	37.4	38.2	39.2
		+Connector	28.3	33.5	33.5	35.2	38.4	35.8	39.4	38.2
	-Negation	-Connector	34.6	38.7	35.4	35.4	36.1	34.8	37.5	38.4
		+Connector	34.1	38.2	36.7	37.2	35.4	36.4	39.5	40.6
	Total		33.9	36.6%	34.2	36.0	36.8	36.7	38.4	38.8

doi:10.1371/journal.pone.0163480.t004

context kernel, contributing 0.7% (for DrugBank part) more in the improvement process compared with the entire list of the invented features.

Our results concerning the proposed subtree (ST) kernel (Table 5) confirm that the dataset containing sentences without negation cues and with clause connectors has the best performance and the best rate of improvement (15.3% for DrugBank part). Although all feature sets improve the performance of the original subtree kernel, the best combination of features is neutral candidate and negation cue and scope features (Table 5), whose improvement is comparable to that of the entire list of features (15.3%). However, for those sentences containing negation cues, scopes, and connectors, no significant improvement was observed, possibly because the original subtree kernel had good performance for that type of sentences.

Finally, Table 6 indicates that the best combination of feature sets for the proposed local context (LC) kernel is neutral candidate with negation cue and scope features, producing slightly more improvement than the entire list of the invented features (68.5% for DrugBank and 48.3% for MEDLINE part). Furthermore, similar to the global context kernel, due to the consideration of tokens in the original version of the LC, negation scope and cue and clause dependency features generate some duplicated features which reduce the performance of the system. The high performance of neutral candidate features lifts up the overall performance of the feature set up to around +5%. Table 7 presents the f-measure results for test parts of the two used datasets as well as p-values which will be defined in the following section.

Table 5. F1-measure results for the subtree kernel with combination of different feature sets: Negation scope and cue (N), Clause dependency (C), and neuTral candidate (T).

	Category		SubTree (%)							
			-	+N	+C	+T	+NC	+CT	+NT	+NCT
DrugBank	+Negation	-Connector	60.9	59.2	59.9	66.9	68.9	59.9	70.2	68.5
		+Connector	63.2	63.1	62.6	63.2	62.7	63.2	63.1	63.3
	-Negation	-Connector	58.6	62.9	59.7	68.5	59.5	68.4	73.9	73.9
		+Connector	36.3	36.3	36.3	38.7	36.3	38.6	36.3	39.1
	Total		47.1	47.6	47.1	51.4	48.7	50.1	51.6	53.0
MEDLINE	+Negation	-Connector	18.7	19.9	20.2	20.8	19.8	22.6	23.5	22.7
		+Connector	17.9	19.4	19.6	19.6	17.3	18.7	20.7	23.1
	-Negation	-Connector	18.9	18.8	19.8	20.9	19.7	19.7	20.7	19.8
		+Connector	18.2	19.6	19.1	19.8	15.8	18.6	20.6	18.4
	Total		18.4	19.8	19.6	19.9	19.6	20.3	21.4	21.5

doi:10.1371/journal.pone.0163480.t005

Table 6. F1-measure results for the local context kernel with combination of different feature sets: Negation scope and cue (N), Clause dependency (C), and neuTral candidate (T).

	Category		Local Context (%)							
			-	+N	+C	+T	+NC	+CT	+NT	+NCT
DrugBank	+Negation	-Connector	62.6	63.4	62.8	66.0	61.5	65.7	65.6	65.2
		+Connector	58.0	52.2	60.9	67.2	50.9	67.8	64.8	63.1
	-Negation	-Connector	64.8	65.9	64.9	66.2	65.7	66.9	68.9	69.5
		+Connector	63.9	65.3	63.9	69.6	64.2	70.0	69.9	69.9
	Total		63.4	64.1	63.7	68.1	63.0	68.5	68.5	68.4
MEDLINE	+Negation	-Connector	39.4	43.4	49.2	51.2	48.5	46.8	48.6	50.6
		+Connector	40.1	44.2	50.2	48.4	53.8	48.9	50.2	50.1
	-Negation	-Connector	44.2	37.2	42.5	42.6	45.9	52.9	44.7	48.7
		+Connector	41.7	48.4	43.2	49.8	46.1	47.3	48.1	44.8
	Total		42.3	43.5	45.7	46.5	47.7	48.3	47.9	48.2

doi:10.1371/journal.pone.0163480.t006

4.3 Sign test

To verify the significance of the proposed method, a sign test was conducted according to the approach of [22]: $P = P(r(X > Y))$; thus, the null hypothesis: $H_0: P = 0.50$ was tested. For a given random pair of predictions by the original and the corresponding improved method (X_i, Y_i), the null hypothesis states that X_i and Y_i are equally prone to be larger than each other.

For calculating the sign test, we trained the systems with the training part of NegDDI-Drug-Bank and MEDLINE parts and tested them with the test part of the datasets. Table 7 depicts the p-values which state probabilities for accepting the null hypothesis.

In Table 7, column M+ shows the number of correct predictions by the improved method which have been incorrectly predicted by the corresponding original method and are considered a success. Column M- presents the number of correct predictions by the original method which has been incorrectly predicted by the corresponding improved method and is considered a failure. For instance, for the local context kernel, the calculated p-value is the chance of observing 480 successes in 553 trials.

Due to the p-value < 0.0001 in all the sign tests for all experiments, the null hypothesis is rejected and, as a result, all the improvements obtained are statistically significant.

4.4 Error analysis

In this subsection, two categories of errors are presented:

- **Inherent word ambiguities.** Although most of the clause connector features were successfully identified by the proposed system during the superficial features extraction process, few clause connectors features that have alternative speech parts in the sentence were identified

Table 7. The f-score and calculated p-values by sign test for the test parts of the two datasets of the three improved and original methods.

	Method	-	+NCT (%)	M+	M-	p-value
DrugBank	GC	61.7	68.3	425	62	9.0e-53
	ST	47.1	53	395	65	3.6e-71
	LC	63.4	68.4	480	73	3.3e-64
MEDLINE	GC	33.9	38.8	143	35	2.0e-23
	ST	18.4	21.5	129	38	2.3e-32
	LC	42.3	28.2	153	34	3.4e-43

doi:10.1371/journal.pone.0163480.t007

with higher error rate. This happened because the extraction process of the superficial features only considers the structure of the texts rather than their semantics. For example, the connector “that” was the most problematic connector feature, due to the possibility of having different speech parts in the sentence, for example, being also a demonstrative pronoun. Thus, that was not used as a clause connector feature for simplicity.

Other clause connectors feature, similar to that, were considered or ignored, due to the common speech roles they take or do not, in scientific medical articles. For instance, the connector feature “when” was considered only as a connector, a common speech role in the mentioned articles, but ignored as an information question word. Consequently, in minor cases, the value of the feature was set to wrong value.

- **Parentheses.** Another source of inaccuracy in the proposed system as well as many of the text mining systems was parentheses. The error analyses of the system demonstrated higher rate of false positive in sentences with parenthesis. Several reasons contribute to the problem. For instance, parentheses are ignored in the negation annotation process, since the scope of annotation continues and cannot separate parenthesis from other parts of sentences. Consequently, the negation related feature was set to wrong value. For example, in some sentences in DrugDDI corpus, there is a clause or explanation containing the drug name that is placed inside parentheses such as the following sentence:

- Although specific drug or food interactions with *mifepristone* have not been studied, on the basis of the metabolism of these drugs by CYP 3A4, it is possible that *ketoconazole*, *itraconazole*, *erythromycin*, and *grapefruit juice* may inhibit its metabolism (increasing serum levels of *mifepristone*).

Ketoconazole and *mifepristone* are two drug names, which have been annotated as true interaction in the corpus. However, owing to the existence of parentheses, their interaction was not detected by the system. A sentence simplification algorithm could be useful to resolve the parentheses issue.

Discussion and Future Works

In this paper, we studied a list of features including clause dependency features and some features for identifying neutral candidates as well as features extracted from negation cues and scopes. Our experiments indicate that the proposed features improve the performance of the relation extraction task combined with other kernel methods.

The obtained results show that the linguistically-oriented and scope-based negation annotation, which identifies negation cue and scope, does not generally yield sufficient information to decide upon negation confidently in the drug-drug interaction extraction. Therefore, one should regard other factors including identifying neutral candidates and clause dependencies. According to the results, neutral candidate feature set is the most useful among all three feature sets. In addition, better results are obtained from the combination of neutral candidate features with the other two feature sets.

Furthermore, as our analyses of the corpus show, sentences with negation cue have more clause connectors in comparison with sentences without negation cue; therefore, taking account of clause connectors and dependent clauses is important to solve the negation.

A stimulating question that has been partially answered in this work is whether all kernel methods benefit from the proposed features here. As our results of the subtree kernel for sentences with negation cues and clause connectors showed, it is possible that more advanced kernels using more informative features from different presentations of the sentence benefit less

from the proposed features. In few experiments, the complete feature set did not yield the best results in comparison with other possible combinations of features. Thus, a suitable feature selection method can improve the results.

Moreover, in this work, some experiments for using a few basic simplification methods were carried out to overcome the complex sentences; for example by using the main clause as a separate feature, no significant improvement was achieved. However, a future work is trying a combination of simplification and pronoun resolution specified for drugs.

Another motivating future work is extension of the definition of the DDI relation and neutral candidate's confidence level. The extension of the confidence level concept to a membership function for a fuzzy DDI relation instead of a crisp DDI relation will enable us to compare and combine extracted results from different sentences. Dissimilar results for a specific DDI candidate extracted from different sentences with different confidence levels can be compared and combined, which will contribute to identify different types of errors, including systematic or human ones. This can lead to boosting the overall performance of the system, which is not possible with a crisp DDI relation. Speculation and deduction cues including modal verbs of possibility, such as may and related adjective and adverbs, such as likely in addition to the proposed rule-based system to identify neutral candidates can be used to calculate the membership function, i.e. the confidence level.

Author Contributions

Conceptualization: BB.

Data curation: BB AD.

Formal analysis: AD.

Funding acquisition: BB HRC AD.

Investigation: BB.

Methodology: AD.

Project administration: AD.

Resources: BB AD.

Software: BB.

Supervision: AD HRC.

Validation: AD.

Visualization: BB.

Writing – original draft: BB.

Writing – review & editing: HRC.

References

1. Shepherd G., Mohorn P., Yacoub K. and May D. W., "Adverse drug reaction deaths reported in United States vital statistics, 1999–2006," *Annals of Pharmacotherapy*, vol. 46, no. 2, pp. 169–175, 2012. doi: [10.1345/aph.1P592](https://doi.org/10.1345/aph.1P592) PMID: [22253191](https://pubmed.ncbi.nlm.nih.gov/22253191/)
2. Loos E. E., Anderson S., Dwight H D. J., Jordan P. C. and g J. D., *Glossary of linguistic terms*, Camp Wisdom Road Dallas: SIL International, 2004.

3. M. Krallinger, "Importance of negations and experimental qualifiers in biomedical literature," in *NeSp-NLP'10 Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, 2010.
4. Vincze V., Szarvas G., Mora G., Ohta T. and Farkas R., "Linguistic scope-based and biological event-based speculation and negation annotations in the BioScope and Genia Event corpora," *Journal of Biomedical Semantics*, vol. 2, no. Suppl 5, p. S8, 2011. doi: [10.1186/2041-1480-2-S5-S8](https://doi.org/10.1186/2041-1480-2-S5-S8) PMID: [22166355](https://pubmed.ncbi.nlm.nih.gov/22166355/)
5. Harris M. and Rowan K. E., "Explaining grammatical concepts," *Journal of Basic Writing*, vol. 8, no. 2, pp. 21–41, 1989.
6. M. Miwa, R. Saetre, Y. Miyao and J. Tsujii, "Entity-focused sentence simplification for relation extraction," in *Proceedings of the 23rd international conference on computational linguistics*, 2010.
7. Segura-Bedmar I., Mart{\'i}nez P. and S{\'a}nchez-Cisneros D., "The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts," *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction*, vol. 761, pp. 1–9, 2011.
8. R. McDonald, "Extracting relations from unstructured text," *Rapport technique*, Department of Computer and Information Science-University of Pennsylvania, 2005.
9. O. Frunza and D. Inkpen, "Extraction of Disease-treatment Semantic Relations from Biomedical Sentences," in *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, Stroudsburg, PA, USA, 2010.
10. C. Giuliano, A. Lavelli and L. Romano, "Exploiting shallow linguistic information for relation extraction from biomedical literature.," in *EACL*, 2006.
11. S. V. N. Vishwanathan and A. J. Smola, "Fast Kernels for String and Tree Matching," in *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 15*, 2003.
12. Airola A., Pyysalo S., Bj{\'o}rne J., Pahikkala T., Ginter F. and Salakoski T., "All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning.," *BMC bioinformatics*, vol. 9 Suppl 11, 2008. doi: [10.1186/1471-2105-9-S11-S2](https://doi.org/10.1186/1471-2105-9-S11-S2) PMID: [19025688](https://pubmed.ncbi.nlm.nih.gov/19025688/)
13. Tikk D., Thomas P., Palaga P., Hakenberg J. and Leser U., "A comprehensive benchmark of kernel methods to extract protein—protein interactions from literature," *PLoS Comput Biol*, vol. 6, no. 7, p. e1000837, 2010. doi: [10.1371/journal.pcbi.1000837](https://doi.org/10.1371/journal.pcbi.1000837) PMID: [20617200](https://pubmed.ncbi.nlm.nih.gov/20617200/)
14. Chowdhury M. F. M. and Lavelli A., "FBK-irst: A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information," *Atlanta, Georgia, USA*, vol. 351, p. 53, 2013.
15. P. Thomas, M. Neves, T. Rockt{\'a}schel and U. Leser, "WBI-DDI: drug-drug interaction extraction using majority voting," in *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, 2013.
16. Kim S., Liu H., Yeganova L. and Wilbur W. J., "Extracting drug—drug interactions from literature using a rich feature-based linear kernel approach," *Journal of biomedical informatics*, vol. 55, pp. 23–30, 2015. doi: [10.1016/j.jbi.2015.03.002](https://doi.org/10.1016/j.jbi.2015.03.002) PMID: [25796456](https://pubmed.ncbi.nlm.nih.gov/25796456/)
17. He L., Yang Z., Zhao Z., Lin H. and Li Y., "Extracting Drug-Drug Interaction from the Biomedical Literature Using a Stacked Generalization-Based Approach," *PLOS ONE*, vol. 8, no. 6, p. e65814, 2013. doi: [10.1371/journal.pone.0065814](https://doi.org/10.1371/journal.pone.0065814) PMID: [23785452](https://pubmed.ncbi.nlm.nih.gov/23785452/)
18. M. Faisal, M. Chowdhury, A. Lavelli and F. B. Kessler, "Exploiting the Scope of Negations and Heterogeneous Features for Relation Extraction: A Case Study for Drug-Drug Interaction Extraction," in *HLT-NAACL13*, 2013.
19. Siddharthan A., "A survey of research on text simplification," *the International Journal of Applied Linguistics*, pp. 259–98, 2014. doi: [10.1075/ijt.165.2.06sid](https://doi.org/10.1075/ijt.165.2.06sid)
20. Y. Peng, C. Tudor, M. Torii, C. Wu and K. Vijay-Shanker, "iSimp: A sentence simplification system for biomedical text," in *Bioinformatics and Biomedicine (BIBM)*, 2012 IEEE International Conference on, 2012.
21. Segura-Bedmar I., Martinez P. and de Pablo-Sanchez C., "A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents," *BMC Bioinformatics*, vol. 12, no. Suppl 2, p. S1, 2011. doi: [10.1186/1471-2105-12-S2-S1](https://doi.org/10.1186/1471-2105-12-S2-S1) PMID: [21489220](https://pubmed.ncbi.nlm.nih.gov/21489220/)
22. Velldal E., Ovreid L., Read J. and Oepen S., "Speculation and Negation: Rules, Rankers, and the Role of Syntax," *Comput. Linguist.*, vol. 38, no. 2, pp. 369–410, #jun# 2012. doi: [10.1162/coli_a_00126](https://doi.org/10.1162/coli_a_00126)
23. Lappin S. and Leass H. J., "An algorithm for pronominal anaphora resolution," *Computational linguistics*, vol. 20, no. 4, pp. 535–561, 1994.
24. Wishart D., Knox C., Guo A., Cheng D., Shrivastava S., Tzur D., Gautam B. and Hassanali M., "Drug-Bank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic acids research*, 2007. doi: [10.1093/nar/gkm958](https://doi.org/10.1093/nar/gkm958) PMID: [18048412](https://pubmed.ncbi.nlm.nih.gov/18048412/)

25. I. Segura-Bedmar, P. Martinez and M. 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: Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, USA, 2013.
26. G. Szarvas, V. Vincze, R. Farkas and J. Csirik, "The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts," in Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, 2008.
27. N. Konstantinova, S. de Sousa, N. Cruz, M. Maa, M. Taboada and R. Mitkov, "A review corpus annotated for negation, speculation and their scope," in Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 2012.
28. R. Morante and E. Blanco, "**SEM 2012 Shared Task: Resolving the Scope and Focus of Negation," in *SEM 2012: The First Joint Conference on Lexical and Computational Semantics—Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation {(SemEval 2012)}, Montr'eal, Canada, 2012.
29. B. Bokharaeian, A. Diaz, M. Neves and V. Francisco, "Exploring Negation Annotations in the DrugDDI Corpus," in Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, Reykjavik, Iceland, 2014.
30. [Online]. Available: http://nil.fdi.ucm.es/sites/default/files/NegDDI_DrugBank.zip.
31. Wishart D. S., Knox C., Guo A. C., Shrivastava S., Hassanali M., Stothard P., Chang Z. and Woolsey J., "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D668–D672, 2006. doi: [10.1093/nar/gkj067](https://doi.org/10.1093/nar/gkj067) PMID: [16381955](https://pubmed.ncbi.nlm.nih.gov/16381955/)
32. R. J. Mooney and R. C. Bunescu, "Subsequence kernels for relation extraction," in *Advances in neural information processing systems*, 2005.
33. "Pattern (java platform SE 6)," 2015. [Online]. Available: <http://docs.oracle.com/javase/6/docs/api/java/util/regex/Pattern.html>.
34. B. Bokharaeian and A. Diaz, "Automatic Extraction of SNP-Trait Associations from Text through Detecting Linguistic-Based Negation," in 4th Joint Iranian Congress of Fuzzy and Intelligent Systems (CFIS2015), 2015.
35. T. Joachims, "Making large scale SVM learning practical," 1999.
36. D. McClosky, M. Surdeanu and C. D. Manning, "Event Extraction As Dependency Parsing for BioNLP 2011," in Proceedings of the BioNLP Shared Task 2011 Workshop, Stroudsburg, PA, USA, 2011.
37. Schmid H., "Treetagger| a language independent part-of-speech tagger," *Institut fur Maschinelle Sprachverarbeitung, Universitat Stuttgart*, vol. 43, p. 28, 1995.

The distance between insanity and genius is measured only by success.

Bruce Feirstein



SNPPhenA: A Corpus for Extracting Ranked Associations of SNPs and Phenotypes from Literature

RESEARCH

Open Access



SNPPhenA: a corpus for extracting ranked associations of single-nucleotide polymorphisms and phenotypes from literature

Behrouz Bokharaeian^{1*}, Alberto Diaz¹, Nasrin Taghizadeh², Hamidreza Chitsaz³ and Ramyar Chavoshinejad⁴

Abstract

Background: Single Nucleotide Polymorphisms (SNPs) are among the most important types of genetic variations influencing common diseases and phenotypes. Recently, some corpora and methods have been developed with the purpose of extracting mutations and diseases from texts. However, there is no available corpus, for extracting associations from texts, that is annotated with linguistic-based negation, modality markers, neutral candidates, and confidence level of associations.

Method: In this research, different steps were presented so as to produce the SNPPhenA corpus. They include automatic Named Entity Recognition (NER) followed by the manual annotation of SNP and phenotype names, annotation of the SNP-phenotype associations and their level of confidence, as well as modality markers. Moreover, the produced corpus was annotated with negation scopes and cues as well as neutral candidates that play crucial role as far as negation and the modality phenomenon in relation to extraction tasks.

Result: The agreement between annotators was measured by Cohen's Kappa coefficient where the resulting scores indicated the reliability of the corpus. The Kappa score was 0.79 for annotating the associations and 0.80 for the confidence degree of associations. Further presented were the basic statistics of the annotated features of the corpus in addition to the results of our first experiments related to the extraction of ranked SNP-Phenotype associations. The prepared guideline documents render the corpus more convenient and facile to use. The corpus, guidelines and inter-annotator agreement analysis are available on the website of the corpus: <http://nil.fdi.ucm.es/?q=node/639>.

Conclusion: Specifying the confidence degree of SNP-phenotype associations from articles helps identify the strength of associations that could in turn assist genomics scientists in determining phenotypic plasticity and the importance of environmental factors. What is more, our first experiments with the corpus show that linguistic-based confidence alongside other non-linguistic features can be utilized in order to estimate the strength of the observed SNP-phenotype associations. Trial Registration: Not Applicable

Keywords: SNP, Phenotype, Relation extraction, Negation, Modality, Degree of confidence

* Correspondence: behrouz.bo@usm.es

¹Facultad informática, Complutense University of Madrid, Calle Profesor José García Santesmases, 9, 28040 Madrid, Spain
Full list of author information is available at the end of the article



© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

Background

Background

An SNP is a single base mutation occurring at the DNA level. Variations in DNA sequences can affect how humans develop diseases and respond to pathogens, chemicals, drugs, and other agents [1]. There exist an approximate ten to thirty million SNPs in humans [2]. As a result of the increasing number of related articles, the use of automatic association extraction in determining the associations of mutations (e.g. SNPs) and their consequences is increasing in biological systems and genotype-phenotype studies.

In genetic epidemiology, GWA study refers to the process of examining several common genetic variants in different people so as to discover a possible correlation between a variant and a phenotype trait. A phenotype is an organism's recognizable characteristics or traits such as its development, biochemical or physiological properties, behavior, and the concomitant products of that behavior [3]. The large amount of data generated from these studies [4] necessitates the need to develop an automatic approach in order to facilitate the study of the extracted associations. Recently, a few corpora and methods have been developed with the aim of extracting mutation and disease associations from texts such as [5] and [6]. There is, on the other hand, no available corpus for extracting the association of SNP-phenotypes from texts annotated with negation, modality, and the confidence degree of such associations. The need for different levels of annotation for biomedical associations has been considered in certain biomedical resources such as PharmGKB [7]. It collects information about the impact of human genetic variations in drug responses that have been annotated with four levels of evidence.

In this paper, we described and discussed the process of constructing ranked SNP-phenotype association corpus (SNPPhenA), inter-annotator agreement analyses and the results of some utilized baseline methods during an initial experiment. In most cases, implementing a biomedical text-mining system is a difficult task as the basic scientific communication components — i.e. journals and databases — are designed to be read by humans, not machines or computers. In order to address this problem, xml was selected as the main format for the produced corpus. Furthermore, biomedical Natural Language Processing (BioNLP) systems (e.g. relation extraction) have been mostly applied to abstracts as, though concise, they are more readily available. Also, abstracts are deemed as good targets for information extraction (IE) because they are a succinct and summarized version of an article [8], hence the selection of abstracts in the present research.

Motivation

Several named entities have been investigated during the biomedical relation extraction task, few of which are

suitable candidates for annotating with confidence degrees, which is the major aim of the research when identifying the strength (severity) of associations or interactions. The reason for this is that there are no adequate biomedical agreements. For instance, Drug-drug Interactions (DDI) or Protein-protein Interactions (PPI) are two biomedical relations discussed by a myriad of researchers. However, it is difficult even for a human expert to reliably classify the strength or severity of DDIs or PPIs according to confidence level, a problem existing due to the variation in the types of related experiments and the paucity associated with the methods of quantifying and estimating the significance of both the research method and the association. Most GWA studies that report SNP-phenotype associations are generally based on case-control researches [9] initially tested for statistically significant differences between the proportion of exposed subjects among cases and controls. Accordingly, to gauge the research significance of the result, researchers are encouraged to, more often than not, report a level of evidence by considering p -values and study size.

Both preparing a reliable corpus annotated with confidence level in associations and developing an automated tool for this purpose are evidently more difficult for a host of other biomedical named entities that may require different models of study [7]. For instance, comparing and finding an acceptable agreement of confidence level for an association reported in a case-control experiment beside to a case study reported association would be more difficult and challenging. In addition, it is difficult to identify the strength and severity of associations (or interactions) in a sentence explaining a biochemical mechanism occurring in many corpora such as DDI and Protein-related associations because every chemical reaction may precipitate different sequences within the body.

Consequently, insofar as NLP, ranked SNP-phenotype association extraction based on confidence level is considered to be a more feasible task in comparison with many other biomedical association extraction tasks. Additionally, it is worth mentioning that specifying neutral candidates and the effects of negation annotated in the corpus is influenced by measured confidence level of association between two entities, elaborated in the following sections. This shows how crucial it is to have reliable annotations for confidence level in associations as well as an automated method for identifying them.

Yet another objective of the present was to identify the association of such phenotypes as quantitative traits instead of diseases with SNPs, variously studied by researchers. Such extension is significant because many phenotypes can be detected during the sub-clinical phase of a disease history, hence determining their association with an SNP entails a more early diagnosis and treatment

of the disease. Certain phenotypes, it should be noted, are important risk factors for the disease.

Related tasks and phenomena

One of the linguistic-based phenomena discussed in this paper is **negation**. According to linguistics [10], negation refers to a morphosyntactic operation wherein a lexical item or construction is denied or whose meaning becomes inverted by another lexical item. Likewise, the lexical item representing the negation is referred to as the negator. Commonly used in clinical and biomedical text documents, negation is a significant cause of low precision in automated information retrieval systems. In the prepared corpus, the marked sentences were annotated with negation scopes and cues. A sample of a negated sentence can be found in Fig. 1, wherein the SNP and phenotypes are written in bold font.

The other linguistically-driven phenomenon employed here is linguistic **modality**. Generally, modal expressions are words that state modality which is the expression of the subjective attitudes and opinions of the presenter about a possible fact or to control a probable action including intentions, possibility, probability, necessity, obligation [11]. In this research, linguistic-based modals and speculation analyses were made use of in order to determine the confidence level of the SNP-phenotype association candidates in the corpus. The linguistic-based confidence level of an extracted biomedical association can provide an estimate for the reliability of the obtained association and the strength of the biomedical association. Figure 2 demonstrates the sample of a sentence in the corpus with three modality markers. The modality analysis of a sentence and the linguistic-based confidence level of associations can be utilized in addition to other non-linguistic features so as to obtain more accurate annotations.

Named Entity Recognition (NER) is the first step towards extracting associations and relations as well as making related corpora within biomedical texts [12]. It is crucial to notice that the characteristics of NER in the biomedical domain are different from those in the news-wire domain [13]. Identifying mutations in texts is among the most difficult NER tasks in *BioNLP*, investigated in a myriad of studies such as [14–16]. *EMU* is another mutation tagger effective in reducing the annotation time of articles candidate for mutation related associations [17]. It should be noted that implementing a state-of-the-art

automated SNP and phenotype NER is not the objective of this research. Rather, it is the first step toward producing an association extraction corpus, where, the product of the automated algorithm is subsequently checked manually.

The rest of the paper is organized as follows: The next section reviews some of the related works; section three presents the methodology of the paper; section four is dedicated to the evaluation and results; and the last section concludes the paper.

Related works

In this section, we are going to introduce some of the relevant works about preparing the datasets used for extracting mutation related entities including disease as well as different methods of annotating negation and levels of confidence in the biomedical domain.

Mutation association extraction methods and corpora

Besides classical relation extraction tasks in the *BioNLP* domain such as protein-protein and gen-disease, certain novel methods and corpora have been developed with the aim of extracting mutation/polymorphism and disease associations, among which, mention can be made of *BRONCO* [18] and *Variome* [19]. *BRONCO* contains more than four hundred variants and their associations with genes, diseases, drugs and cell lines in the context of cancer, all extracted from 108 full-text articles. *Variome* covers 12 types of relations annotated in 10 full-text articles. While *BRONCO* includes more documents, both corpora annotate several types of relations, such as mutation-disease association, as binary relations on a full-text level. On the other hand, the advantages of abstract-level relation extraction over full-text were mentioned in the introduction section. Therefore, the prepared corpus in this research was provided on an abstract level.

PKDE4J [5] and *Dimex* [6] are two methods for extracting mutation and disease association, the latter being a rule-based unsupervised mutation-disease association extraction working on the abstract level. The *PKDE4J*, however, is a supervised method that employs a rich set of rules to detect the used features. Both methods work on usual binary relations that determine whether or not there exist an association; neither method considers the degree of certainty or confidence [20].

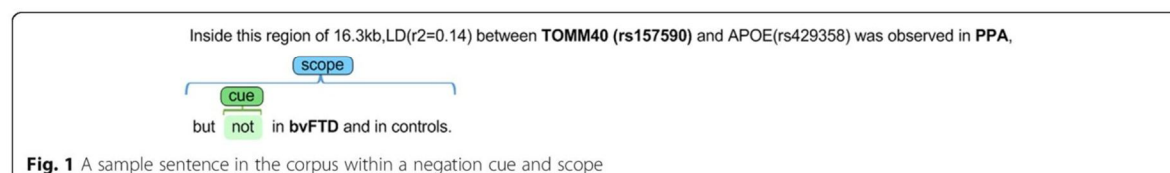
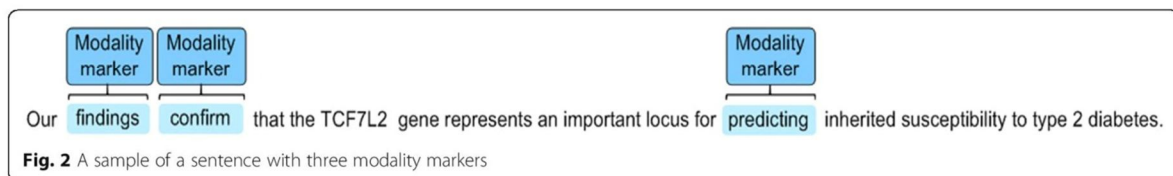


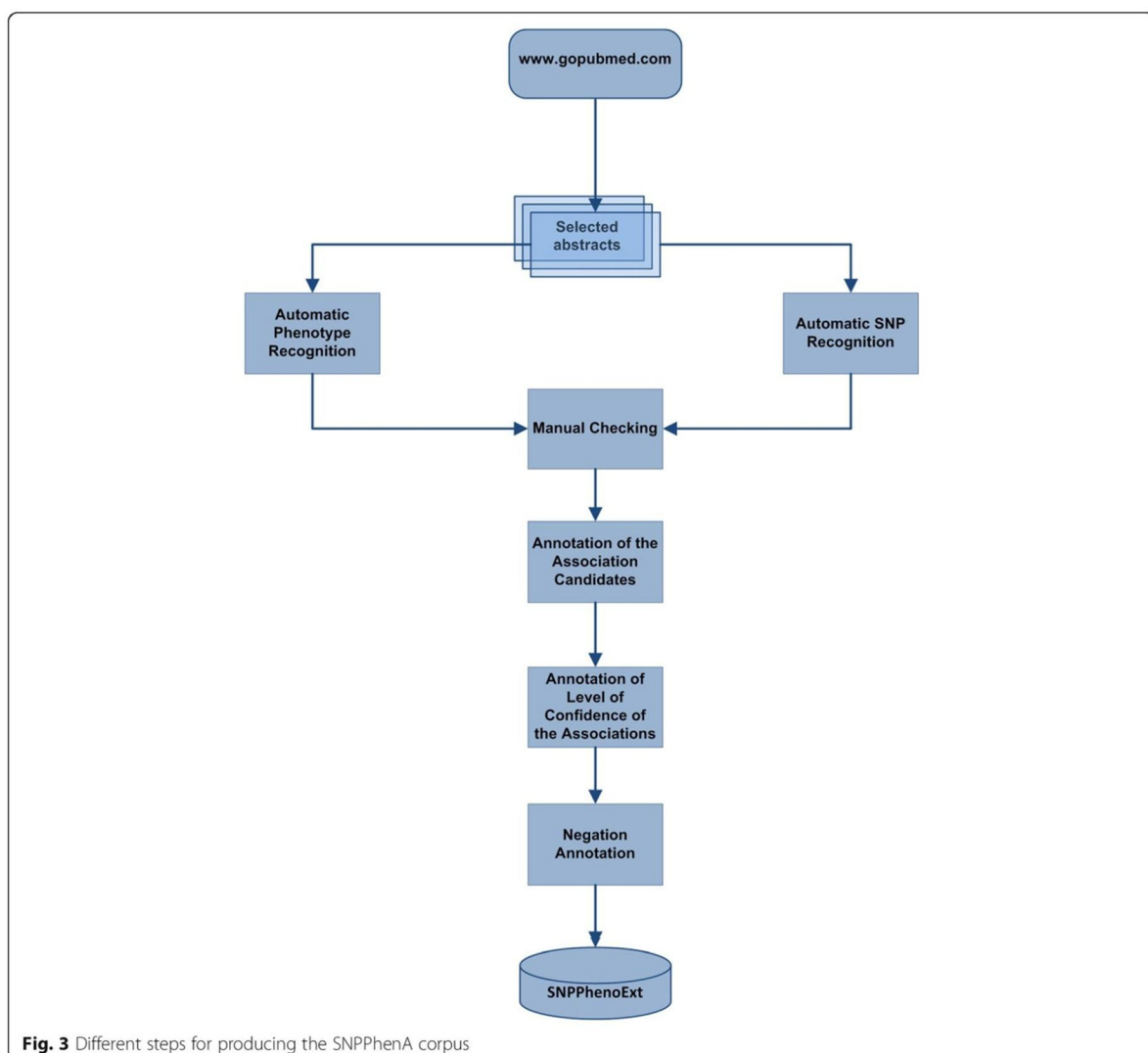
Fig. 1 A sample sentence in the corpus within a negation cue and scope

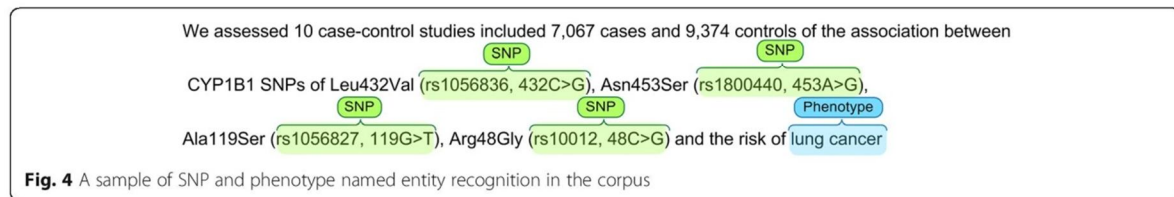


developed another related miner system that gathers heterogeneous data from a variety of literature sources in order to draw new inferences as to the target protein families. Likewise, Ravikumar and his colleagues [21] developed an automated extraction tool in order to obtain protein-specific residue associations from the literature. Another similar automated approach was proposed by [22], which extracts impacts and related

information from literature. In another recent study, Klein et al. proposed the principal infrastructure for the benchmarking of mutation text mining systems [23].

The corpus prepared in this research was annotated with negation cues and scopes, modality markers, and neutral association candidates. Such linguistic features were conducive to the extraction of more accurate information about the extracted SNP-phenotype associations.





Annotating the modality and degree of confidence

As mentioned earlier, “modality” indicates the degree to which a certain observation is possible, probable, likely, certain, permitted, or prohibited. A host of studies have been conducted for the identification of modality and speculation in NLP; very few, however, have been employed for the classification of modality language in bioscience texts.

Although several studies such as [24] have been conducted within the linguistics community as to hedging in scientific texts, in neither is there direct relevance to the task of classifying from an NLP and machine learning perspective.

Light and his colleagues conducted one of the very few direct studies [25], where the speculation identification is introduced using examples from the biomedical domain. They address the question of whether there is sufficient agreement among researches as to what constitutes a speculative assertion that renders the task viable from a computational perspective. Despite the fact that Light attempts to separate the two sides of speculation (strong and weak), he fails to glean sufficient evidence for such a reliable distinction. They conclude that having a reliable distinction between speculative and non-speculative sentences is feasible, and reliable automated methods might also be developed.

Table 1 Some of the most occurred phenotypes in the corpus

Phenotype/phenotypic trait	Num. of abstracts
health risk	40
smoking	33
Obesity	25
metabolic syndrome	16
hypertension	10
insulin sensitivity	9
hypertriglyceridemia	7
glucose metabolism	6
impaired glucose tolerance	5
longevity	4
body mass intake	4
cognitive performance	4
skin pigmentation	3
AIDS	3

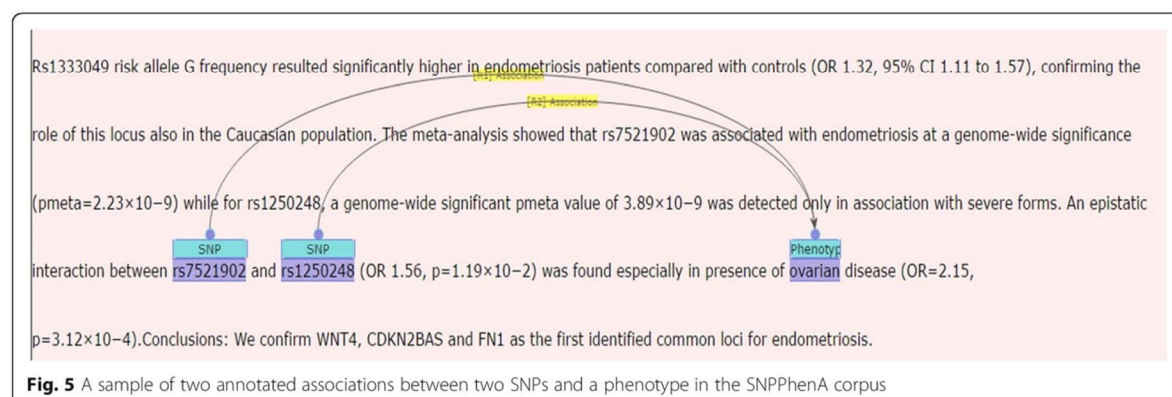
It is noteworthy that in addition to the preponderance of biomedical relation extraction annotations that merely include usual binary association information, there exist certain others containing extra-linguistic information including POS, negation, and speculations information. As an example, the Genia corpus [26], along with biological events, contains annotations for three levels of uncertainty. Nonetheless, to the best of our knowledge, all of the mutation related corpora have only been annotated with binary associations. In the current study, the corpus was enriched through adding more linguistic information such as the linguistic based confidence level of associations, modality markers, and neutral association candidates.

Negation annotation

In general, two negation detection methods have been developed to annotate the employed corpora: A linguistic-based approach and an event-oriented approach. Among other negation annotated corpora, one may refer to the two most well-known: the linguistically-focused, scope-based BioScope [27] and the event-oriented Genia [26]. In BioScope, scopes recognize the position of the key negated event within the sentence, with each argument of the key events coming under the scope, as well. Genia, on the contrary, independently deals with modality within the events. In a Genia event, biological concepts (relations and events) are annotated for negation, yet no linguistic cues are annotated. In fact, the objective of the BioScope corpus is to approach this language phenomenon in a general, task-independent, and linguistically-oriented manner. It can further automatically recognize negation scopes and cues in sentences.

Table 2 Eight of most occurred SNP's in the SNPPhenA corpus and number of contained abstracts

SNP	Number of abstracts
rs12255372	78
rs429358	55
rs7412	46
rs4680	38
rs1051730	25
rs662799	20
rs1799971	18
rs1800629	14



NegDDI-DrugBank is another corpus that was annotated by the authors of the previous work with scopes of negation and negation cues [28]. The automatic extraction of Drug-Drug interactions from the text is held to be highly significant, as two corpus versions (in 2011 and 2013) were prepared in this regard. Concerning the high rate of negated sentences in the DDI corpus, a complete set of sentences within DDI 2011 (with a total of 5806 sentences and 579 files) was automatically annotated with negation scopes and cues. The results were, then, manually checked by three experts to address possible mistakes within the course of the automated process [29]. Adding a new XML negation-tag containing negation cues and negation scopes, the *NegDDI-DrugBank* corpus was established.

Corpus construction

In this section, the steps followed in the construction of the SNPPhenA corpus are explained. The entire process consists of three major steps of collecting documents, automatically and manually recognizing the SNP and phenotypes, and annotating the associations and the related information (Fig. 3). The last step entails annotating the association candidates, the confidence level of associations, the modality markers and the negation scopes and cues of the sentences.

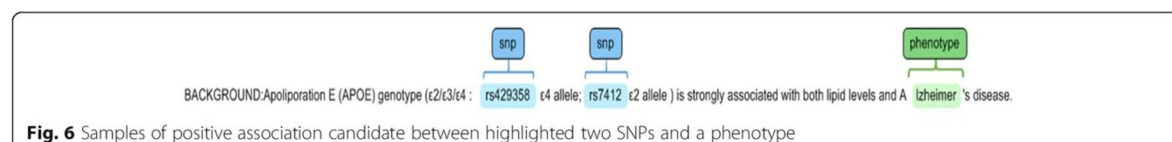
In order to have consistent annotations, all annotators were given the same instruction which includes a pellucid definition of the entities and their relationships, rules and conventions of annotating the confidence level of associations and complete examples for each type of tags. The annotation guideline also contains rules for tackling linguistic phenomena such as negation cues and

modality markers. Moreover, this document presents different types questions raised and retorted by the annotators during the annotation process. The annotation guideline can be found on the website of the corpus.

In the end, 360 XML files were generated comprised of the abstract texts, SNPs, Phenotypes, and the SNP-phenotype associations in the selected sentences. The Phenotypes, SNP names and the association candidates were annotated as xml element tags for each nominated sentence in the abstract. Next, the annotations and the final product were manually checked. The produced SNPPhenA corpus is available for public use¹. So as to better fathom and employ the corpus, brat stand-off annotation format of the files is also available at the website of the corpus. The next subsection is dedicated to the abstracts collection process².

Abstract retrieval

Information provided by the “<http://www.gpubmed.org/>” search engine was used to collect genome-wide association abstracts. *GoPubMed* is a webserver allowing users to explore PubMed search results with Gene Ontology [30]. Twenty popular SNPs were used as query terms enumerated popular by “<http://www.snppedia.com/>” website; the extracted list of abstracts was shortened via selecting those comprised of popular disease names. The list was finally truncated again through choosing those that have candidate sentences consisting of both types of entities. We collected a total of 360 abstracts (including 2625 sentences) with at least one candidate sentence with an SNP and a phenotype name. There were 483 key sentences containing at least one SNP and one phenotype name that were annotated with the xml element



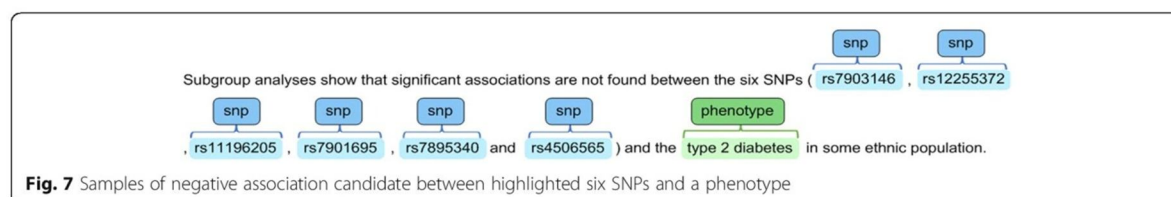


Fig. 7 Samples of negative association candidate between highlighted six SNPs and a phenotype

“SENTENCE”. The total number of SNP names annotated in the SNPPhenA corpus was 875. It is worth mentioning the SNPPhenA is a sentence-level corpus and sentences merely including SNP or Phenotype were not annotated.

The next step was to perform an automatic Named Entity Recognition, followed by a manual checking of sentences with candidate relations for SNPs and phenotype names, as explained in the section below.

Named entity recognition (NER)

An essential part of biomedical NLP is to detect biomedical named entities [31]. During the construction process, two Named Entity Recognitions were done on SNPs and Phenotypes. These two tasks are minutely explained in the two following subsections. A sample of implemented NERs is shown in Fig. 4.

Phenotype NER

A phenotype is the appearance of an organism in terms of its morphology, development, physiology, behavior and its concomitant products [3]. Although there are databases containing disease names and popular phenotype names, no compendious database of phenotypes is yet available.

In this regard, a dictionary-based NER task was implemented by combing two more complete and pertinent databases. The prepared dictionary includes a list from the Comparative Toxicogenomics Database (CTD) for disease names [32]. Also included is the phenotype ontology prepared in the blast project [33]. The collected list of phenotypes includes 65,530 phenotype names along with more than twelve thousand disease names and their synonyms.

The phenotype names were initially recognized automatically by the prepared dataset. Manual checks were subsequently made by two experts in order to identify missed or inexact phenotypes.

A short list of the most frequent phenotypes is shown in Table 1 where the top two phenotypes in the corpus are “health risk” and “smoking”.

SNP NER

The inconsistent description of biological data elements renders the relation extraction tasks challenging. Names associated with polymorphism are particularly problematic because historical or common names are, more often than not, employed instead of standard nomenclature [34], specifically in candidate gene association studies. What is more, it is hard to find the links between historical or common SNP names and refSNP [35]. To address this issue, we implemented a database containing both refSNP(rs) and historical names, matched with their corresponding rsID numbers, while utilizing the *Variant Name Mapper*(VNM) tool [36]. The VNM tool consists of historical names matched with their corresponding rsID numbers extracted from multiple open-access databases, including SNP500Cancer [37], SNPedia [38], pharmGKB [39]. The database was utilized for extracting the different SNP names.

Similar to the phenotype NER process, SNP name annotations were initially checked manually by two biology experts and verified by a third professional annotator. A short list of the most frequent SNPs is shown in Table 2.

Annotating the candidate SNP-phenotype associations

This section deals with the process of annotating the associated candidates which includes the annotation of the SNP-phenotype associations, the confidence level of associated candidates, modality markers, and negation scopes and cues in the negated sentences.

Annotating the SNP-phenotype associations

Following the collection of abstracts and the determination of the SNP and phenotype candidate names, the associations between SNP and phenotype were manually annotated by three gurus in genetics (Fig. 5). The SNP-

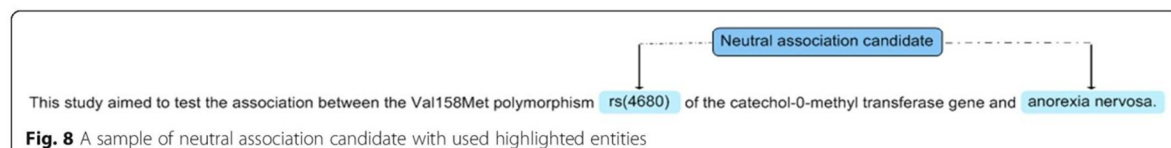


Fig. 8 A sample of neutral association candidate with used highlighted entities

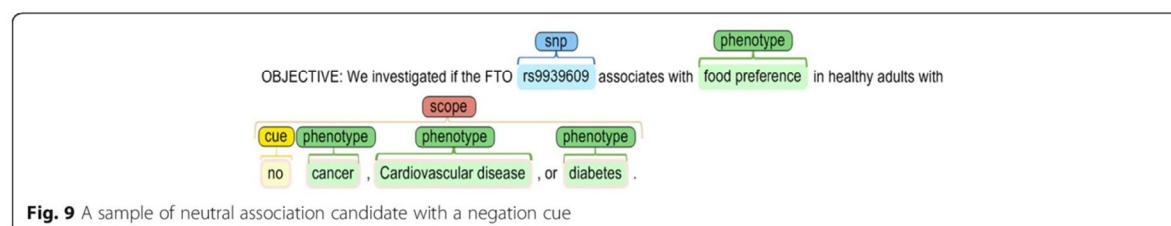


Fig. 9 A sample of neutral association candidate with a negation cue

phenotype candidates were classified into three categories of positive, negative and neutral. The positive SNP-phenotype relation candidates are those with clearly indicated associations (Fig. 6). In contrast, negative SNP-phenotype relation candidates are those in which a lack of association is evident (Fig. 7). In addition to the typical classes of relationships, a neutral class is defined for those that fall between the two other classes, where the presence or absence of association is not remarked in the sentence (see Fig. 8).

As Fig. 8 shows, the presence or absence of association is neither mentioned between “rs4689” and “anorexia nervosa”, nor can it be identified with a high level of confidence, hence, the association between the SNP and the phenotype was annotated as neutral.

In more precise terms, an SNP-Phenotype association candidate is identified as neutral if:

(i) The absence or presence of association between SNP-phenotype cannot be specified from the sentence (or container clause) with a confidence level of more than zero.

(ii) The status of presence or lack of association between the SNP and the phenotype does not change from positive to negative or vice versa if the sentence (or container clause) is negated and SNP and phenotype names are located in the scope of the negation.

(iii) The confidence level of association between SNP and the phenotype does not change if a modal marker is utilized in the sentence and both entities are located in the scope of modality.

The association in Fig. 9, for instance, is neutral and the used negation cue (“no”) does not change the status of the association between the SNP and the phenotypes.

It is worth mentioning that in most relation extraction corpora, neutral candidates were considered to be part of the negative (non-positive) class. Considering them as a separate class of associations allows researchers to conduct different types of experiments. More details as to the role of neutral candidates in biomedical relation extraction tasks can be found in the author’s other study [40].

Similar to the previous steps, the manual checking was initially performed by two experts, and in order to sort out the issue of contradictory confidence levels, the verdict of a third expert annotator was taken into account.

Annotating the level of confidence of the SNP-Phenotype associations

In spite of the fact that genetic components have the instructions for the growth and development of each individual, a person’s phenotype is influenced by environmental factors during embryonic development and throughout life. Environmental factors can stem from a variety of influences such as diet, climate, illness and level of stress. For instance, the capability to taste food is a phenotype estimated, by scientists, to be 85% influenced by genetic inheritance [41]. Nevertheless, environmental factors such as dry mouth or recently eaten food could affect such ability.

“Phenotypic plasticity” is the ability of a genotype to generate more than one phenotype due to various environments [42]. The plasticity is considered to be high if environmental factors have a strong influence. Conversely, if the phenotypic plasticity is low, the genotype can be made use of so as to reliably predict the phenotype. The degree of influence environmental factors have on a person’s ultimate phenotype is, not infrequently, a matter of heated scientific debate.

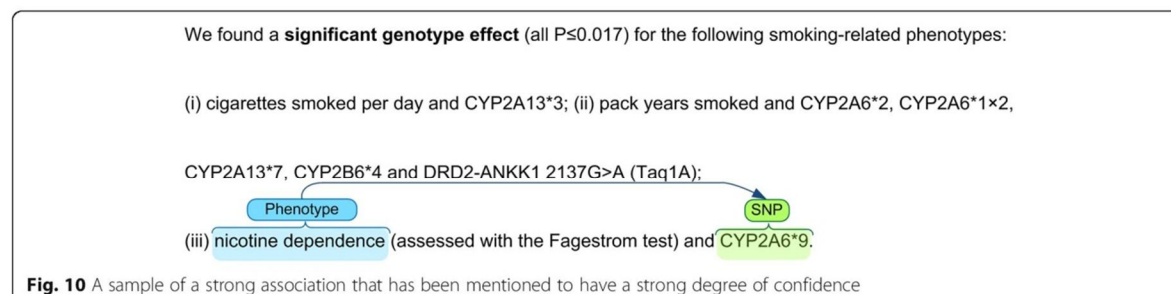
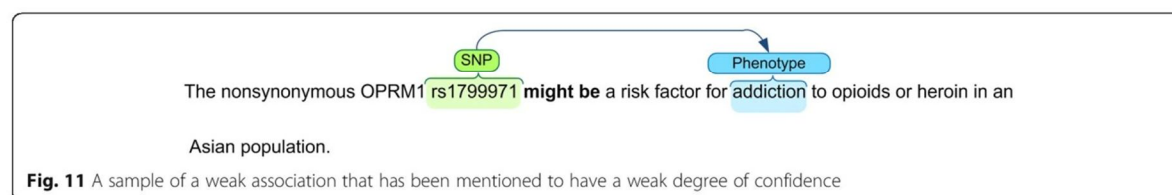


Fig. 10 A sample of a strong association that has been mentioned to have a strong degree of confidence



Differing phenotypic plasticities alongside possible unknown genetic components are the two reasons why GWA study uses confidence level in order to describe the strength of association. The linguistic-based confidence level of the reported association ultimately yields informative data leading to the determination of phenotypic plasticity.

However, there is no available data source or automated method for extracting confidence level from the obtained results. This is when the presence of such a tool and data source is critical and conducive to reviewing literatures.

For this purpose, the confidence levels of positive association candidates in the corpus were annotated by a guru in human genetics. Based on the strength of the linguistic correlation between each individual phenotype and the relevant SNP mentioned in the abstract, the confidence level of associations was categorized into weak, moderate, and strong. Moreover, when the association is neutral (ASSOCIATION = neutral), the degree of confidence is set to “zero”. The confidence levels were assorted considering modality, adverbs and the reported statistical results (p-value). Detailed information about the annotation guidelines can be seen in the guidelines document, available on the website of the corpus. The process, all the same, is demonstrated here via some samples.

The sentence shown in Fig. 10, for example, is considered to have a high confidence level as it indicates “found a significant genotype effect”.

The sample mentioned in Fig. 11, on the other hand, is annotated as having a weak confidence level because of the “might be” clause. However, there exist certain cases that fall under both two categories such as the sample below (see Fig. 12), annotated as moderate.

The annotation of confidence level was carried out by two biology experts both of whom had the same opinion regarding 86% of the association candidates in the whole corpus. In order to sort out the issue of contradictory

confidence levels (14%), the opinion of a third guru annotator was considered.

Linguistic based negation detection and modality markers

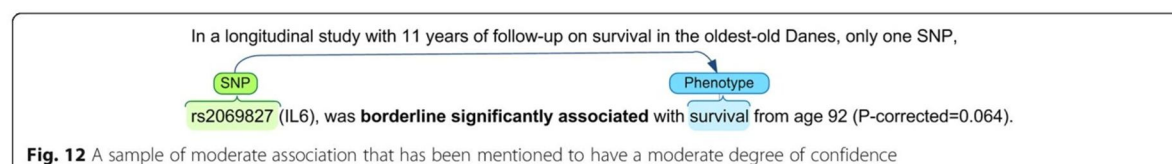
Identifying negative statements is essential in order to obtain accurate information from the text data. The sentence in Fig. 13 demonstrates the importance of considering negation where there is no association between “APOE (rs429358)” and “bvFTD”; however, if the negation had been neglected, an incorrect association might have been identified.

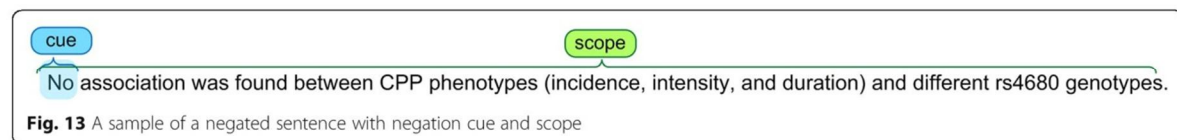
A rule-based system, proposed by [43], was initially utilized in order to annotate the negation scopes and cues. During the process, a set of negation cues such as “not”, “lack”, were detected making use of Bioscope’s guidelines. Negation cues indicate that a negation exists in a sentence. Considering the syntactic context, the scopes of negation and negation cues were subsequently determined, a task already done in a previous work by the authors [28] annotating the DrugDDI 2011 corpus. In order to preclude any possible mistakes, manual checks were made by an expert following the automated process.

In addition to the negation cue and scopes, modality markers were annotated during the annotating process. The employed modality markers obtained from the list were already provided in [44], which is an extension of the list provided by [45] for the biomedical domain. The process includes an automated annotation, followed by an expert performing the manual check. The five more frequent annotated modal markers in the corpus are: “suggest”, “more”, “strong”, “observe”, and “show”.

Evaluation and results

In this section, inter-annotator agreement analyses and the calculated scores are initially presented; then some of the basic statistics of the produced corpus will be





demonstrated; and finally, the results obtained from our first experiment using the corpus are presented.

Inter-annotator agreement

In order to evaluate the quality of the corpus and the reliability of the annotations, the inter-annotator agreement score was measured for the task of classifying candidate sentences into positive, negative and neutral classes, and also for the task of determining the confidence level of the association. As was mentioned before, two annotators had independently tagged the corpus. In the case of disagreement between two tags, a third annotator was asked to decide on the correct one. For the task of classifying candidate sentences, inter-annotator agreement was 91%, which means that in 91% of cases, the two annotators agreed. Additionally, we computed Cohen's Kappa coefficient [46] for the two annotators; this coefficient takes into account the degree of agreement that could be expected to occur by chance and is computed as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Where P_o is the relative observed agreement among annotators, and p_e is the hypothetical probability of chance agreement. The Kappa value was 0.79 for the two annotators. In general, $\kappa=1$ indicates a complete agreement. Furthermore, $\kappa<0$ shows that there is no agreement between annotators other than what would be expected by chance (as given by p_e).

As far as the task of annotating the confidence level of the association with four categories (zero, weak, medium,

strong), annotators agreed in 87% of the occasions; yet the Kappa value was 0.80 which is satisfactory.

Characteristics of the SNPPhenA corpus

This section provides detailed statistics as to the linguistic and nonlinguistic properties of the corpus. The basic properties of the corpus are presented in Table 3 which includes the statistics of the produced corpus in terms of test and training parts. As the table shows, the candidates with a positive association comprise the largest category while the negatively associated candidates constitute the smallest category.

Table 4 provides the detailed analyses concerning the different types of SNP-phenotype association candidates.

Additionally, as mentioned earlier, the key negated sentences in the corpus were annotated with scopes of negation and negation cues. As Table 4 shows, 16.8% of the sentences have at least one negation cue. Further analysis shows that "not" and "no" with respective occurrences of 35 and 38 were the most frequent negation cues. According to the conducted analyses, each sentence in the corpus had an average of 76.9 tokens, 1.7 SNPs, and 1.2 phenotypes.

As illustrated in Table 3, 76.3% of the samples are distinguished (i.e. they are positive and negative association candidates). It can, therefore, be concluded that the annotated sentences were mostly expressed as a direct mechanism or association between one or more SNPs and a phenotype.

Additionally, as Table 4 shows, 63.8% of the candidate sentences have at least one clause connector, while 36.2% do not have one. The result of statistical analysis on the clause connectors further indicates that 9.7% (=87/895) of instances had concessive clauses.

Table 3 Basic statistics of the SNPPhenA corpus in terms of test and train parts

Item	Train	Test	Total
Files	270	90	360
Sentences	1940	685	2625
Key sentences	362	121	483
SNP	691	244	935
Phenotypes	496	158	654
SNP-Phenotype association candidates	935	365	1300
Neutral candidates	142	166	308
Negative candidates	91	29	120
Positive candidates	702	170	872

Table 4 Statistics of different types of SNP-phenotype association candidates in the SNPPhenA corpus

Item	Number	Percentage (%)
Total SNP-phenotype association candidates	1300	100
Candidate with at least one negation cue	218	16.8
Candidates with only one negation cue	188	14.5
Candidates with clause connectors	823	63.8
Candidates without clause connector	470	36.2
Weak degree of confidence candidates	515	39.6
Moderate degree of confidence candidates	124	9.5
Strong degree of confidence positive candidates	233	17.9

Experiment

The results of our first experiments with the corpus are presented in this subsection. Although several mutation-related association extraction methods have recently been developed, automatically measuring the confidence level in an association is a novel task. Consequently, our first experiments were evaluated via certain baseline kernel methods for the two subtasks.

In order to categorize the associations, we employed the two kernel methods that have been expansively made use of in the relation extraction task; the local context kernel [47] and sub-tree kernel [48]. Additionally, the binary Bag of Word (BOW) method was carried out on the corpus so as to predict the degree of confidence for the associations. In all the experiments, the training part of the prepared corpus was used for training the classifier and the test part was employed for testing the system (Tables 5, 6 and 7).

Table 5 shows the performance of the two utilized baseline methods, applied to all three types of candidates. The reported f-score was measured for the detection of positive SNP-phenotype association candidates. Table 6 further indicates the performance of the baseline methods were only applied to the positive and negative association candidates.

The results of the confidence level prediction of associations are presented in Table 7 where the best f-measure is related to the candidate expressions of associations with a weak confidence level, while the worst result is obtained for the moderate confidence level.

The lower performance of identifying the confidence level of association in comparison with the association extraction method demonstrates that the simple features used in the binary BOW may not have enough information to surmount the task and more linguistic features are required. Moreover, the difficulty of the task might be precipitated by the fact that during the annotation process, the annotators employed the mentioned p-value number as a complementary factor for identifying the confidence category, which was the case with 20% of the candidate sentences. It can, accordingly, be concluded that accurately identifying ranked association from biomedical articles requires more linguistic features including dependency parsing, lemmatizing and features related to identifying the significance degree of the biomedical statistical tests.

Table 5 Comparative f-score results for the test SNPPhenA part for two kernel methods with all types of candidates (positive, negative and neutral class)

Method	LCK	Subtree kernel
F1	71.3%	57.7%
Recall	68.7%	51.8%
Precision	69.2%	50.3%

Table 6 Obtained comparative results for the test SNPPhenA corpus for the two investigated kernel methods with non-neutral candidates (positive and negative class)

Method	LCK	Subtree kernel
F1	63.4%	45.7%
Recall	59.8%	41.3%
Precision	56.6%	40.1%

A simple version of the baseline method can be found online ³. It is indispensable to mention that the online system may have a worse performance in comparison with the reported results in this section due to the absence of manual checking during the NER task as well as the omission of the negation detection step.

All the kernel method experiments were carried out by a support vector machine with SMO [49] implementation. Weka API [50] was used as the implementation platform.

Conclusion and future work

In this research, a SNPPhenA corpus was developed in order to extract the ranked associations of SNPs and phenotypes from GWA studies. The process entailed collecting relevant abstracts, Named Entity Recognition, and annotating the associations, negation, modality markers, and the confidence level of the associations.

As opposed to the previous biomedical relation extraction corpora containing true and false types of relations, the annotated associations in the corpus were divided into three classes: positive, negative and neutral candidates. The neutral candidates were those SNP-phenotype candidates that showed no clear evidence as to the presence or lack of association between the SNPs and phenotypes. Identifying neutral candidates is critical for the negation process as the status of such candidates and their corresponding level of confidence do not change when they are located in the scope of negation terms; the status of distinguished association candidates, on the other hand, change in such cases. Similarly, the confidence level, certainty or uncertainty of a neutral candidate, does not change if it is located in the scope of a speculation or modality term. Hence, determining the effect of negation as well as modality terms requires the identification of neutral candidates.

Table 7 Obtained results for the calculating confident interval of the positive association of the test part of the SNPPhenA corpus by bag of words method

Parameter	Weak degree of confidence	Moderate degree of confidence	Strong degree of confidence
F1	69.5%	32.6%	35.3%
Recall	66.4%	30.5%	34.2%
Precision	65.3%	31.6%	32.2%

Not to be forgotten is the fact that the SNPPhenA corpus must be considered as an initial step in extracting graded associations from literature, which could result in the idea of a fuzzy relation extraction task that can be employed so as to construct better biomedical ontologies.

Furthermore, it is important for future researches to employ more linguistic-based and non-linguistic-based factors that could be utilized to determine the confidence of the reported associations. Credibility of the genotyping techniques (such as MLPA or RFLP) and the validity of the research through graph-based network analyses can be employed in the process of identifying the overall confidence level of the reported associations.

Endnotes

¹<https://figshare.com/s/b18f7ff4ed8812e265e8>

²<https://figshare.com/s/f19191317056d6835b38>

³<http://snpphenotypeext-nilg.rhcloud.com/>

Additional file

Additional file 1: Abstract files of SNPPhenA corpus. (ZIP 651 kb)

Acknowledgement

The authors acknowledge Dr. Mariana Neves (Hasso-Plattner-Institute, University of Potsdam, Germany) for her very helpful comments and for advice regarding the usage of brat and pubannotaion tools.

The authors also acknowledge Dr. MT Pilehvar (Cambridge University, UK) for her useful comments and suggestions for organizing the paper.

Funding

Not applicable.

Availability of data and materials

The prepared corpus (SNPPhenA) is available at this address: XML format: <https://figshare.com/s/b18f7ff4ed8812e265e8>; BRAT format: <https://figshare.com/s/f19191317056d6835b38>; Simple online version of the association extractor is available here: (<http://snpphenotypeext-nilg.rhcloud.com/>); Web site of the corpus: <http://nilfdi.ucm.es/?q=node/639>; Annotation guideline: <http://nilfdi.ucm.es/sites/default/files/guidline.pdf>; DTD: http://nilfdi.ucm.es/sites/default/files/SNPPhenA_DTD.zip; Kappa calculation: <https://figshare.com/s/f1fe27ca17022fd4a698>; Document Text Files (Additional file 1): <https://figshare.com/s/47886f335fb0beaf3099>

Authors' contribution

The constructing the corpus was managed by BB, preparing the files of the corpus as well as carrying out the baseline methods was performed by the first author. Moreover the annotating the negation scope and cues was performed by BB. The basic structure of the paper and some details of the experiments and presenting the results were performed by AD. All authors read and approved the final manuscript. NT (University of Tehran, Tehran, Iran) developed a program to optimize the corpus and helped in writing the guideline document. HC (Colorado state university, Colorado, US) helped in preparing the inter-annotator measurement and also preparing and coordinating the two annotator. He also helped in structure of the paper and figures. RC (External Collaborator, Royan Institute for Reproductive Biomedicine, Tehran, Iran) helped in annotation of the corpus as well as give some guidance in biological aspects of the study.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

¹Facultad informatica, Complutense University of Madrid, Calle Profesor José García Santesmases, 9, 28040 Madrid, Spain. ²School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran. ³Department of Computer Science, Colorado State University, Fort Collins, CO 80523, USA. ⁴External Collaborator, Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine, Tehran, Iran.

Received: 5 July 2016 Accepted: 13 January 2017

Published online: 07 April 2017

References

- Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, et al. A general approach to single-nucleotide polymorphism discovery. *Nat Genet.* 1999;23(4):452–6.
- others, I. H. Integrating common and rare genetic variation in diverse human populations. *Nature.* 2010;467:52–8.
- Martin E, and Hine R. *A Dictionary of Biology*, 6 ed. Oxford University Press; 2014.
- Leslie R, O. C. Retrieved May 2016, from GRASP: 2016. <http://grasp.nih.gov/Updates.aspx>. Accessed May 2016.
- Verspoor KM, Heo GE, Kang KY, Song M. Establishing a baseline for literature mining human genetic variants and their relationships to disease cohorts. *BMC Med Inform Decis Mak.* 2016;16(1):37.
- Mahmood AA, Wu T-J, Mazumder R, Vijay-Shanker K. DiMeX: A Text Mining System for Mutation-Disease Association Extraction. *PLoS ONE.* 2016;11(4):e0152725.
- Whirl-Carrillo M, McDonagh E, Hebert J, Gong L, Sangkuhl K. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther.* 2012;92:414–7.
- Seringhaus M, Gerstein M. Manually structured digital abstracts: A scaffold for automatic text mining. *FEBS Lett.* 2008;582(8):1170.
- Lin D, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol.* 2009;33(3):256–65.
- Loos EE, Anderson S, Dwight HDJ, Jordan PC, Wingate JD. Glossary of linguistic terms. *Camp Wisdom Road Dallas: SIL International*; 2004.
- Bybee J and Fleischman S. *Modality in grammar and discourse* (Vol. 32). Philadelphia: John Benjamins Publishing; 1995.
- Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. *Brief Bioinform.* 2005;6(4):357–69.
- Smith L, Tanabe LK, Ando RJ, Kuo C-J, Chung I-F, Hsu C-N, et al. Overview of BioCreative II gene mention recognition. *Genome Biol.* 2008;9 Suppl 2:1–19.
- Thomas P, Rocktaschel T, Hakenberg J, Lichtblau Y, Leser U. SETH detects and normalizes genetic variants in text. *Bioinformatics.* 2016;32(18):2883–5.
- Caporaso JG, Baumgartner WA, Randolph DA, Cohen KB, Hunter L. MutationFinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics.* 2007;23(14):1862–5.
- Wei C-H, Harris BR, Kao H-Y, Lu Z. tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics.* 2013;29:1433–9.
- Doughty E, Kertesz-Farkas A, Bodenreider O, Thompson G. Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. *Bioinformatics.* 2011;27(3):408–15.
- Lee K, Lee S, Park S, Kim S, Kim S, Choi K, et al. BRONCO: Biomedical entity Relation ONcology CORpus for extracting gene-variant-disease-drug relations. *Database.* 2016. doi: 10.1093/database/baw043
- Verspoor K, Yepes A. J, Cavedon L, McIntosh T, Herten-Crabb A, Thomas Z, et al. Annotating the biomedical literature for the human variome. *Database.* 2013. doi: 10.1093/database/bat019.
- Horn F, Lau A, Cohen F. Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics.* 2004;20(4).
- Ravikumar K, Liu H, Cohn JD, Wall ME, Verspoor K. Literature mining of protein-residue associations with graph rules learned through distant supervision. *J Biomed Semantics.* 2012;3:1480–3.
- Naderi N, Witte R. Automated extraction and semantic analysis of mutation impacts from the biomedical literature. *BMC Genomics.* 2012;13(4).
- Klein A, Riazanov A, Hindle M, Baker CJ. Benchmarking infrastructure for mutation text mining. *J Biomed Semantics.* 2014;5:11.

24. Kim LC, Lim JM-H. Hedging in academic writing - a pedagogically-motivated qualitative study. *Procedia Soc Behavioral Sci.* 2015;197:600–7.
25. Light M, Qiu X, Y, & Srinivasan P. The Language of Bioscience: Facts, Speculations, and Statements in Between. *Linking Biological Literature, Ontologies and Databases.* Glasgow; 2004. pp. 17–24.
26. Tateisi Y, Yakushiji A, Ohta T, and Tsujii J. Syntax Annotation for the GENIA corpus. *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-2005), Jeju Island, Korea, October, 2005.* pp. 11–13.
27. Vincze V, Szarvas G, Farkas R, Mora G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics.* 2008;9(11):1.
28. Bokharaeian B, Diaz A, Neves M, and Francisco V. Exploring negation annotations in the DrugDDI Corpus. *Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BIOTxtM 2014).* 2014. Reykjavik.
29. Bokharaeian B, Diaz Esteban A, Ballesteros Martinez M. Extracting Drug-Drug interaction from text using negation features. *Procesamiento del Lenguaje Natural.* 2013;51:49–56. Madrid.
30. Doms A, Schroeder M. GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Res.* 2005;33 suppl 2:W783–6.
31. Nadeau D, Sekine S. A survey of named entity recognition and classification. *Linguisticae Investigationes.* 2007;30(1):3–26.
32. Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegerts TC, Mattingly CJ. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic Acids Res.* 2009;37 suppl 1:D786–92.
33. SF, A., W., G., W., M., EW, M., & DJ., L. Retrieved may 2016, from Basic Local Alignment Search Tool (BLAST): 2015. <https://blast.ncbi.nlm.nih.gov/Blast.cgi>. Accessed May 2016.
34. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* 2000;28(1):352–5.
35. Nicolazzi E, Caprera A, Nazzicari N, et al. SNPchip v. 3: integrating and standardizing single nucleotide polymorphism data for livestock species. *BMC Genomics.* 2015;16:283.
36. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. A navigator for human genome epidemiology. *Nat Genet.* 2008;40(2):124–5.
37. Packer BR, Yeager M, Burdett L, Welch R, Beerman M, Qi L, et al. SNPS00Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Res.* 2006;34 suppl 1:D617–21.
38. Cariaso M, Lennon G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res.* 2012;40(D1): D1308–12.
39. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart J, Altman R, et al. PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res.* 2002;30(1):163–5.
40. Bokharaeian B, Diaz A, Chitsaz H. Enhancing extraction of drug-drug interaction from literature using neutral candidates, negation, and clause dependency. *PLoS ONE.* 2016;11(10):e0163480. doi:10.1371/journal.pone.0163480.
41. Wooding S, Kim UK, Bamshad MJ, Larsen J, Jorde LB, Drayna D. Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. *Am J Hum Genet.* 2004;74(4):637–46.
42. Price TD, Qvarnstrom A, Irwin DE. The role of phenotypic plasticity in driving genetic evolution. *Proc Biol Sci.* 2003;270(1523):1433–40.
43. Ballesteros M, Francisco V, Diaz AJH, Gervas P. Inferring the Scope of Negation in Biomedical Documents. *13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012).* New Delhi: Springer; 2012. p. 363–75.
44. Thompson P, Venturi G, McNaught J, Montemagni S and Ananiadou S. Categorising modality in biomedical texts. *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining.* 2008. pp. 27–34.
45. Hyland K. Talking to the Academy: Forms of Hedging in Science Research Articles. *Writ Commun.* 1996;3(2).
46. Pustejovsky J and Stubbs A. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications.* O'Reilly Media. 2012.
47. Giuliano C, Lavelli A, Romano L. Exploiting shallow linguistic information for relation extraction from biomedical literature. *EACL.* 2006;18:401–8.
48. Tikk D, Thomas P, Palaga P, Hakenberg J, Leser U. A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLoS Comput Biol.* 2010;6(7):e1000837.
49. Joachims T. Making large scale SVM learning practical. In: *Advances in kernel methods.* Cambridge, US: MIT Press; 1999. p. 169–84.
50. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsl.* 2009;11(1):10–8.

Submit your next manuscript to BioMed Central
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit



10

chapter

Extraction of Ranked SNP-Phenotype Associations from Literature through Detecting Neural Candidates, Negation and Modality Markers

When I look upon seamen, men of science and philosophers, man is the wisest of all beings; when I look upon priests and prophets nothing is as contemptible as man.

John F. Tierney

Automatic Extraction of Ranked SNP-Phenotype Associations from Literature through Detecting Neural Candidates, Negation and Modality Markers

. Behrouz Bokharaeian and Alberto Diaz

1. Facultad informatica, Complutense University of Madrid, Madrid, Spain, Email: behrou.bo@usm.es

2. Facultad informatica, Complutense University of Madrid, Madrid, Spain, Email: albertodiaz@fdi.ucm.es.

Abstract

Genome-wide association (GWA) constitutes a prominent portion of studies which have been conducted on personalized medicine and pharmacogenomics. Recently, very few methods have been developed for extracting mutation-diseases associations. However, there is no available method for extracting the association of SNP-phenotype from text which considers degree of confidence in associations. In this study, first a relation extraction method relying on linguistic-based negation detection and neutral candidates is proposed. The experiments show that negation cues and scope as well as detecting neutral candidates can be employed for implementing a superior relation extraction method which outperforms the kernel-based counterparts due to a uniform innate polarity of sentences and small number of complex sentences in the corpus. Moreover, a modality based approach is proposed to estimate the confidence level of the extracted association which can be used to assess the reliability of the reported association.

Keywords: *SNP, Phenotype, Biomedical Relation Extraction, Negation Detection.*

1. INTRODUCTION

A **single-nucleotide polymorphism (SNP)** is a single base mutation that happens in DNA-level [1]. Variations in the DNA sequences can affect how humans develop diseases and respond to pathogens, chemicals, drugs, and other agents. The first successful GWA study dates back to 2005 when Klein and his colleagues carried out the first successful GWAS on patients with age-related macular degeneration. It was the start of a worldwide trend which results in finding thousands SNP associations. Fig 1 shows the increasing numbers of papers that have been published in this field from the year 2001 to 2014 obtained from a PubMed search engine for the query ‘Single Nucleotide Polymorphisms’ (performed in November 2015). SNPs are also important for personalized medicine.

Recently, few methods have been developed recently for extracting mutation and disease associations from text such as [2] and [3]. However, there is no available method for extracting the association of SNP-phenotype from text which consider the neutral candidates and the level of confidence of associations.

A **phenotype** is the organism's recognizable characteristics or traits, such as its development, biochemical or physiological properties, behavior, and products of behavior [4]. An SNP can be “associated” with the phenotype when a specific type of variant (one allele) is frequent within samples obtained from subjects. The degree in which phenotype is determined by genotype is referred as “phenotypic plasticity” [5].

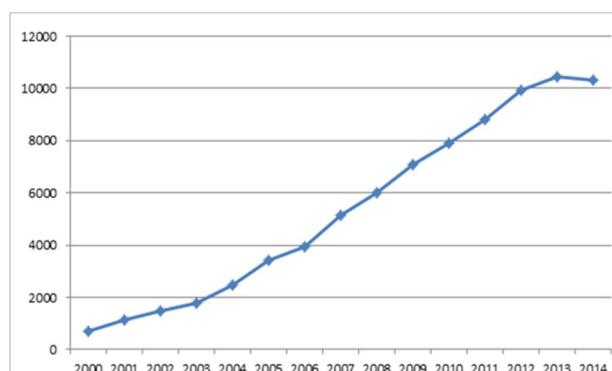


Figure 1. Number of ‘Single Nucleotide Polymorphisms’ publications from 2000 to 2014 in PubMed.

The amount of influence that environmental factors have on a person’s ultimate phenotype is a matter of serious scientific debate.

On the other hand, one of the essential tasks in biomedical text mining is to identify negations which is the more important feature used in our approach. Linguists define **negation** as a morphosyntactic operation [6]. Through this operation a lexical item either denies or inverts the meaning of another item or construction. The importance of negation in biomedical text mining is revealed when we consider the fact that negation is very common in those texts leading to lack of precision in automatic information retrieval systems [7]. For example in the sentence below, there is not any association between “*APOE polymorphisms*” and “*serum HDL-C*”; however, if negation is neglected a wrong association might be identified:

- There were <{ no} associations between *APOE polymorphisms* and *serum HDL-C*, *APO-CIII* and triglycerides>

Linguistic **modality** is another linguistically-driven phenomenon going to be applied in this research. In general, modals are special words stating modality, which expresses the internal attitudes and beliefs of the announcer such as facility, probability, inevitability, commitment, permissibility, capability, wish, and contingency [8]. In current study, we aim to use modals based on linguistic- and speculation analyses for determination of the confidence and strength of the stated SNP-phenotype associations in the corpus.

On the other hand, despite **distinguished** association candidates which include remarks made by the author, a **neutral candidate** does not contain any remarks [10]. In Fig 2, relation status between “*anorexia nervosa*” and “*rs4680*” is neutral since the author has not mentioned their association. In other words, any conclusion about the association between these two entities is not possible with this sentence. McDonald *et al.* are one of the very few groups of researchers, who have investigated

the neutral candidates in relation extraction task [9]. More information about the neutral relation candidates their important role in the biomedical domain can be found at the other work of the authors [10].



Figure. 2. A sample of neutral association candidate with used entities specified with circle

In addition to the pervious subjects, **innate polarity** of the sentences about a relation is an important factor that must be taken into account. However, to the best of our knowledge, no research has been conducted on the effect of innate polarity of the sentences on a relation extraction task.

However, innate polarity of a sentence speaking about a relation expresses whether the assumed relation candidate in the sentence without negation cue and scope exists or not. For instance, the first sample below gives a positive innate polarity on SNP-Phenotype [22], while the second sample provides a negative one.

- The nicotinic acetylcholine receptor polymorphism (**rs1051730**) on chromosome 15q25 is associated with major **tobacco-related diseases** in the general population with additional increased risk of COPD as well as lung cancer.
- We investigated the causal relationship between smoking and symptoms of anxiety and **depression** in the Norwegian HUNT study using the **rs1051730** single nucleotide polymorphism (SNP) variant located in the nicotine acetylcholine receptor gene cluster on chromosome 15 as an instrumental variable for smoking phenotypes.

In this study, we suggest a text-mining approach which extracts association between SNP and phenotypic Phenotypes. The rest of this paper is organized as follows. Section 2 introduces some related research works. The proposed method is explicated in section 3. Afterwards, section 4 presents results and statistical analysis. Finally, section 5 concludes the paper while providing suggestions for further research.

2. Related works

Besides classical relation extraction tasks in the BioNLP domain such as protein-protein and gen-disease tasks, some new methods and corpora been developed for extracting mutation/polymorphism and disease associations. DiMex [3] is a rule-based unsupervised mutation-disease association extraction that works on the abstract level. The PKDE4J [2] is a supervised method that employs a rich set of rules to detect the used features. Another related miner system has been developed by [15] that gather heterogeneous data from a variety of literature sources in order to draw new inferences about the target protein families.

Moreover, one of the few researches that took **negation** into account in the relation extraction task was [16]. In the method, SVM classifier was fed using a list of features such as nearest verb to

candidate entity in the parse tree and some negation cues. Pyysalo *et al.* [18] have conducted a survey wherein negation and **uncertainty** issues were taken into account. They stated that among those corpora **BioInfer** has negative annotation. Numerous studies have been conducted on modality and speculation of identification in NLP [19], but only a few of this research have been employed for classifying speculative language in the bioscience texts. In biomedical study: the vocabularies could be involved in theories, experimental results, hedges, and speculations. Though some studies have been performed within the linguistics community on the use of hedging in scientific text like [20], there is little direct relevance for categorizing task using the perspective of NLP/ML.

2. Method

The proposed association extraction method relies on detecting linguistic-based negation and neutral candidates which are introduced in this section. The basic components of the algorithm can be seen in the flowchart in Fig. 3.

In this section, the process of detecting SNP-phenotypes associations is explained. It is worth mentioning that we have used the SNPPhenA corpus during the research which has been introduced previously [20]. The corpus is available for public use¹.

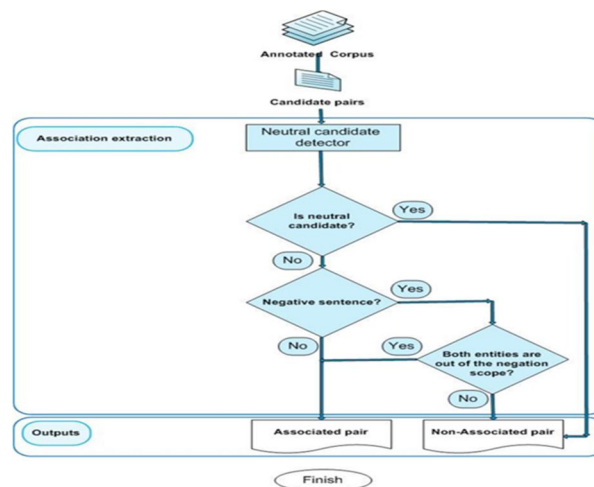


Figure. 3. Flowchart of the different steps of the snp-phenotype association extraction proposed algorithm

3.1. Verifying the Criteria of the corpus

To examine whether the proposed negation based method is applicable to the corpus or not some metrics must be analyzed which are known as verification criteria (See Fig. 4):

- **Complexity of the sentences:** As mentioned in previous sections, complex sentences form a major source of inaccuracy. They reduce the performance of the algorithm in two ways. Firstly, they decrease the performance of the automatic negation detection algorithm; and secondly,

¹<https://figshare.com/s/b18f7ff4ed8812e265e8>

dependent clauses can change the meaning of main clause as mentioned earlier for concessive clauses. Additionally, the number of prominent clause connectors and average number of tokens can be utilized to measure complexity of a sentence.

- **Uniform innate polarity of the sentences regarding to SNP-Phenotype association:** Innate polarity is an important factor in identifying relations from the text. Therefore, the produced corpus is analysed to derive the number of positive and negative innate polarity samples. For an estimation of the ratio of innate positive and negative polarities in the corpus, candidate sentences without negation cue were identified. Selected candidates that express no association between discussed SNP and Phenotype were classified as negatives and the other were identified as positive.

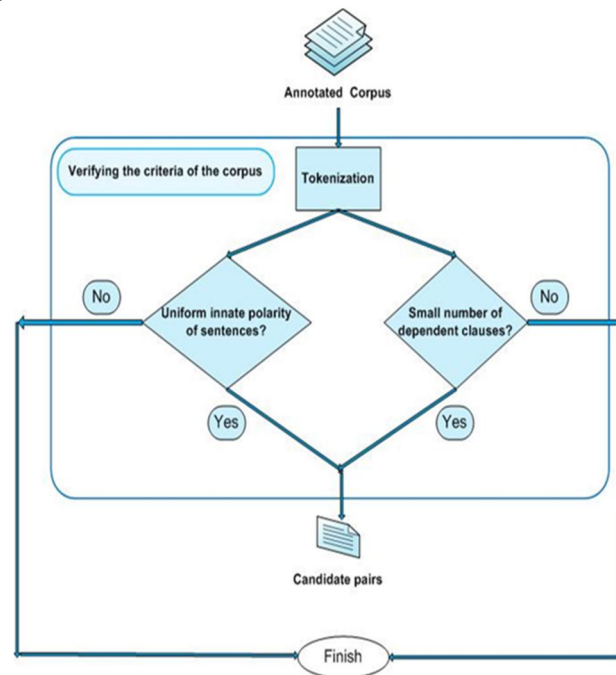


Figure. 4. Verifying the criteria of the corpus

3.2. Proposed association extraction approach

For developing the proposed approach, six Boolean features were extracted from negation cues and scope which have been used. Additionally, to determine possible effects of negation on SNP-Phenotypes relation, **neutral examples** have been identified in the corpus. Negation inverts status of positive or negative relations candidates which are in the negation scope while leaving neutral ones unchanged. As a result, the ratio of neutral candidates to positive or negative ones is of a great significance.

3.2.1. Neutral candidate detector

For automatic detection of neutral candidates we have implemented a neutral candidate detector, As initial experiments shows detecting the neutral candidates are very important in the negation based method. Consequently a neutral candidate detection system has been carried out. The proposed method worked with a **global context method** kernel method, the prepared corpus has been used for training and predicting the neutral candidates.

However, in case of neutral candidates, negation does not change the status of the association and it will remain false. Because of few numbers of neutral candidates in the produced corpus, considering the neutral candidates as negatives still leads to superior performance as will be seen in next section (Table 10). As a result, if the status of the existence of neutral candidate was defined as

- “IsNeutralCand” A Boolean feature which is set as true when association candidate predicted as neutral, while other situation is false.

3.2.2 Negation based association extraction method

As for relation extraction, it must be noticed that negation does not necessarily change the status of a relation between entities. As a matter of fact, the effect of negation on association depends on several factors among which position of entities relative to the negation scope and cue can be directly extracted from extended corpus. For example, consider the following sentence:

- Moreover, the *rs1051730* variant may **not** merely operate as a marker for ***dependence or heaviness of smoking***.

“Dependence or heaviness of smoking” is a phenotypes name inside the negation scope, so as their association relation between SNP (*rs1051730*) and the phenotype name is inverted by the negation. There are 6 different possibilities based on position of SNP and phenotype names relative to the negation scope which are used as 6 features:

- BothInsNegSc: A Boolean feature which is set as true when both SNP and phenotype names are inside the negation scope, while other situations are false.
- OneLeftOneInsNegSc: A Boolean feature which is set as true when one SNP or phenotype name is on the left side (out) of the negation scope and the other one is inside the negation scope, while other situations are false.
- OneRightOneInsNegSc: A Boolean feature which is set as true when SNP or phenotype name is on the right side (out) of the negation scope and the other one is inside the negation scope, while other situations are false.
- Three other Boolean features related to other possibilities.

As table 1 demonstrates and also we have mentioned earlier, almost all of sentences have positive polarity; hence negation can change the relation status from True to False. Consequently, as it is indicated in Fig. 3 if the studied candidate is not a neutral, and one of these three Boolean features (BothinsideNegSc, OneLeftOneInsideNegSc or OneRightOneInsideNegSc) are true, the test association is predicted as false, whereas, any other combination of features lead to a true association.

However, In case of neutral candidates, negation does not change the status of the association and it will remain false. Because of few numbers of neutral candidates in the produced corpus, considering the neutral candidates as negatives still leads to superior performance as will be seen in next section. As a result, if the status of the existence of neutral candidate was defined as

- “IsNeutralCand” A Boolean feature which is set as true when association candidate is neutral, while other situation is false.

The status of association can be calculated as below:

- $$SNPTraitAssociation = (BothInsNegSc \vee OneLeftOneInsNegSc \vee OneRightOneInsideNegSc) \wedge \neg IsNeutralCand$$

We compare the proposed negation neutral based algorithm (NNB) with the three kernel methods. The used kernel methods are global context kernel, local context kernel and subtree kernel. All of the three used kernel methods were trained with train part of the prepared corpus and were tested with test part.

In the next section, we will present the results obtained by the proposed method as well as those given by the kernel methods, so as a comparison can be made.

All of the kernel method experiments were carried out by a support vector with SMO [21] implementation. According to the experiments conducted via SMO approach and comparing the results to those of other implementations of SVM, e.g. libSVM, it was evident that SMO implementation was associated with a faster and better performance. Weka API was used as the implementation platform. A sample version of the proposed system is available online at the address (<http://snpphenotypeext-nilg.rhcloud.com/>).

3.3. Identifying level of confidence of SNP-phenotype association

There are genetic instructions for growing and developing all individuals, but environmental parameters also influence on the phenotype of a person through embryonic growth and life. Environmental parameters can be resulted by a range of effects including nutrition, weather, and disease and stress level. For example, the ability of tasting food is a phenotype, which is estimated as 85% affected via genetic inheritance [22]. On the other hand, this ability could be intervened by environmental parameters including dry mouth or lately eaten food.

The degree in which phenotype is determined by genotype is referred as “phenotypic plasticity” [23]. However, phenotypic plasticity is considered high if environmental factors have a strong influence. Conversely, if phenotypic plasticity is low if genotype can be used to reliably predict phenotype. Overall, the amount of influence that environmental factors have on a person’s ultimate phenotype is a matter of serious scientific debate.

Different phenotypic plasticity as well as other effective unknown genetic components presents two explanations for why a GWA study reports on the importance of degree of confidence for these associations. Consequently, the linguist-based confidence of the reported association will have informative data leading to determination of phenotypic plasticity.

However, there is no available data source or automatic method for extracting level of confidence of the obtained results. Consequently, the presence of such a tool and data source is critical and can be applied to help researchers in reviewing the literature.

We have implemented a modality based supervised method (MMS) for identifying the level of confidence of the extracted association. The proposed method consists of a classifier initially was trained by the related modal markers, the mentioned p-value and the confidence level of the sentence which have been annotated in the corpus. And during the test phase initially modal markers and the container clause were identified. If the sentence doesn't have any modal markers or the entities were not located in the clause that contains the modals, the confident level will set to medium. Otherwise the level of confidence was determined by the trained classifier using the identified modal markers of the candidate sentence.

4. Evaluation

In this section after presenting some statistical analysis regarding the number of different entities and linguistic-based negation cues and clause connectors in the corpus used for evaluation, cooperative validation results are demonstrated. We carried out two types of experiments, first the proposed method were carried out on train data set and were tested using test part and secondly 10 fold cross validation on the whole corpus. We have used three supervised kernel methods as benchmark. For this purpose, the support vector machine was used for this purpose.

The results revealed that the proposed method is superior to counterpart kernel methods. Besides, it eliminates the need for training data avoiding difficulties associated with this step done mostly by related experts.

- **Low proportion of complex sentences in the corpus.** The result of statistical analysis on clause connectors shows, 9.7% (=87/895) of the instances have concessive clauses. Furthermore, considering the table, it could be concluded that the most frequent connectors are “but” and “after”. Additionally, two third of the candidates have clause connector but this ratio is not significant in biomedical domain. While considering that biomedical scientific manuscript contain complex sentences usually mention different situation and condition. However, according to table 6, the average ratio of SNP and phenotype names per sentences is also weak.
- **Similar innate polarities of the sentences,** the polarity analysis shows that most of the sentences have innate positive polarity indicating an “association” between SNP and a phenotype. It is worth mentioning that the polarity analysis regarding associations was carried out on sentences without negation cue (Table 1). For instance the sentence beweak explains an “associated with” implication. Consequently, it has a positive innate polarity proving the existence of an association between operands:
 - In haplotype analysis, the haplotype combination of **rs2254298** A allele, **rs2228485** C allele and **rs237911** G allele was found to be significantly associated with an increased risk of **preterm birth** (OR=3.2 [CI 1.04-9.8], p=0.043).

4.1. Identifying the associations

In this section, the comparative results of the proposed method and local context kernel are presented in terms of F-score to calculate the positive classes.

Table 1. Obtained comparative results for the proposed negation neutral based method (NNB) for the test corpus alongside to the obtained results for the three investigated kernel methods with non-neutral candidates (positive and negative-neutral class)

Method	LCK	Subtree kernel	NNB
F1	60.3%	45.7%	75.6%
Recall	56.7%	41.3%	79.6
Precision	53.5%	40.1%	75.4

Table 2. Obtained comparative 10 fold cross validation results for the proposed NNB method for the LCK kernel methods with two categories of candidates (positive and negative-neutral class).

Method	LCK	Subtree kernel	NNB
F1	94.2	91.5	97.4

The experiments were carried over two groups of the candidates. During the experiments whose results are shown in Tables 1 and 2, neutral candidates have been considered as a part of negative class of candidates as other relation extraction corpora.

To evaluate performance of our proposed method two other schemes are tested as well, namely, local context and subtree kernel methods. As it is shown in Table 1 and 2 the proposed method outperforms the mentioned schemes even when neutral class of samples is ignored. Moreover as tables show, local context kernel shows better performance in comparison with subtree kernel.

The role of neutral samples identification in improving performance of the NNB can be understood via examining Tables 1 and 2. However, table 2 indicates f-measure values for all candidates in the corpus including positive, negative as well as neutral ones.

4.2. Forecasting level of confidence

In addition to the performed experiments for predicting the SNP-phenotype associations, a binary Bag of Word (BOW) method was performed over the corpus as a baseline method to predict degree of confidence for associations

Table 3. Obtained results for the calculating confident interval of the positive association of the test part of the SNPhenA corpus by Bag Of words and the proposed MMS method.

Parameter		Low Level of confidence	Middle Level of confidence	High Level of confidence
BOW	F1	64.2%	16.3%	26%
	Recall	64.6%	14.5%	27.7%
	Precision	63.8%	20%	24.6%
MMS	F1	63.4%	18.8%	54.8%
	Recall	51.9%	10.9%	64.7%
	Precision	81.4%	62.9%	47.6%

The achieved results are presented in Table 3. As the table shows, the best f-measure was achieved in those candidate expressions related to associations with a weak degree of confidence and the worst result was obtained in the medium degree of confidence. The reason for the weak result for the class with medium level of confidence is that there was small number of instances in the class. Moreover, better f-measure results of weak degree of confidence were determined there had been more trained instances. Moreover, the weak performance of the BOW method for two classes with stronger level of confidence can suggest that these classes overlapped with each other and that perhaps two classes of confidence would lead to better performance.

As table 3 shows, the proposed MMS method has better performance in comparison to the BOW method in terms of f-measure, precision and recall. In addition, as it is presented in the table, both methods have weak f-score, recall and precision in the category of middle level of confidences

5. Discussion and Conclusion

In this paper, we proposed a modality based SNP-phenotype association extraction method. The results demonstrate the superior performance of the proposed method. Additionally the results show how the neutral candidates are important category of candidates that can be utilized for implementing better relation extraction methods. Moreover the achieved results show the importance of confidence level of the association as a linguistic-based factor can be used beside to existing methods to obtain more useful information. Although the proposed method shows promising results employing other feature can improve the performance of the confidence estimation of the extracted association. The estimated level of confidence of the association can be used beside to other factor such as abstract and paper confidence to define the overall confidence and credibility of the extracted association.

Although all existing relation extraction corpora and methods utilize **crisp** relations, the authors believe that it is not an efficient model for natural language's relations and they could be replaced with a better mathematical model called **fuzzy relations** (FR). Crisp relations deal with the binary relation between two entities in a sentence while FRs includes sets of fuzzy relations.

Bibliography

- [1] Gabor, T. Marth et al., "A general approach to single-nucleotide polymorphism discovery," *Nature genetics*, vol. 23, no. 4, pp. 452-456, 1999.
- [2] Verspoor, K., Go Eun Heo, Keun Young Kang, and Min Song, "Establishing a baseline for literature mining human genetic variants and their relationships to disease cohorts," *BMC Medical Informatics and Decision Making*, vol. 16, no. 1, p. 37, 2016.
- [3] Ashique, M., Tsung-Jung Wu, Mazumder, R., & Vijay-Shanker, K., "DiMeX: A Text Mining System for Mutation-Disease Association Extraction," *PloS one*, vol. 11, no. 4, p. e0152725, 2016.
- [4] Nature Education. [Online] (2016, July). HYPERLINK
["http://www.nature.com/scitable/definition/phenotype-phenotypes-35"](http://www.nature.com/scitable/definition/phenotype-phenotypes-35)
<http://www.nature.com/scitable/definition/phenotype-phenotypes-35>
- [5] D Price, T., Qvarnstr, A., & E Irwin, D., "The role of phenotypic plasticity in driving genetic evolution," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 270, no. 1523, pp. 1433-1440, 2003.
- [6] Eugene E. Loos, Susan Anderson, Day, Jr. Dwight H., Paul C. Jordan, and J. Douglas Wingate, *Glossary of linguistic terms*. Camp Wisdom Road Dallas: SIL International , 2004.
- [7] Chapman, W., Bridewell, W., Hanbury, P. , Cooper, G.F., and Buchanan, B.G., Evaluation of Negation Phrases in Narrative Clinical Reports, 2002.
- [8] Joan L Bybee , Fleischman, S., *Modality in grammar and discourse*.: John Benjamins Publishing, 1995, vol. 32.
- [9] McDonald, R., "Extracting relations from unstructured text," *Rapport technique, Department of Computer and Information Science-University of Pennsylvania*, 2005.
- [10] Bokharaeian, B., Diaz, A., " Extraction of Drug-Drug Interaction from Literature through Detecting Linguistic-based Negation and Clause Dependency," *Journal of AI and Data Mining*, vol. 4, no. 2, pp. 203-212, 2016.
- [11] Yifan Peng, C.O. Tudor, M. Torii, C.H. Wu, & K. Vijay-Shanker, "iSimp: A sentence simplification system for biomedical text," in *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, Oct 2012, pp. 1-6.
- [12] Bokharaeian, B., Diaz, A., Neves, M., & Francisco, V., "Exploring Negation Annotations in the DrugDDI Corpus," in *Proceedings of the Fourth Workshop on Building and Evaluating Resources for*

Health and Biomedical Text Processing, 2014, pp. 84-91.

- [13] Lee, k., et al., "BRONCO: Biomedical entity Relation ONcology CORpus for extracting gene-variant-disease-drug relations," *Database*, vol. 2016, p. baw043, 2016.
- [14] Verspoor, k., et al., "Annotating the biomedical literature for the human variome," *Database*, vol. 2013, p. bat019, 2013.
- [15] Horn, F., Lau, AL., & Cohen, FE, "Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors," *Bioinformatics*, vol. 20, no. 4, Mar 2004.
- [16] Ravikumar, K., Liu, H., D Cohn, J., E Wall, M., & Verspoor, K., "Literature mining of protein-residue associations with graph rules learned through distant supervision," *Journal of Biomedical Semantics*, vol. 3, October 2012.
- [17] Faisal, Md., Chowdhury, M., Lavelli, A., & Fondazione Bruno Kessler, "Exploiting the Scope of Negations and Heterogeneous Features for Relation Extraction: A Case Study for Drug-Drug Interaction Extraction," in *HLT-NAACL13*, 2013, pp. 765-771.
- [18] Pyysalo, S., Airola, A., Heimonen, J., Bjorne, J., & F. and Salakoski, T. Ginter, "Comparative analysis of five protein-protein interaction corpora," *BMC bioinformatics*, vol. 9, no. Suppl 3, p. S6, 2008.
- [19] Kim, J-D., and Ohta, T., and Tateisi, Y. and Tsujii, J., "GENIA corpus—a semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, pp. i180–i182, 2003.
- [20] Chek Kim, L, and Miin-Hwa Lim, J., "Hedging in Academic Writing - A Pedagogically-Motivated Qualitative Study ," *Procedia - Social and Behavioral Sciences* , vol. 197, pp. 600-607, 2015, 7th World Conference on Educational Sciences. [Online]. HYPERLINK
["http://www.sciencedirect.com/science/article/pii/S1877042815042019"](http://www.sciencedirect.com/science/article/pii/S1877042815042019)
<http://www.sciencedirect.com/science/article/pii/S1877042815042019>
- [21] Thorsten, J., "Making large scale SVM learning practical," Universitat Dortmund, Tech. rep. 1999.
- [22] Bokharaeian, B., Diaz, A., & Chitsaz, H., "The SNPPhenA Corpus: An annotated research abstract corpus for extracting ranked association of single-nucleotide polymorphisms and phenotypes," in *second conference of signal processing and intelligent systems*, Tehran, accepted, 2016.
- [23] Wooding, S., et al., "Natural selection and molecular evolution in PTC, a bitter-taste receptor gene," *The American Journal of Human Genetics*, vol. 74, no. 4, pp. 637-646, 2004.

Part III

Appendices

Appendix 1: SNPPhenA Corpus Guidelines

Annotation Guideline

(Version 1.0)

Abstract

This paper is a guideline that instructs annotator how to think and don in the SNPPhenA corpus. After introducing task description and the background, it explains how to annotate the corpus step by step.

Contents

1. Introduction	2
2. Structure	2
2.1. Markup Conventions	2
2.2. An Example	3
3 Tags.....	4
3.1 Entities.....	5
3.2 Relationship	6
3.3 Features.....	7
3.3.1 Modality	8
3.3.2 Negation.....	8
4. Annotation Process	9
5. Frequently Asked Questions	10
6. Acknowledgement	11
7. References	12

1. INTRODUCTION

This document contains the description of the design principles underlying the SNPPhenA corpus, and detailed information about the encodings, markup conventions and the linguistic annotation with which the corpus was enriched.

The SNPPhenA corpus has been produced during a project aiming at a software package which could automatically determine existence of the association between various polymorphism and relevant phenotypes or traits which are presented in academic articles. Mainly, this corpus can be used for the task of extracting biomedical relations and the degree of the associations from scientific texts with accuracy and high confidence. In general, the task of biomedical relation extraction focuses on the biomedical entities such as proteins, drugs, and genes in the biology related article texts and try to find binary or complex relations between them.

2. STRUCTURE

SNPPhenA corpus consists of 360 documents in xml format and encoded in utf-8; each of which is drawn from the abstract section of the papers published in the journals and conferences related to the area of the life science. These documents have been annotated according to the conventions presented in this reference.

2.1. Markup Conventions

Building blocks and valid tags of the documents can be found on the document type definition file named SNPPhenA.dtd, which is located on the corpus folder.

The SNPPhenA is delivered in UTF-8 encoding. Almost all characters in the corpus are represented directly by the appropriate Unicode character. Some exceptions are as follows:

- the ampersand (&) which is represented by the special string `&`;
- the double quotation mark, which is represented by the special string `"`;
- the arithmetic less-than sign, which always appears as `<`;
- the arithmetic less-than-or-equal sign, which always appear as `≤`;
- the arithmetic greater-than sign, which always appear as `>`;
- the arithmetic greater-than-or-equal sign, which always appear as `≥`;

In the SNPPhenA corpus, name of tags have been written in lowercase; while attribute name is in uppercase. Each document in the corpus has a unique id in the whole corpus, starting from “1000”. IDs of the critical sentences are a composition of the document id followed by a local id which start from 0 such as “1047_1” in the above example. IDs for the other blocks are defined to be unique only on the containing document and so, start from 0.

START and END attributes for each tag shows the index of the beginning and the end of the referring string in the original text.

2.2. An Example

Here is a complete example of an annotated document. The example begins with the start tag for an <abstract> element, which bears an abstract id attribute, the value of which is 1047, and a text attribute, representing the abstract text (Figure 1). The start tag is followed by two <sentence> elements, which provides the critical sentences in the original source text. Sentence elements in turn are followed by zero or more <snp>, <phenotype>, <modality_marker>, <negation_scope> and <pair> elements. Each of these tags are described in the next sections.

```

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE abs SYSTEM "SNPPPhentA.dtd">
<abstract TEXT="OBJECTIVE: There is compelling evidence that the plasma apolipoprotein E (APOE) concentration, in addition to the APOE ε2/ε3/ε4 genotype, influences
plasma lipoprotein levels, but the functional genetic variants influencing the plasma APOE concentration have not been identified. APPROACH AND RESULTS: Genome-wide association
studies in 2 cohorts of healthy, middle-aged subjects identified the APOE locus as the only genetic locus showing robust associations with the plasma APOE concentration. Fine-
mapping of the APOE locus confirmed that the rs7412 ε2-allele is the primary genetic variant responsible for the relationship with plasma APOE concentration. Further mapping of
the APOE locus uncovered that the rs769446 (-427T/C) in the APOE promoter is independently associated with the plasma APOE concentration. Expression studies in 199 human liver
samples demonstrated that the rs769446 C-allele is associated with increased APOE mRNA levels (P=0.015). Transient transfection studies and electrophoretic mobility shift assays i
human hepatoma HepG2 cells corroborated the role of rs769446 in transcriptional regulation of APOE. However, no relationships were found between rs769446 genotype and
plasma lipoprotein levels in 2 cohorts (n=1648 and n=1039) of healthy middle-aged carriers of the APOE ε3/ε3 genotype. CONCLUSIONS: rs769446 is a functional polymorphism
involved in the regulation of the plasma APOE concentration." ABSTRACTID="1262">
- <sentence END="1315" START="1130" ID="1262_0">
  <snp TEXT="rs769446" END="1183" START="1175" ID="0"/>
  <phenotype END="1315" START="1197" ID="0" text="plasma lipoprotein levels"/>
  <modality_marker END="1166" START="1161" text="found"/>
  <negation_scope END="1315" START="1139" text="no relationships were found between rs769446 genotype and plasma lipoprotein levels in 2 cohorts (n=1648 and n=1039) of
healthy middle-aged carriers of the APOE ε3/ε3 genotype.">
    <negation_cue END="1141" START="1139" text="no"/>
  </negation_scope>
  <pair CONFIDENCE="-" ASSOCIATION="negative" SNPID="0" PHENOTYPEID="0" PAIRID="0"/>
</sentence>
- <sentence END="1427" START="1316" ID="1262_1">
  <snp TEXT="rs769446" END="1337" START="1329" ID="1"/>
  <phenotype END="1427" START="1401" ID="1" text="plasma APOE concentration"/>
  <modality_marker END="1327" START="1316" text="CONCLUSIONS"/>
  <pair CONFIDENCE="moderate" ASSOCIATION="positive" SNPID="1" PHENOTYPEID="1" PAIRID="1"/>
</sentence>
</abstract>

```

Figure 1: An example of annotated document in XML format

3 TAGS

A basic element in each document is <sentence>. Any sentences in the original text, which contain at least one SNP and phenotype entity, are considered as a critical sentence and is annotated with the appropriate tags. These tags include entity tags, relationship tag and features tags. At the following each of them are explained.

3.1 Entities

Two main classes of the entities considered in this corpus consist of SNPs and phenotypes. From the scientific view, SNPs are a variation in a single nucleotide that occurs at a specific position in the genome. For the annotation task, all of known SNP names as well as any mention in the text, which refers to a famous gene symbol, are selected as the <snp> entity. These names mostly come from the open-access databases including: SNP500Cancer [1], SNPedia [2], and pharmGKB [3].

In the following example, “rs429358” and “rs7412” are the name of two SNPs, which are clearly expressed in the text. For the simplicity, tags for SNP and phenotypes have been embedded in the original text.

Example 1:

“Apolipoprotein E (APOE) functional haplotypes determined by <snp>rs429358</snp> and <snp>rs7412</snp> SNPs have been extensively studied and found to be one of the most consistent association in human <phenotype>longevity</phenotype> studies.”

Different type of human characteristics should be tagged as <phenotype> which includes wide range of unusual circumstance from trait to disease. Indeed, a phenotype is the appearance of an organism in terms of the traits such as its morphology, development, physiological properties, behavior, and products of that behavior [4]. Two more complete related databases are chosen for this task: a list of Comparative Toxicogenomics Database (CTD) for disease names [5], and the phenotype ontology prepared on the blast project [6]. The collected list of phenotypes includes 65,530 phenotype names with more than twelve thousand disease names and their synonyms. In the example 1, “longevity” is a phenotype. The following table gives an example for the designed entities.

Entity type	Description	Example
phenotype	Name dedicated to each abnormality or features	Coronary heart disease
snp	rs plus number(rsID) and other corresponding historical numbers dedicated to each polymorphism which make probability of association with phenotype	rs499818, A1450G

3.2 Relationship

The main tag under the <sentence> is the <pair> tag. This tag represents the biomedical relation between a pair of SNP and phenotype, which have been annotated with the appropriate tags in that sentence. Attributes defined for the <pair> tag includes an ID for uniqueness, referred phenotype's ID and SNP's ID, association and the strength of the association (degree of confidence) between the entities.

An association indicates the existence or lack of the existence of a correlation between the appointed SNP with the relevant traits. Each pair of SNP-phenotype falls into one of the following categories: 1) positive, 2) negative, and 3) neutral. If the critical sentence expresses an association between SNP and phenotype, in terms of a cause-effect relation between SNP and phenotype, with some probability greater than zero, this pair of entities has a "positive" association. In the other hand, a "negative" association occurs when the SNP-phenotype pair evidently lacks of any association. Additionally, those pairs that their association or lack of association was not remarked in the sentence get "neutral" value for the ASSOCIATION attribute.

Another attribute for the tag <pair> is CONFIDENCE, which is the greatest strength point of the SNPPhenA corpus. This attribute shows the degree of the association described in the ASSOCIATION attribute. When the association is positive, three levels of confidence are defined: "low", "moderate", and "strong". In general, different authors write the same fact about the entities in different linguistic forms and with different confidence. To annotate the confidence level, annotators should note the real value of the degree of the association, if presented in terms of the p-value in the sentence or the paragraph. About 20% of sentences with positive pairs, have this remark. If this matter is not presented in the sentence, tone of the writer and key words like modality markers and negation words should be considered and based on them annotators should decide about the confidence level. These cues are referred as features and are presented at the next section.

If the association between entities is "neutral", then the CONFIDENCE value is "zero". Because, the degree of the association is zero. In "negative" cases, there is no association and so, no strength of association. For simplicity, CONFIDENCE value is presented with "-" in the corpus.

Example 2:

<snp>Apolipoprotein E (ApoE) genotype</snp> has been associated with <phenotype>systemic inflammation</phenotype> and athero-thrombosis however the association with <phenotype>abdominal aortic aneurysm</phenotype> (AAA) has not been previously examined.”

In the above example, it is understandable that there is an association between “ApoE” and “systemic inflammation”; so this pair has positive association. In contrast, the pair “ApoE” and “abdominal aortic aneurysm” belongs to the neutral class. Since, the author says this association has not been examined, and so, this pair cannot be labeled as positive or negative.

Example 3:

“The genetic factors studied were not associated with cognitive status in PD patients. Only age and Hcy plasma levels were found to be independent risk factors predisposing individuals to PD dementia. However, <snp>COMT: rs4680: A>G </snp> and rs4633: C>T polymorphisms were found to significantly affect <phenotype>PD</phenotype> risk, and the <snp>MTHFR 677C>T</snp> polymorphism helped determine <phenotype>plasma Hcy concentrations</phenotype>.”

In the example 3, which is drawn from the conclusion section, there are two SNPs (“COMT: rs4680: A>G” and “MTHFR 677C>T”) and also two phenotypes (“Parkinson's disease (PD)” and “plasma Hcy concentrations”). Result of the analysis has showed that the first SNP effected the PD; while the second one effected the next phenotype. As a result, these two pairs of SNP-phenotype have positive association. In contrast, opposite combination of SNP-phenotype, e.g. “COMT”- “plasma Hcy concentrations” and “MTHFR”-“PD”, have negative association. Meanwhile, two positive pairs should be tagged with the association degree. About the first one, founding showed strong association according to the phrase “were found to significantly affect”. However, second positive pair has weak association, because of the phrase “helped determine”.

3.3 Features

Each critical sentences should be enriched with some informative tags including <modality_marker> and <negation_scope> and <negation_cue>. These kinds of the annotations seem to be effective while

determining the degree of the association between the entities and the class of pairs. As the result, this information may be used in design of the appropriate machine learning algorithms for extracting.

3.3.1 Modality

Modality markers are those words that their job is qualifying the opinion and attitude of the speaker or writer. Writer of an academic paper can make judgments about the truth of a proposition or state a fuzzy proposition, which is true in partial or in some occasions, by utilizing modality marker words such as “may”, “could”, “possibly”, “almost”, “indicate” and “found”.

For example consider the following sentence. In this sentence, the word that indicate the existence or lack of existence of an association between “rs769446” and “plasma lipoprotein levels” is “found”.

Example 4:

“No relationships were <modality_marker>found</modality_marker> between <snp>rs769446</snp> genotype and <phenotype>plasma lipoprotein levels</phenotype> in 2 cohorts (n=1648 and n=1039) of healthy middle-aged carriers of the APOE ε3/ε3 genotype.”

As another example, following sentence represent an association between “APOA5-1131C” and “MI”, by the phrase “strongly affects”.

Example 5:

“The <snp>APOA5-1131C</snp> allele, associated with higher fasting triglyceride levels, <modality_marker>strongly</modality_marker> affects the risk for early-onset <phenotype>MI</phenotype>, even after adjusting for triglycerides.

Each critical sentence which has modal words should be annotated by <modality_marker> tags. The employed modality markers have been obtained from the list which is provided in [7]. This list is an extension of the list presented in [8] for biomedical domains.

3.3.2 Negation

Critical sentences containing any kind of negation are tagged with <negation_cue> and <negation_scope>. Negation is understood as the implication of the non-existence of something.

However, the presence of a negative word does not imply that the pairs of the sentence should be annotated as “negative” cases. The annotators must pay attention to the sentences including negative words.

The scope of a <negation_cue> tag starts right with the keyword, and ends with end of the key word. List of these key words are available to the annotators. Negation cues can occur in different morphological types, such as verbs like (lack), adverb (not), adjective (absent), determiner (no), noun (absence), conjunction (neither), and preposition (without) [9].

The scope of a <negation_scope> tag can extend to the whole sentence containing <negation_cue> tag, or to the certain phrases. Different negation cues in different structures, such as active and passive sentences, have different scope of negation. The instruction for annotating negation scopes is adopted from the rules given in the work of Morante [9].

4. ANNOTATION PROCESS

One simple method for annotating a document is to read the document from start to end and mark the annotation in order they appeared. This does not result the most accurate corpus. In order to gain consistent annotations, it is needed to have a methodology, according which all annotators think the same and do the same. Therefore, all annotators were asked to perform these steps in order:

- 1- *Read the whole document.* Reading the whole document without thinking about entities or relationship, only for getting understanding is necessary.
- 2- *Mark the entities.* If the tags of entities are incorrect, or some entities have no tag, annotators should edit the tags.
- 3- *Mark features.* If tags of features are incorrect, or some features have no tag, annotators should edit the tags.
- 4- *Find critical sentences.* If a sentence or some near sentences, which speak about same entities, contain both SNP and phenotype, mark this sentences as a critical sentence.
- 5- *Find the relations.* Considering entities and features in a critical sentence, focus on relations between each pair of entities to find the existence of associations and confidence level of the author. Annotators should follow the instruction while annotating each tags described in the above.

- 6- *Look again.* Reviewers are asked to be certain that nothing is missed. Specially, annotators should count number of pairs they tagged. If there is x SNP and y phenotype in a critical sentence, they should annotate $x \times y$ pairs.
- 7- *Record any question, or ambiguous situation.* If reviewers have any question that need to be clarified, they should record them.

After the annotators completed their task, a software program was hired to find the inconsistencies and missing items. Furthermore, inter-agreement analysis has been performed to measure the quality of the annotations.

5. FREQUENTLY ASKED QUESTIONS

In this section, the most frequently asked question of the annotators are presented.

A. How to decide about the confidence level, if the *p-value* is presented in the sentences for indicating the strength of the association between entities?

If the *p-value* is lower than or equal to the 0.001 ($p \leq 0.001$), CONFIDENCE is “strong”. If *p-value* is greater than 0.001 and less than 0.01 ($0.01 < p < 0.001$), CONFIDENCE is “moderate”. In the case of *p-value* greater than or equal to the 0.01 ($p \geq 0.01$), CONFIDENCE is “low”. As an example, consider the following sentences. The degree of the association between narcolepsy and rs5770917 is “low”. Because the *p-value* is 0.02.

Example 6:

“<snp>rs5770917</snp>, a SNP located between CPT1B and CHKB, was associated with <phenotype>narcolepsy</phenotype> in Japanese (<snp>rs5770917</snp>[C], odds ratio (OR) = 1.79, combined P = 4.4×10^{-7}) and other ancestry groups (OR = 1.40, P = 0.02).”

B. What is the scope of the negation cues?

Some of negation cues and their function and scope is as follows:

- “No” is the most frequent negation cue in clinical reports. No occupies the first position of a nominal sentence. The full noun phrase in which no is a determiner is under the scope of the

negation. If no modifies a noun, it scopes over the noun phrase and if it modifies an adjective, it scopes over the adjectival phrase.

“<negation_scope> <negation_cue>No</negation_cue> association between COMT and smoking behavior was observed</negation_scope>”

- “Not” is always a negation cue. If it modifies a verb, it scopes over the verb phrase in active sentences and over the clause, in passive sentences. If it modifies other phrases, it scopes over the phrase.

“<negation_scope> <negation_cue>not</negation_cue> confirm the association of CPT1B/CHKB (rs5770917) in the Chinese population”

- In active sentences, negation cues, such as “cannot”, “could not”, “didn’t”, and “exclude” cope over the object of the main verb, and over the subject in passive sentences.

“COMT <negation_scope><negation_cue>didn't</negation_cue> affect CPP </negation_scope>”

- “Neither ... nor” scopes over the full clause, if it coordinates copulative clauses or clauses in passive form.

“<negation_scope> <negation_cue>neither</negation_cue> rs11196218 <negation_cue>nor</negation_cue> rs290487 showed a significant association </negation_scope>”

- The determiner “neither” acts always as a negation cue, which scopes over the full clause.
- The noun “absence” is always a negation cue that its scopes is the prepositional phrase headed by of that is required by absence.
- The adjective “absent” scopes over the noun phrase it modified, or over the copulative clause it participated in it.

C. How the SNPPhenA corpus can be extended?

This corpus can be extended qualitatively by the complementary annotation tags, which may be useful in determining the degree of association, and quantitatively by adding supplementary articles.

6. ACKNOWLEDGEMENT

The author would like to acknowledge the Dr. Mariana Neves (University of Potsdam) and Dr. MT Pilehvar (Cambridge university) for their useful comments.

7. REFERENCES

- [1] Bernice R Packer et al., "SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D617--D621, 2006.
- [2] Michael Cariaso and Greg Lennon, "SNPedia: a wiki supporting personal genome annotation, interpretation and analysis," *Nucleic acids research*, vol. 40, no. D1, pp. D1308--D1312, 2012.
- [3] Micheal Hewett, Diane E Oliver, Daniel L Rubin, Katrina L Easton, and et. al., "PharmGKB: the pharmacogenetics knowledge base," *Nucleic acids research*, vol. 30, no. 1, pp. 163-165, 2002.
- [4] Elizabeth Martin and Robert Hine, *A Dictionary of Biology*, 6 ed.: Oxford University Press, 2014.
- [5] Allan Peter Davis et al., "Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical--gene--disease networks," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D786--D792, 2009.
- [6] (2015) Basic Local Alignment Search Tool (BLAST). [Online]. HYPERLINK
"https://blast.ncbi.nlm.nih.gov/Blast.cgi" <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
- [7] Paul Thompson, Giulia Venturi, John McNaught, Simonetta Montemagni, and Sophia Ananiadou, "Categorising modality in biomedical texts," in *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, 2008, pp. 27-34.
- [8] Ken Hyland, "Talking to the Academy: Forms of Hedging in Science Research Articles.," *Written Communication*, vol. 3, no. 2, 1996.
- [9] Roser Morante, "Descriptive analysis of negation cues in biomedical texts," in *LREC*, 2010.
- [10] Michael Seringhaus and Mark Gerstein, "Manually structured digital abstracts: A scaffold for automatic text mining," *FEBS letters*, vol. 582, no. 8, p. 1170, 2008.
- [11] Wei Yu, Marta Gwinn, Melinda Clyne, Ajay Yesupriya, and Muin J Khoury, "A navigator for human genome epidemiology," *Nature genetics*, vol. 40, no. 2, pp. 124-125, 2008.

Appendix 2: Snapshots of the SNPPhenA Corpus in XML and Brat Formats


```

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE abs SYSTEM "SNPPhenA.dtd">
- <abstract TEXT="BACKGROUND: Apolipoprotein E (APOE) polymorphism is associated with lipid levels. Some studies have reported that blood lipid response to diet or obesity varies depending on APOE genotypes. The aim of this study was to assess the effect of APOE genotypes, the intake of saturated fatty acids(SFA), and obesity on serum lipid levels in Lithuanian adult population. RESULTS: A cross-sectional health survey was carried out in five municipalities of Lithuania. The random sample was obtained from lists of 25-64 year-old inhabitants registered at primary health care centres. The data from 996 subjects (416 men and 580 women) were analysed in this study. Two single-nucleotide polymorphisms (rs429358 and rs7412) were assessed using a real-time polymerase chain reaction. 24-hour recall and food frequency questionnaire were used for evaluation of dietary habits. Serum lipids were determined using enzymatic methods. Men and women with the APOE2 genotype had the lowest level of total cholesterol (TC) (p=0.002 for men, and p=0.02 for women) and low-density lipoprotein cholesterol (LDL-C) (p<0.001). Multivariate linear regression analysis showed that age, genotype APOE2, SFA intake, and body mass index (BMI) were significant determinants of TC and LDL-C level (with p values ranging from 0.043 to 0.001). Our data did not reveal any statistically significant interactions between APOE genotype and SFA intake or between APOE genotype and BMI regarding TC and LDL-C level (all p>0.05). However, the predictive power of the regression model for LDL-C improved when gene-BMI interaction and gene-BMI interaction plus gene-nutrient interaction were added (p=0.04 and p=0.032 for R(2) change, respectively). CONCLUSIONS: APOE genotypes, SFA intake, and obesity were found to be associated with blood lipid levels in Lithuanian adult population. Analysis of gene-diet and gene-obesity interactions did not confirm that the effects of diet and obesity on TC and LDL-C level significantly depended on APOE genotype." ABSTRACTID="1112">
- <sentence END="364" START="189" ID="1112_0">
  <snp TEXT="APOE genotypes" END="255" START="241" ID="0"/>
  <phenotype END="364" START="303" ID="0" text="obesity"/>
  <pair CONFIDENCE="zero" ASSOCIATION="neutral" SNPID="0" PHENOTYPEID="0" PAIRID="0"/>
</sentence>
- <sentence END="1850" START="1713" ID="1112_1">
  <snp TEXT="APOE genotypes" END="1741" START="1727" ID="1"/>
  <phenotype END="1850" START="1759" ID="1" text="obesity"/>
  <modality_marker END="1724" START="1713" text="CONCLUSIONS"/>
  <modality_marker END="1777" START="1772" text="found"/>
  <pair CONFIDENCE="weak" ASSOCIATION="positive" SNPID="1" PHENOTYPEID="1" PAIRID="1"/>
</sentence>
- <sentence END="2018" START="1851" ID="1112_2">
  <snp TEXT="APOE genotype" END="2017" START="2004" ID="2"/>
  <phenotype END="2018" START="1948" ID="2" text="obesity"/>
  <modality_marker END="1920" START="1911" text="confirm t"/>
- <negation_scope END="2017" START="1907" text="not confirm that the effects of diet and obesity on TC and LDL-C level significantly depended on APOE genotype">
  <negation_cue END="1910" START="1907" text="not"/>
</negation_scope>
  <pair CONFIDENCE="-" ASSOCIATION="negative" SNPID="2" PHENOTYPEID="2" PAIRID="2"/>
</sentence>

```

Figure 1: A snapshot of SNPPhenA corpus in XML format

```

T1    SNP 54 69   TOMM40 rs157590
T2    SNP 74 87   APOE rs429358
T3    Phenotype 104 107 PPA
T4    Phenotype 120 125 bvFTD
T5    Modality_Marker 92 100 observed
T6    Negation_Scope 113 142 not in bvFTD and in controls.
T7    Negation_Cue 113 116   not
R1    weak_confidence_association Arg1:T1 Arg2:T3
R2    negative_association Arg1:T1 Arg2:T4
R3    weak_confidence_association Arg1:T2 Arg2:T3
R4    negative_association Arg1:T2 Arg2:T4

```

Figure 2: A snapshot of SNPPhenA corpus in brat format

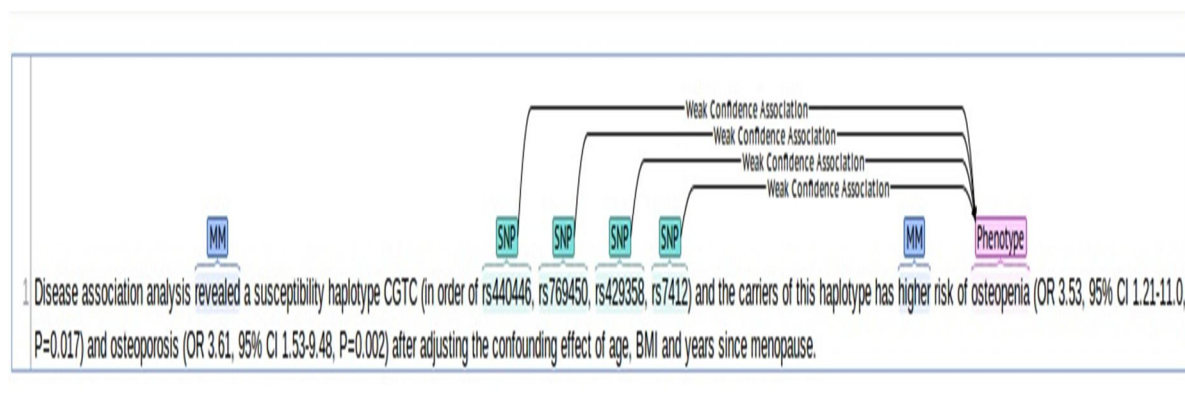


Figure 3: A snapshot of a sentence with four weak confidence associations in the SNPPhenA corpus drawn by Brat

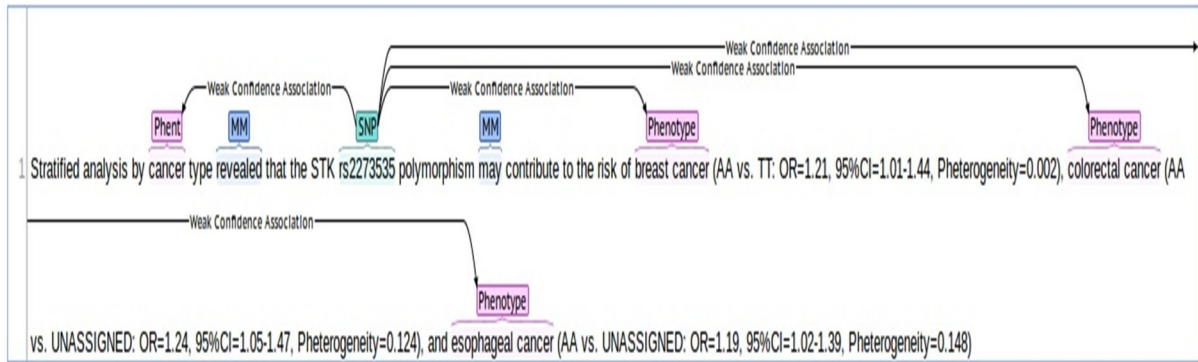


Figure 4: A snapshot of a sentence with four weak confidence association in the SNPPhenA corpus drawn by Brat

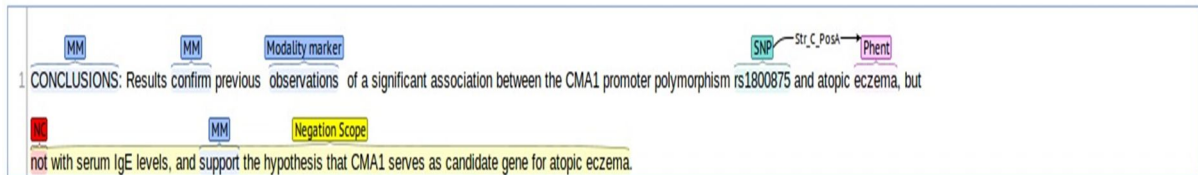


Figure 5: A snapshot of a sentence with a negation cue and scope in the SNPPhenA corpus drawn by Brat

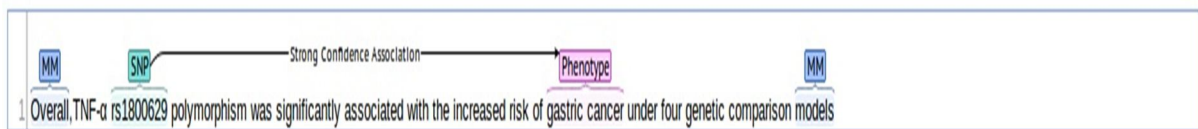


Figure 6: A snapshot of a sentence with a strong confidence association in the SNPPhenA corpus drawn by Brat

Appendix 3: Snapshots of the SNPPhenA Website

SNPPhenA: A corpus for extracting ranked associations of SNP and phenotypes from literature

Submitted by behrouz on Mon, 10/17/2016 - 14:07

The SNPPhenA corpus

The SNPPhenA corpus consists of medical and biological texts annotated for snp-phenotype associations, negation, modality markers and degree of confidence of associations. This was done to allow a comparison between the development of systems for association extraction as well as the degree of confidence and strength of associations. The corpus is publicly available for research purposes.

The annotation guidelines: [pdf](#)

Annotation principles are also discussed in the following paper:

Corpus download

Information provided in the <http://www.gopubmed.org/> search engine was used to collect genome-wide association abstracts. GoPubMed is a webserver that allows users to explore PubMed search results with Gene Ontology. Here is DTD for the xml files containing the annotations: [DTD](#)

Abstracts of the SNPPhenA corpus: xml v1.0

The full corpus in XML and BRAT formats is available in one file: [zip](#)

An online association extraction system that utilizes the SNPPhenA corpus is available [here](#).

Inter-agreement analysis

In order to evaluate the quality of the corpus and the reliability of the annotations, inter-annotator agreement score was measured for the task of classifying candidate sentences into positive, negative and neutral classes, and also for task of determining the confidence level of the association. Two annotators independently have tagged the corpus. In the case of disagreement between two tags, a third annotator was asked to decide about the correct one. For the task of classifying types of association, inter-annotator agreement was 86%, which means that in 86% of cases, the two annotators have agreed. Additionally, we computed Cohen's Kappa coefficient, for two annotators, which takes into account the amount of agreement that could be expected to occur through chance. For our two annotators and the type of association task, the Kappa value was 0.79. For the task of annotating confidence level of the association, the Kappa value was 0.80.

The results show that annotating confidence level of association is a more difficult task than simply classifying candidate sentences to positive, negative and neutral classes.

Corpus statistics

In the table below, some detailed statistics of the linguistic and nonlinguistic properties of the corpus, in terms of test and training parts, are presented.

Figure 1: A snapshot of the website of the SNPPhenA corpus dedicated corpus download and inter-annotator agreement analyses

Corpus statistics

In the table below, some detailed statistics of the linguistic and nonlinguistic properties of the corpus, in terms of test and training parts, are presented.

Item	Train	Test	Total
Abstracts	270	90	360
Key Sentences	362	121	483
SNP	691	244	935
Phenotypes	496	158	654
SNP-Phenotype association candidates	935	365	1300
Neutral Candidates	142	166	308
Positive Candidates	702	170	872
Negative Candidates	91	29	120
Strong degree of confidence candidates	213	20	233
Medium degree of confidence candidates	92	32	124
Weak degree of confidence candidates	390	125	515

Figure 2: A snapshot of the website of SNPPhenA corpus

Appendix 4: Kappa Calculation for Analyzing the Reliability of the SNPPhenA Corpus

Table 1: Kappa calculation for annotation of association

Annotator	B				
A	Association value	0	1	2	Total
	0	101	5	9	115
	1	11	738	80	829
	2	1	9	198	208
	Total	113	752	287	

$$K = (\text{Pr}(a) - \text{Pr}(e)) / (1 - \text{Pr}(e))$$

Where,

Pr (a) - Relative observed agreement,

Pr (e) - Hypothetical probability of chance agreement,

k - Cohen's kappa index value

$$\text{Pr}(a) = 0.900174$$

$$\text{Pr}(e) = 0.5245$$

$$K = 0.79$$

Table 2: Annotator agreement and Kappa calculation for annotation of Confidence level

Annotator	B					
A	Confidence level	0	1	2	3	Total
	0	198	7	0	2	207
	1	62	411	3	8	484
	2	6	15	82	4	107
	3	12	10	10	195	227
	Total	278	443	95	209	

$$K = (\text{Pr}(a) - \text{Pr}(e)) / (1 - \text{Pr}(e))$$

Where,

Pr (a) - Relative observed agreement,

Pr (e) - Hypothetical probability of chance agreement,

k - Cohen's kappa index value

$$\text{Pr}(a) = 0.86439$$

$$\text{Pr}(e) = 0.313686$$

$$K = 0.8024$$