UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÁTICA

Extracting Negative Biomedical Relations from Literature

Leonor Horta Torcato

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:

Professor Doutor Francisco José Moreira Couto

2021

Acknowledgements

Firstly, I would like to thank my teacher and supervisor Professor Doutor Francisco Couto for his valuable expertise, and for always being available to guide me through this thesis. I also would like to warmly thank Diana Sousa for taking the time to answer my questions and give me important feedback. To my family, thank you for supporting and being present in my educational journey.

Abstract

The prevalent source for obtaining scientific knowledge remains the scientific literature. Considering that the focus of biomedical research has shifted from individual entities to whole biological systems, understanding the relations between those entities has become paramount for generating knowledge. Relations between entities can either be positive, if there is evidence of an association, or negative, if there is no evidence of an association. To this date, most relation extraction systems focus on extracting positive relations, therefore few knowledge bases contain negative relations. Disregarding negative relations leads to the loss of valuable information that could be used to advance biomedical research.

This work presents the Negative Phenotype-Disease Relations (NPDR) dataset, which describes a subset of negative disease-phenotype relations from a gold-standard knowledge base made available by the Human Phenotype Ontology (HPO), and an automatic extraction system developed to automatically annotate the entities and extract the relations from the NPDR dataset. The NPDR dataset was constructed by analysing 177 medical documents and consists of 347 manually annotated at the document-level relations, from which 222 were inferred from the HPO gold-standard knowledge base, and 125 were new annotated relations. The main categories of the dataset are the characterization of the entities that participate in the negative relation; the characterization of the sentence that implies the negative relation; and the characterization of the location of the entities and sentences in the article.

The automatic extraction system was created to evaluate the impact of the NPDR dataset on the Named-Entity Recognition (NER), Named-Entity Linking (NEL) and Relation Extraction (RE) text mining tasks. The NER task showed an average of 20.77% more entities annotated when using disease and phenotype synonyms lexica generated from the NPDR dataset, when comparing the number of annotations produced by the OMIM and HPO lexica. The increase in annotated entities also resulted in 15.11% more relations extracted. The RE task performed poorly, with the highest accuracy being 8.84%.

Keywords: Negative Relations, Biomedical Ontologies, Text Mining, Biomedical Literature, Relation Extraction.

Resumo

Texto livre continua a ser, aos dias de hoje, o principal meio de produção e partilha de conhecimento. Mais concretamente, a literatura biomédica é a principal fonte de conhecimento clínico e biológico para investigadores e clínicos. Porém, à medida que a informação contida em texto livre, correspondente ao número de publicações de artigos científicos aumenta a um ritmo exponencial, torna-se difícil para os investigadores manterem-se a par dos desenvolvimentos dos variados domínios científicos. Para além disso, extrair informação textual relevante é uma tarefa laboriosa e morosa para seres humanos, uma vez que a maioria da informação se encontra retida em texto livre não estruturado. Embora esta tarefa possa resultar em erros quando realizada por computadores, só poderá ser alcançada por meio de processos automáticos. Nesse sentido, métodos de prospeção de texto são uma alternativa interessante para reduzir o tempo despendido por especialistas na obtenção de informação relevante, para além de também cobrirem um largo volume de dados provenientes da literatura biomédica.

Métodos de prospeção de texto incluem várias tarefas, tais como *Named-Entity Recognition* (NER), *Named-Entity Linking* (NEL) e Extração de Relações (ER). O NER identifica as entidades mencionadas no texto, o NEL mapeia as entidades reconhecidas a entradas numa base de dados, e o ER identifica relações entre as entidades reconhecidas. Visto que o foco da investigação biomédica mudou de entidades individuais, tais como genes, proteínas ou fármacos, para sistemas biológicos num todo, métodos de ER automáticos tornaram-se fundamentais para entender relações entre entidades, tais como interações proteína-proteína, interações fármaco-fármaco, ou relações gene-doença. Estas relações podem ser classificadas como negativas, caso haja evidência de não associação entre as entidades, ou positivas, caso haja evidência de associação entre as entidades.

ER pode ser efetuada através de múltiplas abordagens que diferem nos métodos que empregam. Essas abordagens podem ser divididas nos seguintes grupos: coocorrência, que é a abordagem mais simples, uma vez que apenas visa a identificação das entidades na mesma frase; baseada em regras, que são definidas manualmente ou automaticamente; e aprendizagem automática, que utiliza *corpora* biomédica anotada para aplicar supervisão distante. Métodos de supervisão distante podem ainda ser categorizados em *feature-based* e *kernel-based*. Aos dias de hoje, a maioria dos sistemas de ER não diferenciam entre relações positivas, negativas ou falsas, porém podem-se salientar algumas excepções, tais como os sistemas Excerbt e BeFree. O primeiro combina análises sintáticas e semânticas com abordagens de regras e aprendizagem automática, e foi adaptado de forma a detetar representações léxicas negadas de itens léxicos (tais como verbos, nomes ou adjetivos) para a anotação do Negatome, uma base de dados

de proteínas que não interagem entre si. O segundo sistema utiliza uma combinação de métodos *kernel-based*, nomeadamente o *Shallow Linguistic Kernel* e *Dependency Kernel*. Para a anotação do *corpus* GAD usando este sistema, também foi treinado um classificador para distinguir entre relações positivas, negativas e falsas entre genes e doenças.

Estima-se que 13.5% das frases de resumos da literatura biomédica possuem expressões negadas. Desconsiderar expressões que poderão, potencialmente, conter relações negativas pode levar à perda de informação valiosa. Porém, a maioria das bases de dados de extrações de relações biomédicas visam apenas recolher relações positivas entre entidades biomédicas. No entanto, exemplos negativos e positivos são igualmente importantes para treinar, afinar e avaliar sistemas de extração de relações. Contudo, uma vez que os exemplos negativos não se encontram tão documentados como os positivos, poucas bases de dados os contêm. Para além disso, a maioria das bases de dados de extração de relações biomédicas não diferencia entre relações falsas, em que duas relações não estão relacionadas, e negativas, em que existe afirmação de não associação entre duas entidades. Adicionalmente, alguns datasets de padrão prata (compostos por dados gerados de forma automática) também contêm relações negativas falsas que são desconhecidas ou não estão documentadas. Logo, a exploração dessas relações é um bom ponto de partida para expandir as bases de dados de relações biomédicas e populá-las com exemplos negativos corretos.

Este trabalho produziu um dataset de anotações de fenótipos e doenças humanas e as suas relações negativas, o dataset *Negative Phenotype-Disease Relations* (NPDR), e um módulo de anotação automática de entidades e relações. Para a realização da primeira etapa da criação do dataset NPDR, foi necessário realizar a recolha dos identificadores PubMed (PMIDs) associados à relações negativas descritas numa base de dados padrão-ouro, disponibilizada pela *Human Phenotype Ontology* (HPO). A partir desses PMIDs foi possível extrair artigos completos que foram subsequentemente analisados manualmente. Essa análise consistiu na descrição das entidades que participam na relação negativa, que compreende a análise dos fenótipos, doenças e os seus genes associados; a descrição das frases que sugerem a relação a negativa, que engloba a caracterização do *token* de negação usado na frase e a coocorrência das entidades; e a descrição da localização das entidades e frases no artigo. O dataset NPDR contem um total de 347 relações anotadas ao nível do documento, das quais 222 foram obtidas a partir da base de dados padrão-ouro da HPO, e 125 são novas relações.

De forma a avaliar o impacto do dataset NPDR na anotação e extração automática de entidades e as suas relações, a partir dos artigos reunidos para o desenvolvimento da criação do dataset, um *pipeline* que realiza NER, ER e extrai frases de negação foi implementado. NER reconhece fenótipos humanos e doenças, e ER extrai e classifica a relação entre as entidades. De modo a obter os artigos num formato que fosse legível por máquina, dois métodos foram empregues. O primeiro método consistiu em reunir os PMIDs a partir do dataset NPDR, para os converter nos seus identificadores PubMed Central (PMCIDs) correspondentes, de forma a extrair os artigos completos usando a API do PubMed. O segundo método consistiu na conversão dos artigos reunidos para a construção do dataset NPDR em formato PDF para formato de texto, utilizando a ferramenta de extração de texto PDFMiner. A etapa NER foi realizada usando a ferramenta *Minimal Name-Entity Recognizer* (MER) para extrair menções de fenótipos,

doenças e genes a partir dos artigos. Por fim, utilizando uma abordagem de supervisão distante, a base de dados padrão-ouro da HPO foi usada para obter as relações obtidas pela ocorrência de fenótipos nas frases que sugerem a relação negativa, e a ocorrência de doenças e genes relacionados presentes no artigo. As relações foram marcadas como *Conhecida* se a relação estivesse descrita na base de dados, ou *Desconhecida* caso contrário.

Para a anotação de fenótipos dois léxicos foram utilizados, um de termos oficiais da HPO, e outro de sinónimos obtidos a partir do dataset NPDR. Para a anotação de doenças e genes, o léxico principal foi obtido a partir da base de dados da *Online Mendelian Inheritance in Man* (OMIM), e os restantes léxicos foram construídos a partir de sinónimos e abreviaturas de doenças presentes no dataset NPDR. A adição dos léxicos provenientes do dataset NPDR permitiram anotar, em média, mais 20.77% de entidades, comparativamente à anotação de entidades com os léxicos da HPO e OMIM. Este maior número de entidades também se refletiu num aumento de 15.11% de relações anotadas. A tarefa de ER teve um desempenho fraco, sendo que a precisão de relações negativas detetadas foi de 8.84%.

Keywords: Relações Negativas, Mineração de Texto, Ontologias Biomédicas, Literatura Biomédica, Extração de Relações.

Contents

1 Introduction **1**

1.1 Motivation . . . 1

1.2 Objectives . . . 4

1.3 Methodology . . . 5

1.4 Contributions . . . 5

1.5 Document Structure . . . 6

2 Related Work **7**

2.1 Text Mining for Biomedical Literature . . . 7

2.1.1 Natural Language Processing . . . 7

2.1.2 Text Mining Tasks . . . 8

2.1.3 Text Mining Approaches for Relation Extraction . . . 9

2.1.4 Text Mining Tools for Relation Extraction . . . 10

2.1.5 Text Mining Tools for Negation Detection . . . 12

2.2 Textual Resources for Biomedical Natural Language Processing . . . 13

2.2.1 Ontologies . . . 13

2.2.2 Corpora . . . 14

2.2.3 Databases . . . 17

3 A Dataset of Negative Phenotype-Disease Relations **19**

3.1 Methods . . . 19

3.1.1 Disease Annotations . . . 20

3.1.2 Phenotype Annotations . . . 21

3.1.3 Negative Relation Sentence Annotations . . . 22

3.1.4 Entities and Sentences Location . . . 23

3.2 Evaluation . . . 25

3.2.1 Automatic Extraction Method . . . 25

3.2.1.1 Negation Extraction . . . 26

3.2.1.2 Named Entity Recognition and Named Entity Linking . . . 26

3.2.1.3 Relation Extraction . . . 27

3.3 Results and Discussion . . . 28

3.3.1 NPDR Dataset 28
3.3.2 Automatic Extraction 33

4 Conclusion **37**
4.1 Future Work 38

References **41**

List of Figures

1.1 Example Sentences for Positive and Negative Relations 2
1.2 HPO Ontology Excerpt 4

3.1 NPDR Pipeline 25

List of Tables

2.1 Evaluation of Biomedical RE Systems 11

2.2 Evaluation of Biomedical Tools for Negation Detection 13

2.3 Biomedical Corpora Relevant to RE Systems 16

3.1 NPDR Dataset Categories 20

3.2 Clinical Table Example 23

3.3 Negation Words and Their Location 23

3.4 Summary of Article Locations and Their Equivalent Location 24

3.5 Source, Method and Final Number of Documents for Each Corpus 26

3.6 NPDR Dataset Categories Statistics 29

3.7 NPDR Dataset Negation Tokens Statistics 31

3.8 NPDR Dataset Co-occurence Statistics 32

3.9 NPDR Dataset Locations Statistics 33

3.10 Final Number of Disease, Gene and Phenotype Annotations 34

3.11 Final Number of Extracted Relations 35

3.12 Evaluation of the Relation Extraction Classifier 35

Chapter 1

Introduction

1.1 Motivation

Free text is still, to this day, the prevailing mean of producing and sharing knowledge. More specifically, biomedical literature is the fundamental source of clinical and biological knowledge for researchers and clinicians. However, as the information concealed in free text, corresponding to the number of publications of scientific articles, increases at an exponential rate [Jensen et al., 2006], it becomes extremely demanding for researchers to keep up-to-date with the developments of numerous fields. Furthermore, retrieving relevant textual information is a gruelling and time-consuming enterprise for humans since most of the information is trapped in free, unstructured text. Although this task can be prone to errors when conducted by machines, it can only be attainable using an automatic process. Therefore, text mining tools offer an interesting solution for reducing the time spent by experts on obtaining purposeful information, while covering a larger amount of data from the biomedical literature.

Text mining tools comprise several tasks, such as Named-Entity Recognition (NER), Named-Entity Linking (NEL) and Relation Extraction (RE). The aim of NER is to identify entities mentioned in text, of NEL is to map these entities to entries in a knowledge base, and of RE is to identify the relations between the entities [Sousa et al., 2019]. Considering that the focus of biomedical research shifted from individual genes, proteins or chemicals, to whole biological systems, automatic relation extraction methods became paramount for understanding the relations between these entities, such as protein-protein interactions [Federico and Monti, 2021], drug-drug interactions [Hou et al., 2021], and disease-gene relations [Zhou et al., 2021].

True relations can be either classified as negative, if there is evidence of no association between the entities, or positive, if there is evidence of an association between the entities [Sousa et al., 2020]. For example, the sentence *Wilson disease (WD) is an autosomal-recessive disorder caused by mutations in the copper (Cu)-transporter ATP7B* is an illustration of a positive relation between the entities **Wilson**

Disease and **ATP7B**. On the other hand, *In conclusion, we found no association between the Gln-Arg 191 polymorphism of the human paraoxonase gene and coronary artery disease in Finns* represents a negative relation between the entities **human paraoxonase gene** and **coronary artery disease** (Figure 1.1).

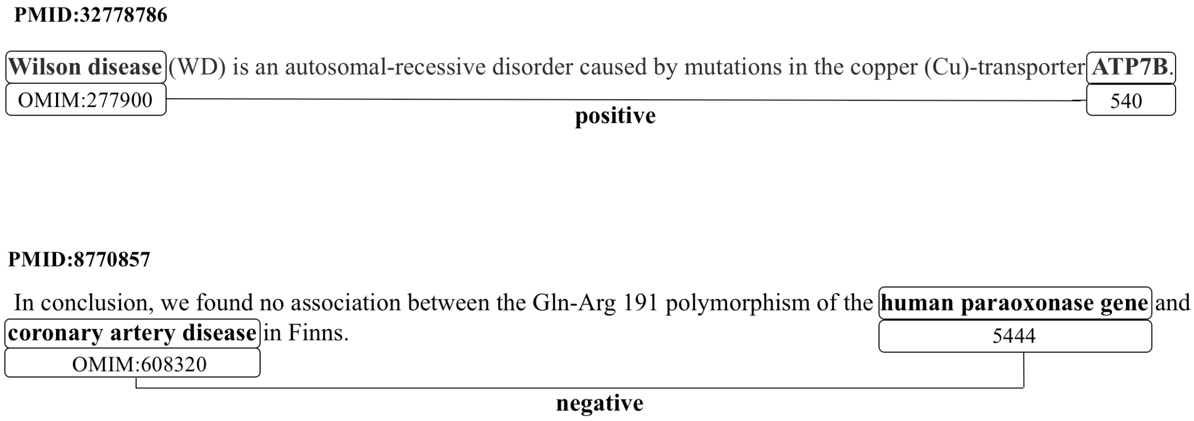
PMID:32778786
Wilson disease (WD) is an autosomal-recessive disorder caused by mutations in the copper (Cu)-transporter ATP7B.
OMIM:277900
540
positive
PMID:8770857
In conclusion, we found no association between the Gln-Arg 191 polymorphism of the human paraoxonase gene and coronary artery disease in Finns.
5444
OMIM:608320
negative

Figure 1.1: Example sentences for positive and negative relations: positive (PMID:32778786), and negative (PMID:8770857). Also present, the identified entities (NER) for each sentence, and their identifiers (NEL) in the National Center for Biotechnology Information (NCBI) (for genes) and Online Mendelian Inheritance in Man (OMIM) (for diseases).

RE can be achieved through multiple approaches that differ in their techniques. These approaches can be divided into the following groups: co-occurrence, which is the simplest approach since it aims at identifying co-occurring entities in a sentence; rule-based, which uses manually or automatically defined rules or patterns; and Machine Learning (ML), which uses annotated biomedical corpora to apply supervised learning. Supervised methods can be further categorised in feature-based and kernel-based methods [Zhou et al., 2014]. To this date, most biomedical RE systems do not differentiate between positive, negative and false relations, but some exceptions are worth mentioning, such as the Excerbt and BeFree systems. The former combines syntactic and semantic analyses with rule-based and ML approaches, and was adapted to detect negated lexical representation of argument-taking lexical items (e.g. verbs, nouns or adjectives) for the annotation of the Negatome, a database of non-interacting proteins [Blohm et al., 2013]. The later uses a combination of kernel-based methods, namely the Shallow Linguistic Kernel [Giuliano et al., 2006a], and the Dependency Kernel. For the annotation of the GAD corpus using this system, a classifier was also trained in order to distinguish between positive, negative and false relations between genes and diseases [Bravo et al., 2015].

It is estimated that 13.5% of the sentences in abstracts from biomedical literature contain negated expressions [Vincze et al., 2008]. Disregarding expressions that may possibly hold negative relations can lead to the loss of valuable information. Still, most biomedical relation extraction databases seek only to collect positive relations among biological entities, losing relevant information by not annotating negative ones. Negative and positive examples are equally important for training, tuning, and evaluating relation extraction systems. The more examples that are fed into the system, the better it will be at extracting new relations from the literature. However, since negative examples are not as documented as positive ones, few knowledge bases contain them. Additionally, most biomedical relation extraction datasets do not differentiate between false relations, where two entities are not related, and negative ones, where there is an affirmation of no association between two entities. Besides, some silver standard datasets, which are composed by automatically annotated data, also have false negative relations that are unknown or undocumented [Sousa et al., 2020]. Therefore, the exploration of these relations is a good point of departure for expanding biomedical relations knowledge bases and populate them with accurate negative examples. The annotation of negative relations is equally valuable for clinicians and researchers, since they restrict the search space by providing relations that were already refuted and enhance the quality of available information to the scientific community.

Even though automatic text mining tools make information extraction from biomedical literature more efficient, one of the major challenges it faces is the transformation of natural language into a structured representation, which can be easily processed by computer programs [Lamurias and Couto, 2019]. Furthermore, the language of biomedicine is populated with various synonyms, abbreviations, and acronyms that can point out to the same concept, making information extraction even more challenging for accurate computer-aided processing. Ontologies are formal descriptions of a set of concepts within a specific domain and the relations between them, organized in a hierarchical manner. These descriptions are determined by entities, classes, attributes, and relations. Thus, ontologies produce a framework for integrating large amounts of data from multiple sources, making available relevant information in a machine-readable way. They also provide a unique identifier for each concept or entity in a domain, in addition to connect concepts with related meanings, which makes them an essential feature of biomedical research fields. Moreover, their organisation in a machine-readable format makes their integration in relation extraction models more straightforward [Robinson and Bauer, 2011]. Therefore, ontologies help detect and classify relations between entities since they make available underlying characteristics between entities with is-a relations. For example, *astigmatism* (HP:0000483), a phenotypic abnormality of the curvatures on the anterior and/or posterior surface of the cornea, **is-a** *abnormality of the curvature of the cornea* (HP:0100691), and **is-a** *abnormal cornea morphology* (HP:0000481), which in turn **is-a** *abnormal anterior eye segment morphology* (HP:0004328) (Figure 1.2). Combining the knowledge of different domain-specific ontologies, such as the Gene Ontology (GO) [Ashburner et al., 2000], the Human Phenotype Ontology (HPO) [Köhler et al., 2018b] and the Human Disease Ontology (DO) [Schriml et al., 2018] to support automating searching for relations can contribute to the discovery of new relationships between entities. Therefore, ontologies have become an essential tool for biomedical research in which a vast amount of data is handled.

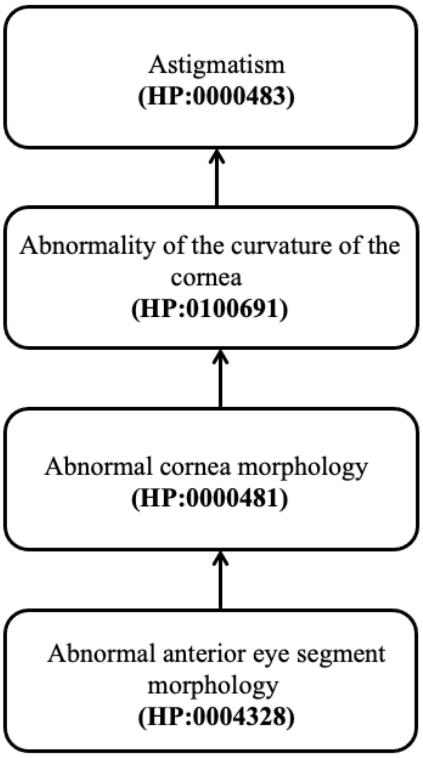
Astigmatism
(HP:0000483)
Abnormality of the curvature of the cornea
(HP:0100691)
Abnormal cornea morphology
(HP:0000481)
Abnormal anterior eye segment morphology
(HP:0004328)

Figure 1.2: An excerpt of the HPO ontology showing the first ancestors of *astigmatism*, using **is-a** relationships.

1.2 Objectives

The lack of annotated datasets that accurately characterise negative relations makes it harder to develop new systems for extracting negative, false, and unknown relations from biomedical literature. Therefore, this work aims to create a Negative Phenotype-Disease Relations (NPDR) dataset containing human phenotype and disease annotations, and their negative relations.

Thus, the main objective and contribution of this work is the creation of a dataset that characterises the negative relations from a gold-standard knowledge base of relations[1] provided by the Human Phenotype Ontology (HPO), the Negative Phenotype-Disease Relations (NPDR) dataset.

The general **hypothesis** of this dissertation is that it is possible to extract negative relations from biomedical literature, in order to create a knowledge base that can enable the development of better and more accurate text mining applications for negation identification, and RE of negative relations.

[1] http://purl.obolibrary.org/obo/hp/hpoa/phenotype_annotation_negated.tab

1.3 Methodology

The NPDR dataset was created by first gathering, from the HPO gold-standard knowledge base of negative phenotype-disease relations, the available Pubmed PMIDs associated with the negative relations. From these PMIDs it was possible to retrieve full-text articles that were subsequently manually examined. The analysis consisted in the description of the entities present in the negative relation, which comprises the analysis of phenotype entities, disease entities and their associated genes; the description of the sentence implying the negative relation, which comprises the characterization of the token of negation used in the sentence and the co-occurrence of the entities; and the description of the location of the entities and sentences in the article.

To evaluate the impact of the NPDR dataset on the automatic annotation and extraction of the entities and their negative relations from the articles retrieved for the creation of the dataset, a pipeline that performs NER and RE was implemented. NER recognizes human phenotypes and diseases entities, and RE extracts and classifies the relation between the identified entities. In order the obtain the articles in machine-readable format, two methods were used. Firstly, the PMIDs gathered from the NPDR dataset were converted to their corresponding PubMed Central ID (PMCID), to retrieve full-text articles using the PubMed API. Secondly, PDF articles were converted to text using the PDFMiner text extraction tool[2]. For the NER stage, a dictionary lookup solution, the Minimal Name-Entity Recognizer (MER) tool [Couto and Lamurias, 2018], was used for the annotation of the entities terms from the articles. More precisely, diseases and their related genes were annotated from the abstract and body of the article, and phenotypes were annotated from the sentences in the article that contained an adverb of negation, such as *no, not, negative, without* or *none*, or a word that implied negation of association, such as *impaired, impairment, normal, lack of* or *present exclusively*. The MER tool was chosen for its simplicity and efficiency, since it only requires a lexicon as a text file, with a list of terms containing the entities of interest as input for the annotation process [Couto and Lamurias, 2018]. Finally, the HPO gold-standard knowledge base was used to mark the negative relations obtained by the occurrence of phenotype entities in the sentences that implied a negative relation, and the occurrence of diseases and related genes present in the article. The relations were marked *True*, if the relation was in the knowledge base, or *False* otherwise.

1.4 Contributions

The main objective of this thesis is the analysis of negative phenotype-disease relations that can be found in biomedical literature. Related to this work, the overall contributions are as follow:

1. The creation of a negative phenotype-disease relation corpus that thoroughly characterizes these relations, the NPDR dataset.

2. 125 new negative phenotype-disease relations manually annotated at the document level.

[2] https://github.com/pdfminer/pdfminer.six/

1.5 Document Structure

Additionally, to the present introductory chapter, this document is structured in four chapters as follows:

- **Chapter 2** (Related Work) presents the basic concepts and resources of text mining for biomedical literature, text mining applications for negation identification, and an overview of knowledge bases that can aid in the annotation of negative relations between biomedical entities.
- **Chapter 3** (A Dataset of Negative Phenotype-Disease Relations) presents the work developed to create a dataset that characterizes the negative phenotype-disease relations present in the HPO gold-standard knowledge base.
- **Chapter 4** (Conclusion) presents the main conclusions of this work, and the ideas for future work.

Chapter 2

Related Work

This chapter presents the fundamental concepts of text mining for biomedical literature, text mining applications for negation identification, and an overview of knowledge bases that aid in the annotation of negative relations between biomedical entities.

2.1 Text Mining for Biomedical Literature

Text mining is an interdisciplinary field that gathers knowledge from data mining, information extraction, Machine Learning (ML), computational linguistics and statistics, to process and analyse unstructured text [Hotho et al., 2005]. Considering that a substantial amount of information is stored as text, research in text mining has been very active. Consequentially, various tools and applications have been developed in order to handle different types of documents. The expansion of these tools is especially relevant for biomedical text mining, given the abundance and heterogeneity of scientific literature. Different types of texts can have different forms depending on their nature (e.g., clinical report, journal paper, patent, book) [Friedman et al., 2002], and a vast range of terms can be used with distinctive styles of spellings, abbreviations and database identifiers [Lamurias and Couto, 2019]. Therefore, extracting facts and relationships in a structured form to derive meaning from multiple domains, is a challenging task. Nonetheless, text mining has generated a lot of interest from the bioinformatics community, since patterns and knowledge extracted from texts can be used to derive new facts or hypotheses, that can later be validated experimentally.

2.1.1 Natural Language Processing

Natural language processing (NLP) is a field of computer science that studies how computers can understand and derive meaning from human language [Manning and Schutze, 1999]. This Section will assess NLP techniques relevant to text mining pre-processing tasks [Lamurias and Couto, 2019], which will be described in Section 2.2.

- **Tokenization**: is the first step in NLP, and consists of the segmentation of a given text into elementary units for subsequent analysis. These tokens are defined as a sequence of characters and can be words, symbols, numbers, or phrases. For example, sentence splitting consists of the identification of the sentence boundaries of a text, i.e., splitting the text into sentences in order to extract the meaning of an independent sentence. This task can be challenging due to the differences in punctuation marks, such as a period corresponding to the end of a sentence or an abbreviation. The simplest form of tokenization is the separation of words by spaces in a text. However, regarding biomedical literature, this process needs to be more refined since the structure of scientific information is different from general language. This makes tokenization in biomedical literature particularly challenging since its terminology is inconsistently spelt, texts can be ungrammatical (i.e., they do not conform to grammatical rules) and often include abbreviations and acronyms. Biomedical terms also contain digits, special symbols (such as hyphens), greek or latin letters and capitalized letters within words. Therefore, biomedical tokenization demands specially designed strategies in order to minimize the propagation of errors in successive NLP analysis pipeline [Cruz Díaz and Maña López, 2015].

- **Part-of-Speech (POS) Tagging**: aims at attributing one or more categories to each token, corresponding to its syntactic functions (such as noun, verb, adjective or punctuation) and semantic context.

- **Lemmatization and Stemming**: consists of removing suffixes and inflexions of a token, in order to reduce it to its base form. The fundamental principle of these techniques is to group similar words by the same root or the same canonical citation [Nunzio and Vezzani, 2018]. The lemma determines the canonical form of the word, which is always a real word, and the stemma corresponds to a one and only fragment of a word. For example, the lemma of *amusement* will be *amuse* and the stem *amus*.

2.1.2 Text Mining Tasks

The text mining tasks described below [Jurafsky and Martin, 2009] are common to all sources and domains of text, but their performance may vary according to the domain they are applied to. Nonetheless, all of them have a general goal of helping us to identify useful knowledge.

- **Topic Modelling**: is the classification of documents according to their themes, so that they can be organised in order of relevance to a given topic.

- **Named-Entity Recognition (NER)**: seeks to identify and classify the entities, i.e., pieces of text relevant to a given domain, which can be composed of one or more tokens, specified in the text, locating it by the offset of its first and last character. The class assigned to the entity depends on the concept it is being referred to in the text.

- **Named-Entity Linking (NEL)**: will link the entities to a formal identifier that can be found in an external database or ontology.

- **Relation Extraction (RE)**: identifies the relations between the entities that participate in a relation described in the text. Conventional tools will mainly extract the relation between two entities that are in the same sentence, by applying co-occurrence. Another more complex approach would be document-level RE, where the relationship among the entities is extracted from a paragraph [Han and Wang, 2020].

- **Event Extraction**: is an extension of the relationship extraction task, and has the purpose of identifying the label of the relationship and the role of each participant.

2.1.3 Text Mining Approaches for Relation Extraction

A wide range of approaches has been applied in the biomedical field for extracting relations. These techniques can be broadly categorised into three groups, which are rule-based (including pattern-based), co-occurrence and Machine Learning (ML)-based [Lamurias and Couto, 2019]:

- **Co-occurrence**: identifies co-occurring entities in a sentence, and is the simplest approach to identify or extract relations between entities.

- **Rule-based**: uses manually as well as automatically defined rules or patterns to extract relations. Manual patterns are defined by domain experts, and automatic patterns use bootstrapping or are directly generated from corpora.

- **Machine Learning (ML)-based**: takes advantage of large annotated biomedical corpora to apply supervised learning, in which RE tasks are modelled as classification problems. Mainly, these tasks consist of pre-processing, parsing and RE [Muzaffar et al., 2015]. Supervised methods can be further categorised in feature-based and kernel-based methods. **Feature-based methods** extract a set of features from textual analysis, and represent them in a feature vector that will be presented to a classifier, either for training or classification. These methods require heuristic choices, hence, to maximise the performance of the classifier, these features have to be selected on a trial-and-error basis and setting on the most favourable set of features can prove to be difficult since some of them are not good indicators of entity relationships. **Kernel-based methods** use specialised kernels designed for RE in order to utilise representation of data. The main goal is to assess the similarity between the different data instances and compute the similarity of their representations. These methods can be string kernels, where the similarity is computed for two strings at the character level; bag-of-features kernel, where the similarity is computed at the word level, and word-context around entities can also be used to extract relations; and tree kernels, where the similarity is computed between structured shallow parse trees built on the sentence. Although distant supervised methods can yield high performances, this metric will greatly depend on the quality of the designed features or kernels. On the one hand, the pre-processing data stage required to provide the necessary labeled data is error prone, and on the other hand, these methods are difficult to extend to new unlabelled entity-relation types [Bach and Badaskar, 2007].

Recently **deep learning** approaches have gained significant interest since they can directly extract features from large-scale data, and consequently rely less on NLP tools. The goal of these approaches is to create a large network that is able to capture the features in the data. Deep learning studies were originated from artificial neural networks research, in which an interconnected combination of processing units can generate knowledge from information from the environment. This approach is more effective than conventional ML since it can handle more non-linear and abstract representations [Leng and Jiang, 2016]. Some common concepts across most deep learning models are words embeddings, which consist of similarity vectors measured by semantic relevance; Convolutional Neural Networks (CNN), which are formed by an input layer, multiple hidden layers, and an output layer; and Recurrent Neural Networks (RNN), which make use of sequential information that is used to process sequences of inputs. RNNs execute the same task for every element of a sequence, and the output of one calculation is dependent on the previous output. This model has two types, Long Short-Term Memory Networks (LSTM) and Bidirectional RNN [Xue et al., 2018].

2.1.4 Text Mining Tools for Relation Extraction

Below are briefly described some of the state-of-the-art tools used for biomedical RE.

- **Textpresso**: is a biomedical information extraction technique based on ontologies and regular expressions. It can be used as a curation tool, as well as a search engine for researchers [Müller et al., 2004].

- **jSRE**: uses a combination of shallow linguistic kernel functions in order to combine different information sources, the whole sentence where the relation appears and the local contexts around the interacting entities. Therefore, this kernel considers tokens, POS, and lemmas around each entity [Giuliano et al., 2006b]. This system has been used to extract drug-drug relations [Segura-Bedmar et al., 2011].

- **Excerbt**: is a text mining tool based on semantic sentence analysis combined with rule-based and ML approaches, that classifies relations between genes, proteins, phenotypes and diseases [Blohm et al., 2013].

- **BeFree**: is a text mining system that applies a kernel-based approach using both morpho-syntactic and dependency information to identify drug-disease, drug-target and gene-disease relations from text [Bravo et al., 2015].

- **DeepDive**: applies DS to perform RE tasks in order to identify gene-gene relations [Mallory et al., 2015].
- **IBRel**: is a Distant Supervision (DS)-based multi-instance learning tool that extracts microRNA-gene relations from text [Lamurias et al., 2017].
- **BioBERT**: is a pre-trained language representation model for biomedical text mining based on BERT [Devlin et al., 2019] that has been used to effectively extract human phenotype-gene relations [Lee et al., 2019].
- **SciBERT**: is a pre-trained BERT-based language model that leverages unsupervised pretraining on a large multi-domain corpus of scientific publications to improve scientific NLP tasks [Beltagy et al., 2019].
- **BiOnt**: is a deep learning system that uses four types of biomedical ontologies, such as the Gene Ontology, the Human Phenotype Ontology, the Human Disease Ontology and the Chemical Entities of Biological Interest in order to extract phenotype-gene relations, drug-drug interactions and chemical-induced disease relations [Sousa and Couto, 2020].
- **DEMMT**: is a document-level entity mask method with type information, that masks each mention of the entities by special tokens [Han and Wang, 2020].
- **PubMedBERT**: is a pre-trained BERT-based language model pretrained using abstracts from PubMed, which achieves state-of-the-art performance on several biomedical NLP tasks, as shown on the Biomedical Language Understanding and Reasoning Benchmark [Gu et al., 2020].

Table 2.1 describes the performance of some of the RE systems mentioned above. It should be noted that these systems cannot be directly compared, since each one was evaluated on a different corpus to classify relations between different biomedical entities.

Table 2.1: Evaluation of biomedical RE systems.

Name	Approach	Evaluation Corpus	Precision	Recall	F-Measure
Textpresso	Rule-based	Eight full-text journal articles	0.618	0.354	0.450
jSRE	ML	AImed	0.609	0.572	0.590
IBRel	ML/DS	TransmiR	0.359	0.486	0.413
BeFree	ML	EU-ADR	0.840	0.710	0.760
DeepDive	ML/DS	100 000 full-text PLOS articles	0.760	0.490	0.596
BioBERT	ML	NCBI disease	0.883	0.890	0.886
BiOnt	ML	PGR	0.842	0.666	0.744

2.1.5 Text Mining Tools for Negation Detection

Negation is a linguistic phenomenon defined as *an assertion that some event, situation, or state of affairs does not hold. Negative clauses usually occur in the context of some presupposition, functioning to negate or counter-assert that presupposition* [Payne, 1997]. Since the biomedical domain has an abundance of negation statements, negation detection has become an important subtask of information extraction from texts, such as clinical reports and scientific articles. Yet, this task is not trivial because the complexity of language makes it challenging to identify the polarity of the information. Since the meaning of a concept can significantly be affected by the scope of negation, identifying negated concepts based on the presence of a negation token can lead to inaccurate information. For example, in the following sentence: *Radial dysplasia is not commonly associated with malformations of the lower limbs.* (PMID:7077621), the negation cue indicates that radial dysplasia is not associated with malformations of the lower limbs, therefore, the negation token *not* affects the whole sentence. But in the sentence: *Investigation of plasma K^+ concentrations in the members of the family revealed a similar syndrome in two of the three children subjects C and D but not in the husband* (PMID:2766660), the negation cue only partially negates the sentence. Finally, in sentence: *The present cases might be similarly characterized but extrapulmonary manifestations could not be excluded in these living patients* (PMID:1190822), the negation cue is not being used to negate the scope of the sentence.

Negation can be intuitive for humans but delineating the scope of negation can be difficult for computer-based systems. The initial approaches for tackling this task were made using rules and designed heuristics but given the complexities of natural language other methods were developed, such as ML systems. Below are described state-of-the-art tools for negation detection, and Table 2.2 summarizes the performance for some of the systems.

- **NegExpander**: is an algorithm that uses syntactic processing techniques to identify noun phrases or conjunctive phrases that define negation boundaries [Aronow et al., 1999].
- **Negfinder**: is a rule-based system that recognises negated patters in biomedical text. This system uses a lexical scanner called *lexer*, that uses regular expressions in order to generate a finite state machine, and a parser [Mutalik et al., 2001].
- **NegEx**: is a simple regular expression algorithm that implements phrases indicating negation, filters out sentences containing phrases that falsely appear to be negation phrases, and limits the scope of negation phrases [Chapman et al., 2001].
- **NegHunter**: is a negation detection algorithm that identifies negation triggers in clinical practice guidelines [Gindl et al., 2008].
- **DepNeg**: is a dependency parsed-based negation algorithm that uses the dependency structure of a target named entity in a sentence, instead of a fixed negation scope. This system applies manual

negation rules based on the patterns of dependency paths between the targets and negation cues [Sohn et al., 2012].

- **DEEPEN**: is a negation algorithm developed to decrease NegEx's false positives by considering the dependency relationship between negation words and concepts within a sentence using Stanford dependency parser. This system was developed and evaluated in electronic health records data from Indiana University (IU) and Mayo Clinic datasets [Mehrabi et al., 2015].
- **NegMiner**: is a tool that exploits basic syntactic and semantic information to deal with contiguous and multiple negations [Elazhary, 2017].
- **NegBERT**: is a model that uses BERT's transformer-based architecture to negation detection and scope resolution. This model trains Deep Learning systems on corpora, such as the BioScope Corpus, the Sherlock Dataset and the SFU Review Corpus [Khandelwal and Sawant, 2020].

Table 2.2: Evaluation of biomedical negation detection systems.

Name	**Approach**	**Evaluation Corpus**	**Precision**	**Recall**	**F-Measure**
Negfinder	Rule-based	Surgery notes & Discharge summaries	0.918	0.957	0.929
NegEx	Rule-based	Discharge summaries	0.845	0.778	0.804
NegHunter	Rule-based	Clinical practice guidelines	0.675	0.835	0.747
DepNeg	Dependency parser	Mayo clinical notes	0.967	0.739	0.838
DEEPEN	Dependency parser	IU dataset	0.966	0.964	0.965

2.2 Textual Resources for Biomedical Natural Language Processing

Corpora are paramount to train, test, compare or develop text mining systems. On the other hand, datasets and ontologies are also valuable resources for research on linguistic analysis of scientific and clinical literature. This section describes some of the state-of-the-art ontologies, corpora and datasets used in biomedical NLP.

2.2.1 Ontologies

An ontology is *an explicit specification of a conceptualisation of a domain*, as defined by Gruber [Gruber, 1993]. Therefore, an ontology defines a structure in which common vocabulary is used to represent shared knowledge. The domain knowledge in biomedicine keeps on increasing in size and complexity, consequently various bio-ontologies have been developed to overcome the need to merge and organise all the information available. Phenotypes and diseases are some of the biomedical entities structured in publicly available ontologies:

- **The Human Phenotype Ontology (HPO)**: is a standardised vocabulary of phenotypic abnormalities associated with diseases, and serves as a database for deep phenotyping in the field of rare diseases. The organization responsible for this ontology also provides a gold-standard relations file of positive gene-phenotype relations and negative disease-phenotype relations [Köhler et al., 2018b].
- **The Human Disease Ontology (DO)**: is a standardised ontology of human diseases that provides descriptions of human disease terms, phenotype characteristics, and related medical vocabulary disease concepts through cross mapping of DO terms to MeSH, International Classification of Diseases (ICD), National Cancer Institute's (NCI) thesaurus, Systematised Nomenclature of Medicine Clinical Terms (SNOMED) and OMIM [Schriml et al., 2018].

2.2.2 Corpora

A corpus is a collection of written texts on a specific topic, within a meaningful context and with a general purpose, upon which a linguistic analysis is based. Most of them are annotated with NLP techniques and domain-experts and are often assembled to answer a research question. Hence, high quality annotated corpora are paramount for developing and evaluate text mining tools, and training ML models. Even though document retrieval is achieved by automatic processes, domain experts should manually annotate and curate the annotations, which are arduous efforts. Therefore, text mining tools may be used to assist curators by supplying automatic annotations as a baseline.

There are many types of corpora available, which makes it important to choose the right one for the type of annotation to be extracted. For example, for RE it is necessary to have a description in the text of the relations between the entities and experts from multiple domains. Also, not all corpora come at the same cost. Annotated corpora for RE are amongst the most expensive type of corpus since it first requires the recognition of the entities present in the text by the annotators, and then the relationships described between them. Nevertheless, the development of annotated corpora is important for the promotion of RE research in the biomedical field, even if it is a time-consuming and error-prone task.

Below are described some of the state-of-the-art biomedical corpora, that have been developed to provide reference material for applying biomedical RE, and for negation detection.

- **GENIA**: is a semantically annotated corpus of biological literature, consisting of a collection of 2000 abstracts extracted from the MEDLINE database. It contains more than 400 000 words and almost 100 000 annotations hand-coded by two domain experts for biologically meaningful terms, which have been semantically annotated with descriptors from the GENIA ontology [Kim et al., 2003].
- **BioText**: is a corpus of 100 titles and 40 abstracts from Medline with annotated diseases and treatments (both drug and medical treatments) at the sentence level, and positive and negative relations

between the entities [Rosario and Hearst, 2004].

- **BioInfer (Bio Information Extraction Resource)**: is an annotated corpus of biomedical English, containing 1100 sentences from abstracts of biomedical research articles, annotated for relationships (positive and negative), named entities, and syntactic dependencies [Pyysalo et al., 2007].

- **Bioscope**: is a corpus composed of free texts, biological full papers and biological scientific abstracts. It contains annotations at the token level for negative and speculative keywords and at the sentence level for their linguistic scope. The corpus consists of over 20 000 sentences, and over 10% of them contain one or more linguistic annotations that suggest negation or uncertainty [Vincze et al., 2008].

- **EU-ADR corpus**: is a corpus of 300 abstracts with drugs, disorders and targets annotated with their inter-relationships (target-disease, target-drug and drug-disease) [van Mulligen et al., 2012].

- **NCBI disease corpus**: is a corpus annotated with disease names and their corresponding Medical Subject Headings (MeSH) and/or OMIM identifiers. This platform contains almost 800 PubMed abstracts and more than 6800 disease mentions linked to unique concepts, providing a large foundation for improving text-mining research on NER, namely by serving as a training corpus for ML models [Doğan et al., 2014].

- **NegDDI-DrugBank** is an expansion of the DrugDDI corpus, a corpus of pharmacological substances and drug–drug interactions [Herrero-Zazo et al., 2013], annotated with negation cues and scopes, following the BioScope guidelines. It contains 1448 sentences with at least one negated scope [Bokharaeian et al., 2014].:

- **Psychiatric disorders Gene association NETwork (PsyGeNET)**: is a high-quality resource of psychiatric diseases and their associated genes. The information contained in the database is extracted from MEDLINE abstracts with text mining tools, namely the BeFree system, and manualy curated by experts in psychiatry and neurosciences. This knowledge base includes positive as well as negative relations [Gutiérrez-Sacristán et al., 2015].

- **Biomedical entity Relation ONcology COrpus (BRONCO)**: is a gold-standard corpus that contains over 400 genomic variants and their relations with genes, diseases, drugs, and cell lines in the context of cancer anti-tumor drug screening research. The variants and their corresponding relations were manually extracted from 108 full-text articles [Lee et al., 2016].

- **DisGeNET**: is a knowledge base of genotype-phenotype relations, that integrates data from various sources, namely GWAS catalogues, UniProt, ClinVar, The Comparative Toxicogenomics Database, Orphanet, The Genetic Association Database and The Mouse Genome Database. This platform contains a comprehensive collection of more than 1 000 000 associations of genes and variants of human diseases, such as Mendelian, complex, rare, and environmental. The terms are annotated with controlled vocabularies and community-driven ontologies but are not validated by experts [Pinero et al., 2016].

- **SNPPhenA**: is a corpus of associations of single-nucleotide polymorphisms (SNPs) and phenotypes, annotated with linguistic-based negation, modality markers, neutral candidates, and confidence level of associations. [Bokharaeian et al., 2017].
- **Phenotype-Gene Relations (PGR) corpus**: is a silver standard corpus of human phenotype and gene annotations and their relations. This corpus was generated in a fully automated manner and partially evaluated by eight curators. It consists of 1712 abstracts, 5676 human phenotype annotations, 13835 gene annotations and 4283 relations [Sousa et al., 2019].
- **PGxCorpus**: is a manually annotated corpus, designed to enable automatic RE of pharmacogenomics relationships from text. It consists of 945 sentences from 911 PubMed abstracts, mainly annotated with gene variations, genes, drugs and phenotypes, and their relationships [Legrand et al., 2020].

To this date, annotated corpora focus mainly on positive relations, as shown in Table 2.3, where a summary of biomedical corpora relevant to RE systems is provided. The lack of corpora for negative relations annotated at the semantic level hinders the development of systems able to extract these relations.

Table 2.3: Biomedical corpora relevant to RE systems.

Name	**Entities**	**Relation Type**
BioText	Diseases and treatments	Positive & **Negative**
BioInfer	Genes, proteins and RNAs	Positive & **Negative**
EU-ADR	Drugs, diseases and targets	Positive
NCBI disease corpus	Diseases	Positive
NegDDI-DrugBank	Drugs	**Negative**
PsyGeNET	Genes and psychiatric disease phenotypes	Positive & **Negative**
BRONCO	Genes, diseases, drugs and cell lines	Positive
DisGeNET	Genes and disease phenotypes	Positive
SNPPhenA	SNPs and phenotypes	Positive & **Negative**
PGR	Genes and phenotypes	Positive
PGxCorpus	Genes, phenotypes and drugs	Positive

2.2.3 Databases

Databases are another resource that can be as useful as corpora and serve as resources to aid in biomedical text mining. Some of the state-of-the-art databases in biomedical NLP, and detection of negated events are listed below.

- **Online Mendelian Inheritance in Man (OMIM)**: is a relation knowledge base of human genes and genetic disorders, with more than 24 600 entries, that aims to support human genetics research and the practice of clinical genetics. It provides relations between phenotype terms, genetic disorders (diseases), and genes [Hamosh et al., 2005].
- **BioN∅T**: is a searchable database of biomedical negated sentences. It contains approximately 32 million negated sentences extracted from over 336 million biomedical sentences, from full-text biomedical articles in Elsevier and PubMed Central, and abstracts in PubMed [Agarwal et al., 2011].
- **Orphanet (ORPHA)**: is a European relation dataset of disease-gene relations that aims to help improve the diagnostic process, care and treatment of patients with rare diseases. This knowledge base establishes a link between rare genetic diseases and any published information regarding them, and genes are cross-referenced with UniProt, OMIM, HGNC, and Genatlas [Rath et al., 2012].
- **Negatome**: is a database of proteins and proteins domains that are unlikely to interact physically. It is extracted by manual curation of literature and by analysing three-dimensional structures of protein complexes. The manual annotation process is also aided with the text mining tool Excerbet. This database contains both positive and negative examples [Blohm et al., 2013].
- **Phenotype Annotation Negated (HPO)**: is a gold-standard relations file of negative phenotype-disease relations that contains over 1500 annotations. The main categories of this file are the name of the database from which comes the disease, the disease identifier and name, the phenotype's HPO identifier and the reference for the negative relation [Köhler et al., 2018b].

The biomedical literature contains a vast amount of information regarding the associations of biomedical entities, and the availability of published literature in electronic format makes the development of text mining applications even more appealing for understanding the relations among biological systems. But biomedical RE is a task that requires high quality annotated corpora (gold-standard corpora) which are scarce since its construction is laborious and costly. Furthermore, when extracting negative relations, in addition to having to correctly identify and name the entities present in text (NER), the presence of a token of negation that implies no association between the entities also must be taken into consideration. To this date, and to the best of my knowledge, the only gold-standard knowledge base for negative phenotype-disease relations is a document made available by the HPO. Even though this gold-standard relations file is a valuable resource for detecting negative phenotype-disease relations, it does not contain sufficient information to enable the automatic extraction of new negative phenotype-disease relations from biomedical literature, such as synonyms or the negation token that indicates the negative relation.

Chapter 3

A Dataset of Negative Phenotype-Disease Relations

In order to aid in the automatic annotation of negative relations a dataset that characterizes some of the relations from the Human Phenotype Ontology (HPO) knowledge base, in addition to new manually annotated negative relations, was created. The Negative Phenotype-Disease Relations (NPDR) dataset comprises a thorough description of the entities present in the text, their location in the article, and the location of the sentences that contain the negative relation.

3.1 Methods

The HPO is an ontology of phenotypic abnormalities that occur in human diseases. The developers of the HPO also made available a gold-standard knowledge base that links these abnormalities to diseases to which they are not associated. These negative relations are derived by text mining from texts in the Online Mendelian Inheritance in Man (OMIM) and Orphanet (ORPHA) databases, where all phenotype abnormalities that are not associated with a disease are assigned to a disease in the knowledge base. The negative phenotype-disease relations from this document were used in the creation of the NPDR dataset.

From the HPO knowledge base of phenotype-disease relations it was possible to link a subset of 237 negative relations to 169 PubMed IDs (PMID), which led to the retrieval, in PDF format, of 169 articles and clinical reports, seven supplementary material documents, plus one book chapter, that were subsequently manually examined. The NPDR dataset was then created from this analysis, and comprises 347 characterised negative relations, from which 222 derive from the HPO knowledge base of phenotype-disease relations, and 125 are new manually annotated relations. It should be noted that the investigation of every annotation was performed at the document-level.

The main components of the dataset are the characterization of the **entities** (diseases, their associated genes, and phenotypes), the characterization of the **sentences** that imply a negative phenotype-disease relation, and the **location** of the entities and sentences in the text. Each negative relation can also be linked to the article PMID and PubMed Central ID (PMCID) (if available) from where it was annotated, and an evidence code tag used by the HPO. The contents of the dataset are shown in Table 3.1.

Table 3.1: Summary description of the data categories of the NPDR dataset.

Categories	**Description**
pmid_reference	Article PMID
pmc_reference	Article PMCID
evidence_code	Evidence code tag used by HPO
disease_db	Database from which the disease is sourced (OMIM)
disease_name	OMIM designation
disease_abbreviation	Disease abbreviation found in text
associated_gene_or_locus	Associated gene or locus to the disease
disease_synonym	Synonym of the disease found in text
disease_id	OMIM identifier
disease_id_in_text	Presence of the disease identifier in text
gene_id_in_text	Gene OMIM identifier and its presence in text
hpo_name	HPO designation
hpo_id	HPO identifier
hp_synonym	Synonym of the phenotype found in text
negation	Word of negation in sentence
not_abnormal	Presence of the word *normal* in sentence
sentence	Sentence containing the negative relation
sentence_location	Sentence location in text
disease_location	Disease location in text
gene_location	Gene location in text
co_occurrence	Co-occurrence of entities in sentence
comments	Further comments

3.1.1 Disease Annotations

Disease entities were characterised in terms of their official OMIM disease **designation** (*disease_name*), OMIM disease **identifier** (*disease_id*), **synonym** (*disease_synonym*), and **abbreviation** (*disease_abbreviation*). In the case where the gene or locus associated to the disease participating in the negative phenotype-disease relation was mentioned in the text, as well as referenced in the OMIM database, its OMIM **desig-**

nation (*associated_gene_or_locus*) and **identifier** (*gene_id_in_text*) were also annotated. Furthermore, the presence of the OMIM identifiers were characterized in terms of their presence in the article. For diseases, if the identifier was present, the column *disease_id_in_text* would contain *yes*, and *no* otherwise. For genes, if the identifier was present in the text, the column *gene_id_in_text* would contain the identifier, and *no* otherwise.

In the HPO gold-standard knowledge base, the disease name corresponds to the main disease OMIM term. However, in the NPDR dataset, the official designation corresponds to the closest name of the entity found in text. Thus, whenever an official OMIM designation was not present, the disease name was chosen based on the synonym found in text. Example 3.1 shows how the disease name *ADENOSINE DEAMINASE 2 DEFICIENCY; DADA2* was chosen for article PMID:26867732. Also, it is worth mentioning that the official OMIM designation is composed by the disease's full name (*ADENOSINE DEAMINASE 2 DEFICIENCY*) followed by its abbreviation (*DADA2*).

The disease abbreviation could either be an abbreviation chosen by the article's authors, an official OMIM abbreviation, or the name of the associated gene or locus. In some cases, an exact match between the three would occur. For example, in article PMID:21057262, the OMIM disease abbreviation for *MYD88 DEFICIENCY; MYD88* (MIM 612260) is *MYD88D*, as well as the disease abbreviation used by the authors, and the gene name. On the other hand, in article PMID:31630789, the OMIM disease abbreviation for *SPONDYLOEPIMETAPHYSEAL DYSPLASIA, ISIDOR-TOUTAIN TYPE; SEMDIST* (MIM 618728) is *SEMDIST*, the disease abbreviation chosen by the authors is *SEMD*, and the gene name is *RPL13*.

Example 3.1 Example of how a disease name is chosen.

- **Article PMID:** 26867732
- **Disease Identifier:** 615688
- **Disease Synonym:** *ADA2 deficiency*
- **Chosen Disease Name (OMIM alternative title):** *ADENOSINE DEAMINASE 2 DEFICIENCY; DADA2*
- **Disease Name from HPO dataset (main OMIM term):** *VASCULITIS, AUTOINFLAMMATION, IMMUNODEFICIENCY, AND HEMATOLOGIC DEFECTS SYNDROME; VAIHS*

3.1.2 Phenotype Annotations

Phenotype entities were characterized in terms of their HPO **designation** (*hpo_name*), HPO **identifier** (*hpo_id*), and phenotype **synonym** (*hp_synonym*). Since the HPO gold-standard knowledge base only contained the HPO identifier of the phenotype that participates in the negative relation, if the official

HPO designation was not found in text, the name to which the entity was referred was annotated and corresponds to its synonym. Like disease entities, phenotype synonyms are not an official HPO synonym, and phenotype names were also chosen to match as closely as possible the phenotype synonym. Therefore, the closest designation can either be the main HPO term of a given phenotype, or an official synonym. Example 3.2 illustrates how the phenotype name *degeneration of cerebellum* was chosen for article PMID:11706389.

Example 3.2 Example of how a phenotype name is chosen.

- **Article PMID:** 5571218
- **Phenotype Identifier:** HP:0000662
- **Negative Relation Sentence:** *There has been no further subjective loss of vision in the past 30 years, and he denies any impairment in night vision or color vision*
- **Phenotype Synonym:** *impairment in night vision*
- **Main HPO term:** *nyctalopia*
- **Chosen Phenotype Name (HPO official synonym):** *poor night vision*

3.1.3 Negative Relation Sentence Annotations

The sentences were identified based on two criteria. First, they had to contain the name or synonym of the phenotype that participated in the negative phenotype-disease relation, and second, a word that implied negation of association had to be present. Since the HPO is a database of phenotypic abnormalities [Köhler et al., 2018a], if the word *normal* appeared in the sentence and was related to the phenotype, the negation of association was established. Alternatively, adverbs of negation were also considered as suggesting a negation of association.

Occasionally, a negative relation could be established in the absence of a sentence. This exception would arise in the presence of a table summarising the clinical features of all the individuals affected by the same disease. Table 3.2 illustrates this situation, where a negative relation can be established between Phenotype 2 and the Clinical Syndrome. In a situation like the one described before, the name(s) of the phenotype(s) that appeared in the table would be considered as the negative relation sentence. A list of the words of negation found in the sentences, as well as tables, can be found in Table 3.3.

Co-occurrence of the entities in the sentence was also considered, and the type of entity term that co-occurred with the phenotype entity was annotated. For example, in the sentence (PMID:26996948): *The first reported* ***inherited GPI deficiency (IGD)*** *was* ***PIGM (MIM 610273)*** *deficiency in individuals*

Table 3.2: Example of how a negative relation can be implied from a clinical feature table. In this example, a negative relation between Phenotype 2 and the Clinical Syndrome would be implied.

Clinical Features	**Clinical Syndrome**		
	Patient 1	**Patient 2**	**Patient 3**
Phenotype 1	no	no	yes
Phenotype 2	**no**	**no**	**no**
Phenotype 3	yes	yes	yes

Table 3.3: Words of negation and their most common location.

Location	**Negation Words**
Sentence	*absence, absent, denied, excluded, free of, impaired, impairment, lack of, negative, no, none, not, present exclusively, rules out, without*
Table	*neg, N, NP*

suffering from portal thrombosis and seizures without ***intellectual disability***, the disease, gene and phenotype entities co-occur. More specifically, the entities types are the disease synonym (*inherited GPI deficiency*), the disease abbreviation (*IGD*), the gene name (*PIGM*), the gene identifier (*MIM 610273*), and the phenotype name (*intellectual disability*).

Finally, every negative association between a disease and HPO term can be linked to an HPO evidence code. These codes are: inferred from electronic annotation (**IEA**), published clinical study (**PCS**), individual clinical experience (**ICE**), inferred by text mining (**ITM**) and traceable author statement (**ITAS**) [Köhler et al., 2013].

3.1.4 Entities and Sentences Location

Regarding the location of the entities and sentences in the articles, the main categories under consideration were *title, abstract, body* and *appendices* (this last category was only applied to sentences). If the *body* category was composed by other subsections, they were also annotated, but for analysis purposes were subsequently attributed to an equivalent section, such as *introduction, materials and methods, results, discussion*, and *tables*. These attributed sections were chosen depending on the context of the content of the original section. It should be noted that when a negative relation sentence was not found in an article and its supplementary materials were made available, then the latter document was retrieved to be analysed. A list of every article location and attributed location can be viewed in Table 3.4.

Table 3.4: Summary of article locations from the NPDR dataset and their equivalent location.

Equivalent Location	**Article Location**
Introduction	Background
	Biological Relevance
	Introduction
	Summary
Materials and Methods	Case report(s)
	Candidate HSP Genes Identified by Network Analysis
	Clinical and diagnostic relevance
	Clinical Characteristics
	Clinical Relevance
	Clinical report(s)
	Diagnosis
	Extending Results to Larger HSP Cohort
	Family history
	Functional Testing Candidates with Expression and Zebrafish
	Genetic Counselling
	Genetically Related Disorders
	HSP-Related Proteins Interact Within a Network
	Implicated Causal Genes Suggest Modules of HSP Pathology
	Link Between HSP and Neurodegenerative Disease Genes
	Material(s) and Methods
	Methods
	Molecular Genetics
	Multiple Genes Are Implicated in HSP
	Patients and Methods
	Subjects and Methods
	Variants
Discussion	Discussion
	Future Prospects
	Results and Discussion
Results	Results
Table	Tables
Appendices	Supplemental Data
	Supplementary Data
	Supplementary Material
	Supporting Information

3.2 Evaluation

To test if it was possible to automatically annotate the entities described in the NPDR dataset, and extract their negative relation, a system that performs Named Entity Recognition (NER), Named Entity Linking (NEL) and Relation Extraction (RE) was developed. The information from the NPDR dataset, the OMIM knowledge base, the HPO database and the HPO gold standard knowledge base was used as input for this system, in order to compare the difference in the number of annotations and relations extracted by using the HPO and OMIM knowledge bases or complementing their information with the NPDR dataset. Figure 3.1 shows the overall workflow of the developed system.

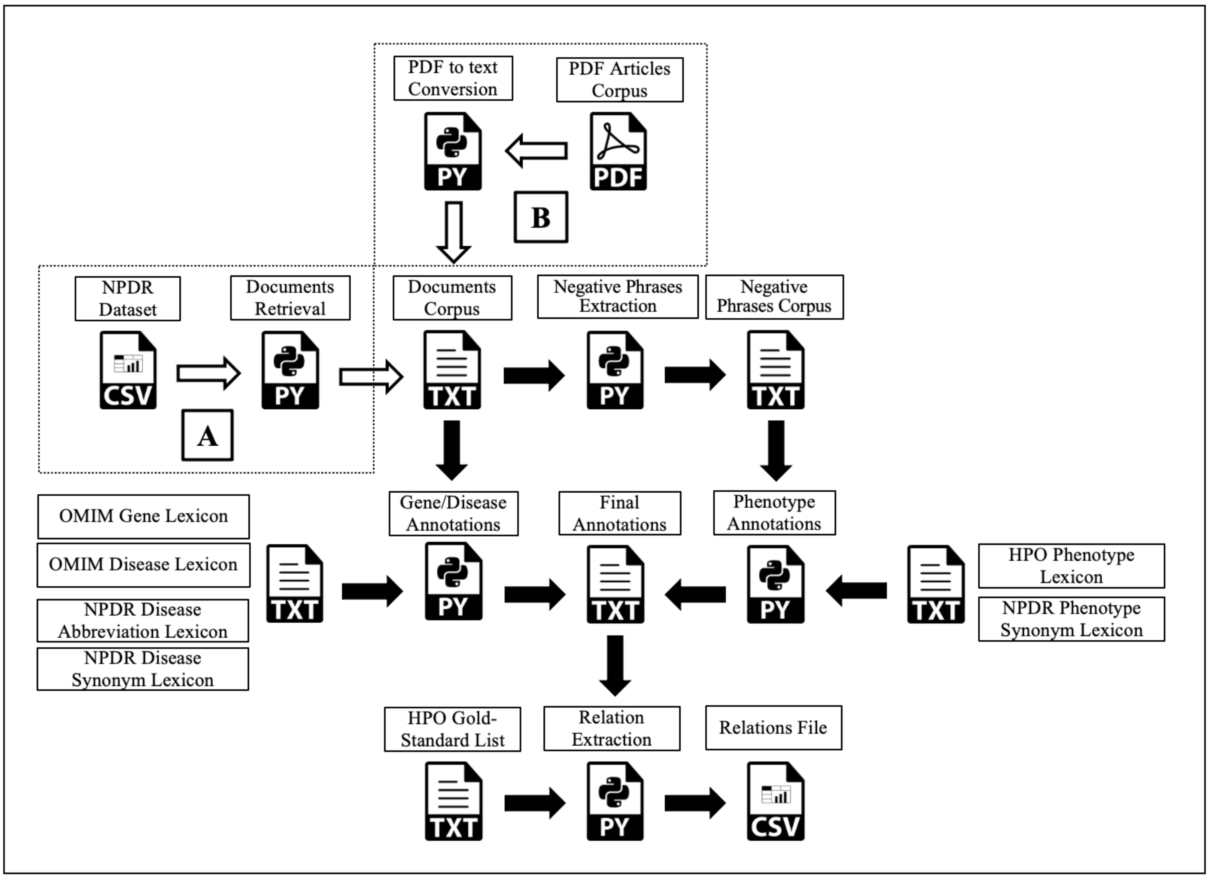

Figure 3.1: Workflow of the NPDR pipeline with the XML Corpus (**A**) and PDF Corpus (**B**).

3.2.1 Automatic Extraction Method

From the annotation of the NPDR dataset, it was possible to observe that most negative relation sentences were in the article's body. Therefore, the main constraint for the annotation of negative relations is that articles have to be available as full-texts. Consequently, in order to retrieve the full-text, the PubMed

PMIDs of the articles from the NPDR dataset were converted to their corresponding PubMed Central PMCID. Using the Entrez Programming Utilities (E-utilities) web service[1] and the 169 PMIDs from the NPDR dataset as keywords, it was possible to obtain the PMCID of 104 articles. Unfortunately, for 80 of these 104 PMC articles *The publisher of this article does not allow downloading of the full-text in XML form*, which led to the retrieval of 24 articles' bodies and 24 abstracts (**XML Corpus**). These documents were retrieved on May 4, 2021. All documents were filtered to only retrieve articles in English with a correct XML format and content.

Since so few articles were retrieved from the E-utilities web service, the 177 PDFs articles, clinical reports and book chapter that were gathered during the development of the NPDR dataset were converted to text, using the PDFMiner text extraction tool[2]. From this conversion it was possible to obtain 168 full-text articles, with their abstracts included (**PDF Corpus**).

Table 3.5: Summary description of the source of the documents, method used to retrieve them and final number of documents retrieved for the XML and PDF Corpora.

Corpus	Source	Method	Final number of documents
XML	PMCIDs from NPDR dataset	E-utilities web service	24 articles & 24 abstracts
PDF	article PDFs	PDFMiner tool	168 complete articles

3.2.1.1 Negation Extraction

After retrieving the documents, the next step was to apply a simple regular expression algorithm to detect negation phrases. The patterns used to extract sentences indicating negation were the tokens of negation found in the negation sentences from the NPDR dataset (Table 3.3).

3.2.1.2 Named Entity Recognition and Named Entity Linking

For the entities' annotation, the Minimal Named-Entity Recognition (MER) tool [Couto and Lamurias, 2018] was used. MER is a dictionary-based NER tool that only requires a lexicon or ontology (e.g., text or OWL file) with the list of terms containing the entities of interest and an input text in order to return the recognized entities, their location, and links to their respective classes.

Several lexica were used to annotate diseases with MER. The main lexicon was created from the file of OMIM's Synopsis of the Human Gene Map (morbidmap.txt[3]). This file contains 8052 phenotype-gene relationships of the Human Genome (as of July 12, 2021), more specifically the name of the phenotype (i.e., the disease's name), its MIM number (if different from that of the gene/locus), its gene symbol, the

[1] https://www.ncbi.nlm.nih.gov/books/NBK25501/
[2] https://github.com/pdfminer/pdfminer.six/
[3] https://https://www.omim.org/downloads

MIM number of the gene/locus and its cyto location. Since several diseases can also be identified by their gene or associated locus, besides a list of diseases and their identifier, a list of genes and their identifiers was also created. Considering that one gene can be associated to more than one disease, i.e., one gene name can be attributed to several MIM numbers, the *get_entities.sh* file from MER had to be modified to annotate gene names with more than one identifier. The other two lexica used for the annotation of diseases were generated from disease synonyms and diseases abbreviation synonyms from the NPDR dataset. Disease annotations were made at the sentence level for the article's bodies and full-text for abstracts.

Phenotype entities were annotated from two lexica. The main lexicon was created from the HPO[4], and the second from a list of phenotype synonyms from the NPDR dataset. Phenotypes were annotated from negation sentences only.

3.2.1.3 Relation Extraction

Relations were extracted using a distant supervision approach with the HPO gold standard knowledge base. The relations were marked as *True* if they were present in the knowledge base, or *False* if they were not identified or did not exist. To achieve this classification, the system extracted pairs of phenotype entities present in each negative sentence, and disease entities from the abstracts and article's body from where the negative sentence was retrieved (Example 3.2).

To compare the impact of the NPDR dataset in the RE phase, two methods were applied. The first method (**Labels Method**) consisted of extracting the relations between the entities annotated with the OMIM and HPO lexica. The second method (**AllLabels Method**), was achieved by adding the annotated entities from the NPDR phenotype synonyms, disease abbreviations and disease synonyms lexica to the other two lexica mentioned above.

The evaluation of the classifier was done, firstly, by automatically verifying which of the extracted relations was referenced in the HPO gold standard knowledge base. If the relation was marked as *True* and existed in the knowledge base, the classification was correct. Since this approach assumes a closed-world assumption, i.e., every unknown relation that is not contained within the knowledge base is false, a manual evaluation was achieved by randomly selecting a sample from the total extracted relations, to confirm if the identified relation existed in the article from which it was annotated. The sample size (n_0) for each method was determined using the Cochran [Cochran, 1977] formula with a confidence level of 95%:

$$n_0 = \frac{Z^2 pq}{e^2} \tag{3.1}$$

where *e* is the margin of error, *p* is the estimated proportion of the population which holds the attribute at issue (in this case the considered value was 0.5), *q* is *1-p*, and the *Z-score* for a confidence level of

[4] http://purl.obolibrary.org/obo/hp.owl

95% is 1.96. Lastly, the accuracy for the *True* relations was calculated in order to determine how many correct relations were detected:

$$Accuracy = \frac{True\ Relations}{Total\ Relations} \times 100 \tag{3.2}$$

Example 3.2 Relation extraction.

- **Article PMID:** 32163377
- **Sentence:** *Hemophagocytic lymphohistiocytosis HLH was suspected, but* ***no*** *evidence of* ***hemophagocytosis*** *was found in the bone marrow*
- **Disease:** *mycobacterial disease*
- **Disease Identifier:** 618963
- **Phenotype:** *hemophagocytosis*
- **Phenotype Identifier:** HP:0012156
- **Relation: True**

3.3 Results and Discussion

3.3.1 NPDR Dataset

The main results of the statistical analysis regarding the NPDR dataset are shown in Table 3.6.

Regarding **disease entities**, the statistical analysis shows that 73.2% of the articles contain a disease synonym, consequently, in the other 26.8%, an exact OMIM disease term could be found. For disease abbreviations, 16.7% of the articles contained an exact match with the official OMIM disease abbreviation, but only 3.2% of the gene or locus names matched the official disease abbreviation. It is also worth mentioning that in 79% of the articles the disease could be identified by its associated OMIM gene or locus term. Furthermore, in 59.6% of the articles that included a disease synonym, instead of an OMIM designation, the associated gene or locus term could be found. Thus, using gene terms related to monogenic diseases, as an alternative disease designation, would greatly improve the success of the NER task. This is especially relevant in articles that use disease names that refer to a broader disease term. For example, article PMID:23434115 is about a heterogeneous group of inherited disorders called *Congenital Macrothrombocytopenia* (MIM 155100), but the negative relation regards ACTN1-mutated individuals. Consequently, the entity that participates in the negative relation is *MACROTHROMBOCYTOPENIA, AUTOSOMAL DOMINANT, ACTN1-RELATED* (MIM 615193). In this case, the disease

Table 3.6: Statistics of the NPDR dataset categories.

Categories	Number of records	Percentage (%)
total relations	347	100
HPO dataset relations	222	64
new relations	125	36
disease_synonym	254	73.2
exact match disease_name with OMIM designation	93	**26.8**
hp_synonym	198	57.1
exact match hpo_name	148	**42.6**
disease_abbreviation	200	57.6
no disease_abbreviation	147	42.4
exact match disease_abbreviation with OMIM abbreviation	58	**16.7**
associated_gene_or_locus	274	79
no associated_gene_or_locus	73	21
exact match associated_gene_or_locus with OMIM abbreviation	11	**3.2**
disease_id_in_text	42	12.1
no disease_id_in_text	305	**87.9**
gene_id_in_text	121	34.9
no gene_id_in_text	226	**65.1**
sentences	285	82.1
negation	198	57.1
not_abnormal	87	25.1
co_occurrence	50	14.4
no co_occurrence	297	**85.6**

entity was identified by the gene term. Moreover, the highest percentage of articles contained the gene identifier related to the disease (34.9%), instead of the disease's identifier (12.1%).

Respecting **phenotype entities**, 42.6% of the articles included an exact match between the name of the phenotype found in text and an official HPO designation. Like diseases, phenotype terms can be complex and heterogeneous entities since they can be composed of multiple words. The fact that a higher percentage of phenotype names, when compared to disease names, matched an official designation could be the result of HPO holding numerous synonyms for their main phenotype terms. Nonetheless, some phenotype terms can still be difficult to catch. For example, in article PMID:31175426, the identified HPO term is *Global developmental delay*, and even though the HPO holds over twenty synonyms for this term, none of them matched the phenotype name found in text, which is *gross motor development* (Example 3.3.1). It is also worth mentioning that contrary to disease and gene terms, very few documents contained phenotype identifiers. The absence of identifiers in biomedical texts is another constraint for the annotation of entities, especially for ones with complex designations, and makes it difficult to distinguish between phenotypes and diseases terms. For example, the term *ocular albinism* can be linked to the HPO identifier HP:0001107 and identifier MIM 300500. During the development of the NPDR dataset this problem did not arise, since the phenotypes were only annotated from negative relation sentences, but this situation should be taken into consideration for further annotations.

Example 3.3.1 Example of high heterogeneity in phenotype entities.

- **Article PMID:** 11706389
- **Identified Phenotype Entity:** *Gross motor development*
- **Phenotype Identifier:** HP:0001263
- **HPO Name:** *Global developmental delay*
- **HPO Synonyms:** *Cognitive delay; Delayed cognitive development; Delayed development; Delayed developmental milestones; Delayed intellectual development; Delayed milestones; Delayed psychomotor development; Developmental delay; Developmental delay in early childhood; Developmental delay, global; Developmental retardation; Lack of psychomotor development; Mental and motor retardation; Motor and developmental delay; Psychomotor delay; Psychomotor development deficiency; Psychomotor development failure; Psychomotor developmental delay; Retarded development; Retarded mental development; Retarded psychomotor development*

In 82.13% of the articles a **negative relation sentence** could be found. This corresponds to a total of 285 negative relations that could be detected through a sentence. The other 62 relations that did not have an associated sentence were all identified from tables, by the process exemplified in Table 3.2. It should be noted that from these relations, 41 (68.33%) came from new manual annotations, which was expected since they would be arduous to detect through an automatic process. As for adverbs of negation, they appeared in 57.1% of the articles, making them more frequent than the word *normal*. But when comparing the percentage of individual words of negation and the word *normal* (Table 3.7), the latter was the most frequent token (30.5%), followed by *no* (28.8%). The other most frequent adverbs were *not* (17.2%), *none* (5.2%), *without* (4.5%), *absent* (3.5%) and *negative* (2.8%).

Co-occurrence of entities in the same sentence appeared in only 14.4% of the articles, and when it was observed, the entities that co-occurred more frequently where the disease abbreviation term and the

Table 3.7: Statistics of the NPDR dataset negation tokens.

Words of Negation (including *normal*)	**Number of records**	**Percentage (%)**
absence	8	2.8
absent	10	3.5
denied	1	0.4
excluded	1	0.4
free of	2	0.7
impairment	2	0.7
lack of	2	0.7
N	3	1.0
negative	8	2.8
neg	1	0.4
no	82	**28.8**
none	15	5.2
normal	87	**30.5**
not	46	16.1
NP	1	0.4
present exclusively	2	0.7
rules out	1	0.4
without	13	4.5
total	285	100

disease associated gene or locus (Table 3.8). Hence, co-occurrence should not be used as an approach for extracting negative relations between phenotype and disease entities. The fact that these relations cannot be detected at the sentence level is one of the main constraints for detecting them. Another approach would be to use document-level RE, which is a more arduous task since relations are extracted from multiple sentences.

Concerning the **location** of the sentences and entities in the articles (Table 3.9), the most common location for both entities and sentences was the *body* section. More specifically, the *Discussion* section for diseases (54.8%), the *Introduction* section for genes (50.1%), and the *Results* section for sentences (28.5%). Disease and gene terms were also common in the *abstract* section (80.1% for diseases, and 62.5% for genes), and 20.5% of the negative relations could be inferred by tables (present in the *body* or *Appendices* section). It is also worth mentioning that to be considered a true negative relation, the negative association between the phenotype and the disease can only be inferred after analysing all cases. Thus, sentences located in the *Discussion* section are more likely to imply a true negative relation than

Table 3.8: Statistics of the NPDR dataset negation tokens.

Co-occurrence	Number of records	Percentage (%)
disease_name	4	8
disease_synonym	6	12
disease_abbreviation	20	**40**
associated_gene_or_locus	22	**44**
gene_id	1	2

sentences in the *Materials and Methods* section, since they usually refer to all individuals that were characterised in the paper. For example, in article PMID:33772159, two sentences (located in the *Results* section) can be found containing both a phenotype entity and a token of negation. If a negative relation had been annotated from these sentences, it would have been a false positive relation, since each sentence only applies to the description of one patient (Example 3.3.2). Furthermore, because the true negative relation from this article was implied by a table, which means a negative relation sentence could not be found, the chances of automatically annotating a false relation would have been considerable. Therefore, the occurrence in the same article of several sentences that imply a negative relation could, paradoxically, point to a false negative relation because they would be referring to one patient at a time.

Example 3.3.2 Example of False Negative Relations.

- **Article PMID:** 33772159
- **Sentence 1:** ***Nerve conduction*** *studies and EMG were entirely* ***normal***
- **Sentence 2:** ***Nerve conduction*** *studies showed* ***normal*** *sensory and motor conduction velocities, with low-amplitude motor responses*
- **Disease:** *NEURODEVELOPMENTAL DISORDER WITH HYPOTONIA, NEUROPATHY, AND DEAFNESS; NEDHND*
- **Disease Identifier:** 618963
- **Phenotype:** *abnormal nerve conduction*
- **Phenotype Identifier:** HP:0040129
- **Relation: False**

Table 3.9: Statistics of the NPDR dataset regarding entities and sentences location.

Location	Diseases (%)	Genes (%)	Sentences (%)
title	68.6	54.5	0
abstract	80.1	62.5	6.6
body	**93.9**	**78.1**	**90.5**
Introduction	49.9	**50.1**	1.4
Materials and Methods	36.3	47	12.7
Results	40.9	46.7	**28.5**
Discussion	**54.8**	47.8	9.5
Tables	7.2	10.7	20.5
Appendices	N.A	N.A.	2.9

3.3.2 Automatic Extraction

The final number of phenotype and disease annotations for each corpus are shown in Table 3.10. From these tables it is possible to see that by using the NPDR dataset as a supplementary lexicon, the final number of disease annotations increased by 15.1% for the XML Corpus, and 21.2% for the PDF Corpus. Regarding phenotype annotations, 23% of the total annotations were added by the NPDR lexicon for the XML Corpus, and 23.8% for the PDF Corpus. It is also worth mentioning that the OMIM lexicon yielded the larger number of annotations for both gene and disease entities, but since disease terms are more complex and heterogeneous than genes, 74.6% of the disease annotations for the XML Corpus, and 91.0% for the PDF Corpus, did not have an identifier and, consequently, were discarded for the RE task. Therefore, from the initial 819 disease annotations from the OMIM lexicon for the XML Corpus, and 7430 annotations for the PDF Corpus, only 208 and 668 were used for the XML Corpus and PDF Corpus, respectively (Table 3.11). Even though the final number of disease annotations used for this task decreased significantly, most diseases could still be linked to their correspondent identifier through their associated genes (Example 3.2.2).

Regarding the annotation of negative sentences, the total number of phrases for the XML Corpus was 1106, and for the PDF Corpus was 5945. Also, since most sentences showed a regular pattern such as *phenotype entity followed by negative expression* (e.g., ***Ectopia lentis*** *was* ***not*** *observed*), the regular expression algorithm was able to detect every negative relation sentence described in the NPDR dataset.

Example 3.3.2 Example of missed NER for a disease term, but correct NEL through its associated gene term.

- **Article PMID:** 25038750
- **Annotated Disease Entity:** *autoimmune disease*

Table 3.10: The final number of disease and gene annotations from the OMIM lexicon, disease abbreviations and synonyms and phenotype synonyms from the NPDR lexicon, phenotype annotations from the HPO lexicon, and total number of annotations for each corpus.

Lexicon	**Entity**	**Number of Annotations**		**Total Annotations**	
		XML Corpus	**PDF Corpus**	**XML Corpus**	**PDF Corpus**
OMIM	gene	7873	41381	8692	48811
	disease with id	208	668		
	disease without id	611	6762		
NPDR	disease abbreviation	433	2502	1713	14550
	disease synonym	1113	10603		
	phenotype synonym	167	1445		
HPO	phenotype	560	4619	560	4619

- **Missed Disease Entity:** *autoimmune disease, multisystem, infantile-onset, 1*
- **Annotated Gene Entity:** *STAT3*
- **Disease and Gene Identifier:** 615952

From the NPDR dataset it was possible to observe that only 7.8% of the articles contained simultaneously an exact match between the entities' terms found in text and their correspondent official designation, and a negative relation sentence. Therefore, it was not expected to find every relation present in every document that served as an input for the automatic extraction system. So, to improve the RE task, the NER and NEL tasks also had to be ameliorated. From Table 3.11, it is possible to observe that the addition of the synonyms lexica from the NPDR dataset led to the extraction of more relations. The total added relations was **124** for the XML Corpus and **545** for the PDF Corpus, when comparing the Labels Method to the AllLabels Method. These added relations also translated in an improvement of the accuracy for the *True* relations, as is shown in Table 3.12. Also, the XML Corpus achieved better results, which can be explained by the fact that articles from the corpus were retrieved automatically, in a correct XML format. Therefore, the NER, NEL and RE tasks were performed on a more regular text, whereas in the PDF Corpus some documents did not retain the original article's format, since they were converted from PDF to text, which compromised the annotation of some entities and relations.

During the manual evaluation of the extracted relations from the four samples (corresponding to each method and corpus), it was possible to observe that every disease annotation that appeared in a *True* relation came from the associated gene or locus. This was expected, since most disease terms could not be annotated during the NER phase, as described above. This poses two main problems for the extraction of relations. Firstly, some genes are shared among multiple diseases, which can lead to the extraction

Table 3.11: The final number of disease and gene annotations from the OMIM lexicon, disease abbreviations, disease synonyms and phenotype synonyms from the NPDR lexicon, phenotype annotations from the HPO lexicon, True and False relations, and total number of relations for each Method and Corpus.

Method	Corpus	Annotations			Relations		
		Gene	Disease	Phenotype	True	False	Total
Labels	XML	7873	208	560	8	694	702
	PDF	41381	668	4619	31	3695	3726
AllLabels	XML	7873	1754	727	25	811	826
	PDF	41381	13773	6064	74	4197	4271

Table 3.12: The number of True and False relations detected, the total relations and the accuracy for the True relations, for the automatic and manual evaluation for each Method and Corpus.

Method	Corpus	Automatic Evaluation				Manual Evaluation			
		True	False	Total	Accuracy	True	False	Total	Accuracy
Labels	XML	56	646	702	7.98	10	238	248	4.03
	PDF	28	3698	3726	0.75	6	343	349	1.72
AllLabels	XML	73	753	826	8.84	17	246	263	6.46
	PDF	125	4146	4271	2.93	10	343	353	2.83

of false negative relations, since the wrong disease identifier can be attributed from the incorrectly associated gene. For example, the gene *FIG4* is associated to *AMYOTROPHIC LATERAL SCLEROSIS 11; ALS11* (MIM 612577), as well as *CHARCOT-MARIE-TOOTH DISEASE, TYPE 4J; CMT4J* (MIM 611228). Secondly, if the associated gene is not referenced in the article, it cannot be annotated. For instance, in article PMID:25038750, the annotation of the disease *PREKALLIKREIN DEFICIENCY; PKK DEFICIENCY; FLETCHER FACTOR DEFICIENCY* (MIM 612423) was missed because of the complexity of the term, and although this entity possesses an associated gene term (*KLKB1*), the later was not referenced in text and, consequently, the correct disease identifier was also not annotated (Example 3.3.3).

Overall, the results obtained from the evaluation show that it is possible to extract negative relations from text using the HPO gold standard knowledge base, but also that many of them are missed, mainly due to problems with the NER task.

Example 3.3.3 Example of missed disease annotation due to absent gene entity in article.

- **Article PMID:** 25038750
- **Disease Entity:** *PREKALLIKREIN DEFICIENCY; PKK DEFICIENCY; FLETCHER FACTOR*

DEFICIENCY

- **Annotated Disease Term:** *kallikrein*
- **Missed Disease Identifier:** *612423*
- **Absent Gene Entity:** *KLKB1*

Chapter 4

Conclusion

To this day, scientific literature continues to be the prevailing source of biomedical knowledge. This knowledge comprises the numerous relations between individual genes, proteins, or chemicals, which constitute whole biological systems. These relations can either be classified as positive, if there is an association between entities, or negative, if no association was found. Negative relations also carry scientific significance by holding information about disproven relations, which limit the search space for researchers and clinicians. Yet, most biomedical relations databases still focus on only annotating positive relations, overlooking the knowledge encoded in negated findings. Moreover, the lack of high-quality knowledge bases harbouring annotated negative relations, hinders the development of text mining systems capable of automatically annotating new negative relations. Even though the Human Phenotype Ontology (HPO) made available a gold standard-knowledge base of negative phenotype-disease relations, the information it yields is not sufficient as training data for biomedical text mining models. Therefore, this work made an important contribution in the understanding of biomedical negative relations present in text, by creating a dataset that thoroughly characterizes negative phenotype-disease relations from biomedical literature, the Negative Phenotype-Disease Relations (NPDR) dataset.

The initial hypothesis of this thesis was that it is possible to produce a knowledge base of negative relations from the scientific literature. To this date, the NPDR dataset is the only resource available that contains 347 manually annotated at the document-level negative relations, from which 222 derive from the HPO knowledge base, and 125 are new annotated relations.

For the creation of the dataset, a subset of 237 relations from the HPO knowledge-base was linked to to 169 PubMed IDs (PMID), which led to the retrieval of 177 biomedical documents (169 articles and clinical reports, seven supplementary materials documents and one book chapter). These documents were manually examined to study the negative relations present in each article. The NPDR dataset comprises three main categories: the characterization of the entities present in the negative relation, which englobes the analysis of phenotype entities, disease entities and their associated genes; the characterization of the sentence containing the negative relation, which comprises the description of the token of negation used in the sentence and the co-occurrence of the entities; and the characterization of the location of the entities and sentences in the article.

Even though negation occurs frequently in scientific literature, detecting negated events is still an arduous task. Some results from the dataset can explain this situation. Firstly, 73.2% of disease entities found in text were described through a non-official synonym, and for phenotype entities this happened in 55.9% of the articles. Secondly, 16.87% of the articles did not contain a negative relation sentence, and when a negative phrase was found, co-occurrence of the entities appeared in only 14.4% of the sentences. Thirdly, 90.5% of the sentences were present in the article's body. Whereas disease entities could be identified in the abstract, the full-text was needed in order to detect the negative relation and phenotype entities.

To evaluate the impact of the NPDR dataset on the Named-Entity Recognition (NER), Named-Entity Linking (NEL) and Relation Extraction (RE) tasks, an automatic system that performed these tasks, and detected negative sentences, was developed. This system used as input the corpus of documents retrieved for the creation of the NPDR dataset. Ideally, to enable an automatic process to perform the NER task, an article should include simultaneously an exact match for both phenotype and disease terms with an official HPO and OMIM designation, respectively, which happened in 12.4% of the articles. Therefore, to improve the NER task, three lexica of disease abbreviations, disease synonyms and phenotypes synonyms from the NPDR dataset were built. To the annotations from the OMIM and HPO lexica, it was possible to increase by 18.15% and 23.4% the number of disease and phenotype entities annotated, respectively. The added annotations also translated in an increase of 15.11% of relations extracted. Even though every negative relation sentence described in the NPDR dataset was successfully captured by the system, the RE task yielded poor results, with the highest accuracy only reaching 8.84%.

The main contribution of this work, the NPDR dataset, as well as the code for the automatic system, are available on GitHub (https://github.com/lasigeBioTM/NPDR).

4.1 Future Work

The results from the NPDR dataset showed that negative relations from biomedical literature offer certain constraints, making them arduous to detect through automatic systems. Specifically, during the NER task many entities were missed due to the complexity and heterogeneity of disease and phenotype terms. Integrating Machine-Learning (ML) methods into the NER task could greatly improve the chances of correctly identifying the entities. For example, a tool such as the Identifying Human Phenotypes (IHP) system [Lobo et al., 2017], which combine a ML approach with a dictionary-based and manual rules-based method, is specifically designed to detect HPO entities in unstructured text. For the RE task, a document-level relation extraction method could be applied, by using, for example, the document-level entity mask method with type information (DEMMT) [Han and Wang, 2020]. This method is BERT-based model that can predict relations between entities at the document-level. The improvements of these tasks, combined with the knowledge encoded in the NPDR dataset, could be used to create a corpus of negative phenotype-disease relations, to improve negative relation extraction systems. Finally, future work can include manually expanding the dataset, by exploring more negative relations from the HPO gold-standard knowledge base.

References

Agarwal, S., Yu, H., and Kohane, I. (2011). BioN∅t: A searchable database of biomedical negated sentences. *BMC Bioinformatics*, 12(1):420. 17

Aronow, D. B., Fangfang, F., and Croft, W. B. (1999). Ad hoc classification of radiology reports. *Journal of the American Medical Informatics Association*, 6(5):393–411. 12

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29. 3

Bach, N. and Badaskar, S. (2007). A survey on relation extraction. Technical report, Language Technologies Institute, Carnegie Mellon University. 10

Beltagy, I., Cohan, A., and Lo, K. (2019). Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676. 11

Blohm, P., Frishman, G., Pawel, S., Goebels, F., Wachinger, B., Ruepp, A., and Frishman, D. (2013). Negatome 2.0: A database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Research*, 42. 2, 10, 17

Bokharaeian, B., Díaz, A., Neves, M., and Francisco, V. (2014). Exploring negation annotations in the drugddi corpus. 15

Bokharaeian, B., Diaz, A., Taghizadeh, N., Chitsaz, H., and Chavoshinejad, R. (2017). SNPPhenA: a corpus for extracting ranked associations of single-nucleotide polymorphisms and phenotypes from literature. 8(1). 16

Bravo, À., Piñero, J., Queralt-Rosinach, N., Rautschka, M., and Furlong, L. I. (2015). Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16(1). 2, 10

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310. 12

Cochran, W. G. (1977). *Sampling Techniques, 3rd Edition*. John Wiley. 27

Couto, F. M. and Lamurias, A. (2018). MER: a shell script and annotation server for minimal named entity recognition and linking. *Journal of Cheminformatics*, 10(1). 5, 26

Cruz Díaz, N. P. and Maña López, M. (2015). An analysis of biomedical tokenization: Problems and strategies. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 40–49, Lisbon, Portugal. Association for Computational Linguistics. 8

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. 11

Doğan, R. I., Leaman, R., and Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10. 15

Elazhary, H. (2017). NegMiner: An automated tool for mining negations from electronic narrative medical documents. *International Journal of Intelligent Systems and Applications*, 9(4):14–22. 13

Federico, A. and Monti, S. (2021). Contextualized protein-protein interactions. *Patterns*, 2(1):100153. 1

Friedman, C., Kra, P., and Rzhetsky, A. (2002). Two biomedical sublanguages: A description based on the theories of zellig harris. *Journal of Biomedical Informatics*, 35(4):222–235. 7

Gindl, S., Kaiser, K., and Miksch, S. (2008). Syntactical negation detection in clinical practice guidelines. *Studies in Health Technology and Informatics*, 136:187–192. 12

Giuliano, C., Lavelli, A., and Romano, L. (2006a). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 401–408, Trento, Italy. Association for Computational Linguistics. 2

Giuliano, C., Lavelli, A., and Romano, L. (2006b). Exploiting shallow linguistic information for relation extraction from biomedical literature. In McCarthy, D. and Wintner, S., editors, *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. The Association for Computer Linguistics. 10

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220. 13

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. ArXiv. 11

Gutiérrez-Sacristán, A., Grosdidier, S., Valverde, O., Torrens, M., Bravo, À., Piñero, J., Sanz, F., and Furlong, L. I. (2015). PsyGeNET: a knowledge platform on psychiatric disorders and their genes: Table 1. *Bioinformatics*, 31(18):3075–3077. 15

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl_1):D514–D517. 17

Han, X. and Wang, L. (2020). A novel document-level relation extraction method based on bert and entity information. *IEEE Access*, 8:96912–96919. 9, 11, 38

Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., and Declerck, T. (2013). The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. 46(5):914–920. 15

Hotho, A., Nurnberger, A., and Paass, G. (2005). A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20:19–62. 7

Hou, W.-J., , and Ceesay, B. (2021). Exploring the adaptation of recurrent neural network approaches for extracting drug–drug interactions from biomedical text. *International Journal of Machine Learning and Computing*, 11(4):267–273. 1

Jensen, L. J., Saric, J., and Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, 7(2):119–129. 1

Jurafsky, D. and Martin, J. H. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, Upper Saddle River, N.J, 2nd ed edition. OCLC: 213375806. 8

Khandelwal, A. and Sawant, S. (2020). NegBERT: A transfer learning approach for negation detection and scope resolution. In *LREC*. 13

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA Corpus—A semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182. 14

Köhler, S., Carmody, L., Vasilevsky, N., and et al., J. O. B. J. (2018a). Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, 47(D1):D1018–D1027. 22

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N. L., Matentzoglu, N., McMurry, J. A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J. P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., Segal, M., Jansen, A. C., Muaz, A., Chang, W. H., Bergerson, J., Laulederkind, S. J. F., Yüksel, Z., Beltran, S., Freeman, A. F., Sergouniotis, P. I., Durkin, D., Storm, A. L., Hanauer, M., Brudno, M., Bello, S. M., Sincan, M., Rageth, K., Wheeler, M. T., Oegema, R., Lourghi, H., Rocca, M. G. D., Thompson, R., Castellanos, F., Priest, J., Cunningham-Rundles, C., Hegde, A., Lovering, R. C., Hajek, C., Olry, A., Notarangelo, L., Similuk, M., Zhang, X. A., Gómez-Andrés, D., Lochmüller, H., Dollfus, H., Rosenzweig, S., Marwaha, S., Rath, A., Sullivan, K., Smith, C., Milner, J. D., Leroux, D., Boerkoel, C. F., Klion, A., Carter, M. C., Groza, T., Smedley, D., Haendel, M. A., Mungall, C., and Robinson, P. N. (2018b). Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, 47(D1):D1018–D1027. 3, 14, 17

Köhler, S., Doelken, S. C., Mungall, C. J., and et al., S. B. (2013). The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(D1):D966–D974. 23

Lamurias, A., Clarke, L. A., and Couto, F. M. (2017). Extracting microRNA-gene relations from biomedical literature using distant supervision. *PLoS ONE*, 12(3):e0171929. 11

Lamurias, A. and Couto, F. M. (2019). Text mining for bioinformatics using biomedical literature. In *Encyclopedia of Bioinformatics and Computational Biology*, pages 602–611. Elsevier BV. 3, 7, 9

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. 11

Lee, K., Lee, S., Park, S., Kim, S., Kim, S., Choi, K., Tan, A. C., and Kang, J. (2016). BRONCO: Biomedical entity relation ONcology COrpus for extracting gene-variant-disease-drug relations. *Database*, 2016. 15

Legrand, J., Gogdemir, R., Bousquet, C., Dalleau, K., Devignes, M.-D., Digan, W., Lee, C.-J., Ndiaye, N.-C., Petitpain, N., Ringot, P., Smail, M., Toussaint, Y., and Coulet, A. (2020). Pgxcorpus, a manually annotated corpus for pharmacogenomics. *Scientific Data*, 7:3. 16

Leng, J. and Jiang, P. (2016). A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm. *Knowledge-Based Systems*, 100:188–199. 10

Lobo, M., Lamurias, A., and Couto, F. M. (2017). Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, 2017:1–8. 38

Mallory, E. K., Zhang, C., Ré, C., and Altman, R. B. (2015). Large-scale extraction of gene interactions from full-text literature using DeepDive. *Bioinformatics*, page btv476. 11

Manning, C. D. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA. 7

Mehrabi, S., Krishnan, A., Sohn, S., Roch, A. M., Schmidt, H., Kesterson, J., Beesley, C., Dexter, P., Schmidt, C. M., Liu, H., and Palakal, M. (2015). DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *Journal of Biomedical Informatics*, 54:213–219. 13

Müller, H.-M., Kenny, E. E., and Sternberg, P. W. (2004). Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*, 2(11):e309. 10

Mutalik, P. G., Deshpande, A., and Nadkarni, P. M. (2001). Use of general-purpose negation detection to augment concept indexing of medical documents: A quantitative study using the UMLS. *Journal of the American Medical Informatics Association*, 8(6):598–609. 12

Muzaffar, A. W., Azam, F., and Qamar, U. (2015). A relation extraction framework for biomedical text using hybrid feature set. *Computational and Mathematical Methods in Medicine*, 2015:1–12. 9

Nunzio, G. M. D. and Vezzani, F. (2018). A linguistic failure analysis of classification of medical publications: A study on stemming vs lemmatization. In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, pages 182–186. Accademia University Press. 8

Payne, T. E. (1997). *Describing morphosyntax: A guide for field linguists*. Cambridge University Press, Cambridge. 12

Pinero, J., Bravo, A., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., Garcia-Garcia, J., Sanz, F., and Furlong, L. I. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1):D833–D839. 16

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., and Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1). 15

Rath, A., Olry, A., Dhombres, F., Brandt, M. M., Urbero, B., and Ayme, S. (2012). Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Human Mutation*, 33(5):803–808. 17

Robinson, P. and Bauer, S. (2011). *Introduction to Bio-Ontologies*. Chapman & Hall/CRC Mathematical and Computational Biology. CRC Press. 3

Rosario, B. and Hearst, M. A. (2004). Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 430–es, USA. Association for Computational Linguistics. 15

Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., Bisordi, K., Campion, N., Hyman, B., Kurland, D., Oates, C. P., Kibbey, S., Sreekumar, P., Le, C., Giglio, M., and Greene, C. (2018). Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Research*, 47(D1):D955–D962. 3, 14

Segura-Bedmar, I., Martínez, P., and de Pablo-Sánchez, C. (2011). Using a shallow linguistic kernel for drug–drug interaction extraction. *Journal of Biomedical Informatics*, 44(5):789–804. 10

Sohn, S., Wu, S., and Chute, C. G. (2012). Dependency parser-based negation detection in clinical narratives. *AMIA Joint Summits on Translational Science Proceedings*, 2012(1-8). 13

Sousa, D. and Couto, F. (2020). BiOnt: Deep learning using multiple biomedical ontologies for relation extraction. In *42nd European Conference on Information Retrieval (ECIR 2020)*, volume 12036. 11

Sousa, D., Lamurias, A., and Couto, F. M. (2019). A silver standard corpus of human phenotype-gene relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1487–1492, Minneapolis, Minnesota. Association for Computational Linguistics. 1, 16

Sousa, D., Lamurias, A., and Couto, F. M. (2020). Improving accessibility and distinction between negative results in biomedical relation extraction. *Genomics & Informatics*, 18(2):e20. 1, 3

van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., Kors, J. A., and Furlong, L. I. (2012). The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, 45(5):879–884. 15

Vincze, V., Szarvas, G., Farkas, R., Móra, G., and Csirik, J. (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S11). 2, 15

Xue, L., Qing, S., and Pengzhou, Z. (2018). Relation extraction based on deep learning. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*. IEEE. 10

Zhou, D., Zhong, D., and He, Y. (2014). Biomedical relation extraction: From binary to complex. 2014:1–18. 2

Zhou, K., Wang, Y., Cohen, K. B., Kim, J.-D., Ma, X., Shen, Z., Meng, X., and Xia, J. (2021). Bridging heterogeneous mutation data to enhance disease gene discovery. 22(5). 1