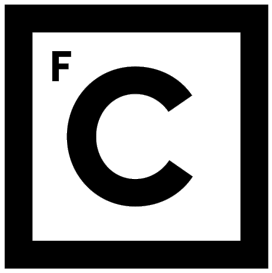
UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS



Ciências
ULisboa

Development of a Text Mining Approach to Disease Network Discovery

"Documento Definitivo"

Doutoramento em Biologia
Especialidade de Biologia de Sistemas

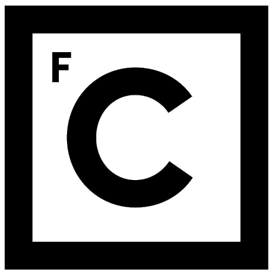
André Francisco Martins Lamúrias

Tese orientada por:
Prof. Doutor Francisco José Moreira Couto e Prof. Doutor Luka Alexander Clarke

Documento especialmente elaborado para obtenção do grau de doutor

2019

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS



Ciências
ULisboa

Development of a Text Mining Approach to Disease Network Discovery

Doutoramento em Biologia
Especialidade de Biologia de Sistemas

André Francisco Martins Lamúrias

Tese orientada por:
Prof. Doutor Francisco José Moreira Couto e Prof. Doutor Luka Alexander Clarke

Júri
Presidente:

Doutor Rui Manuel dos Santos Malhó, Professor Catedrático e Presidente do Departamento de Biologia Vegetal da Faculdade de Ciências da Universidade de Lisboa.

Vogais:

- Doutor João Miguel da Costa Magalhães, Professor Associado, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa;
- Doutor José Luís Guimarães Oliveira, Professor Associado com Agregação, Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro;
- Doutor João André Nogueira Custódio Carriço, Investigador Auxiliar, Faculdade de Medicina da Universidade de Lisboa;
- Doutor Francisco José Moreira Couto, Professor Associado, Faculdade de Ciências da Universidade de Lisboa (orientador);
- Doutor Ricardo Pedro Moreira Dias, Investigador Auxiliar Convidado, Faculdade de Ciências da Universidade de Lisboa.

Financiado pela Fundação para a Ciência e Tecnologia (SFRH/BD/106083/2015)

Documento especialmente elaborado para obtenção do grau de doutor

2019

Acknowledgements

- My supervisors: Prof. Francisco Couto for giving me the opportunity to work with him in this area and trusting in me; Prof. Luka Clarke for overall guidance;
- Mario David and João Pina from INCD, and João Carriço from IMM for giving me access to computing clusters that made it possible to run my analysis;
- FCT and the BioSYS PhD programme;
- my colleagues from LASIGE and BioSYS, former and current: those that contributed directly to my work, and those that contributed indirectly through informal discussions, help with technical issues, and socializing. I also want to thank the support staff of both institutions for helping with the bureaucratic issues;
- my family, who were always there for me.

Abstract

Scientific literature is one of the major sources of knowledge for systems biology, in the form of papers, patents and other types of written reports. Text mining methods aim at automatically extracting relevant information from the literature. The hypothesis of this thesis was that biological systems could be elucidated by the development of text mining solutions that can automatically extract relevant information from documents. The first objective consisted in developing software components to recognize biomedical entities in text, which is the first step to generate a network about a biological system. To this end, a machine learning solution was developed, which can be trained for specific biological entities using an annotated dataset, obtaining high-quality results. Additionally, a rule-based solution was developed, which can be easily adapted to various types of entities.

The second objective consisted in developing an automatic approach to link the recognized entities to a reference knowledge base. A solution based on the PageRank algorithm was developed in order to match the entities to the concepts that most contribute to the overall coherence.

The third objective consisted in automatically extracting relations between entities, to generate knowledge graphs about biological systems. Due to the lack of annotated datasets available for this task, distant supervision was employed to train a relation classifier on a corpus of documents and a knowledge base. The applicability of this approach was demonstrated in two case studies: microRNA-gene relations for cystic fibrosis, obtaining a network of 27 relations using the abstracts of 51 recently published papers; and cell-cytokine relations for tolerogenic cell therapies, obtaining a network of 647 relations from 3264 abstracts. Through a manual evaluation, the information contained in these networks was determined to be relevant. Additionally, a solution combining deep learning techniques with ontology information was developed, to take advantage of the domain knowledge provided by ontologies.

This thesis contributed with several solutions that demonstrate the usefulness of text mining methods to systems biology by extracting domain-specific information from the literature. These solutions make it easier to integrate various areas of research, leading to a better understanding of biological systems.

Keywords: Text Mining; Information Extraction; Systems Biology; Machine Learning

Resumo

O estudo de sistemas biológicos é uma tarefa de elevada dificuldade devido à complexidade dos seus componentes e mecanismos. A biologia de sistemas estuda as diferentes componentes de um sistema e como elas interagem como um todo, em vez de focar individualmente em cada componente. A compreensão destes sistemas biológicos pode levar ao desenvolvimento de modelos de previsão de processos metabólicos, de forma a melhorar a descoberta de novos fármacos e medicina personalizada. Doenças humanas são sistemas biológicos com alto interesse para a comunidade científica. Ao combinar o conhecimento sobre várias doenças humanas, podemos gerar uma rede de doenças. Estas redes são úteis para compreender os mecanismos moleculares comuns a mais do que uma doença.

A literatura é uma das maiores fontes de conhecimento biomédico atuais, na forma de artigos, patentes e outros tipos de relatórios. Para elucidar um sistema biológico, é necessário integrar vários estudos, sendo esta uma tarefa dispendiosa, devido à quantidade cada vez maior de artigos publicados. Uma possível abordagem para resolver este problema é através de prospeção de texto. Métodos de prospeção de texto visam extrair automaticamente informação relevante da literatura. Este métodos conseguem identificar entidades mencionadas no texto, bem como relações descritas entre estas entidades. A informação extraída pode ser depois usada para responder a questões do utilizador, para melhorar a curação de bases de dados e para contruir grafos de conhecimento

Ontologias são usadas para organizar o conhecimento sobre um determinado domínio, sendo estas usadas frequentemente pela comunidade científica. Uma ontologia pode ser definida por um vocabulário de termos e uma especificação formal dos seus significados. Mais recentemente, a comunidade científica tem dado destaque a grafos de conhecimento, que permitem ligar diversas fontes de informação e estabelecer ligações com significado, podendo ser usadas para responder a perguntas do utilizador. Estes grafos são úteis para biologia de sistemas devido à quantidade de conceitos e mecanismos envolvidos em sistemas biológicos e à necessidade de organizar este conhecimento.

Devido aos vários aspectos que podem ser explorados na prospeção de texto, várias tarefas foram estabelecidas. Uma destas tarefas é o reconhecimento de entidades, que consiste em identificar as entidades mencionadas relevantes num dado texto. Nos sistemas biológicos, estas entidades podem ser genes, proteínas, doenças, fenótipos ou farmácos, por exemplo. O desafio desta tarefa é ter em

conta a grande variabilidade de algumas nomenclaturas e da linguagem escrita, bem como o facto de haver sempre novas entidades, e, por isso, os vocabulários nunca estarem totalmente atualizados. Outra tarefa, que é geralmente sequencial à anterior, consiste em corresponder as entidades extraídas a uma referência externa, de modo a ligar diversas fontes de informação. Neste caso há que ter em conta os vários sinónimos e acrónimos que podem ser usados para mencionar conceitos. Finalmente, outra tarefa consiste em extrair relações descritas no texto entre entidades. O objetivo é classificar se cada par de entidades representa uma relação ou não, e, se assim for, de que tipo.

Abordagens automáticas para cada uma destas tarefas podem ser avaliadas individualmente usando um conjunto de teste, com documentos anotados manualmente. No caso do reconhecimento de entidades, as anotações representam a localização das entidades no texto. No caso da ligação de entidades, as anotações são o conceito que representa cada entidade, e é avaliado se o conceito a que cada entidade foi ligado é o mais correto. Finalmente, para extração de relação, as anotações são as relações descritas nos documentos, sendo estas depois comparadas com as relações extraídas automaticamente. O sucesso dos métodos de prospeção de texto nestas tarefas depende do domínio ao qual são aplicadas. Considerando que a linguagem científica é em geral mais complexa de compreender do que outros domínios, por isso os resultados também tendem a ser inferiores.

A hipótese desta tese foi se é possível elucidar sistemas biológicos através de soluções de prospeção de texto que fazem extração automática de informação de documentos. De forma a testar esta hipótese, três objetivos foram estabelecidos, tendo em conta as tarefas previamente mencionadas.

O primeiro objetivo consistiu em desenvolver soluções de prospeção de texto para o reconhecimento de entidades biomédicas em texto, que é o primeiro passo para gerar uma rede sobre sistemas biológicos. Para isso, uma abordagem baseada em aprendizagem automática foi explorada, a qual pode ser treinada para entidades biológicas específicas, obtendo resultados com elevada qualidade. Além disso, uma abordagem baseada em regras foi explorada, a qual pode ser adaptada para vários tipos de entidades. Estas abordagens foram comparadas uma com a outra, bem como com o estado da arte. A primeira obteve resultados com elevada qualidade e a segunda resultados num tempo consideravelmente curto. Tendo em conta as vantagens e desvantagens de cada uma destas abordagens, é possível usar uma ou outra conforme as necessidades do utilizador.

O segundo objetivo consistiu no desenvolvimento de uma abordagem automática para ligar as entidades reconhecidas a uma base de conhecimento de referência. Desta forma, a rede gerada pode ser melhorada com recurso a fontes de infor-

mação externas, pois variações dos termos usados para os mesmos conceitos são combinados num só. Um método baseado no algoritmo *PageRank* foi desenvolvido, de forma a fazer corresponder as entidades aos conceitos que mais contribuem para a coerência global. Para isso, foi estabelecida uma medida de coerência que tem em conta o conteúdo de informação de cada conceito, bem como a semelhança semântica entre conceitos, calculada na ontologia. Combinando estes fatores, esta medida de coerência obtém resultados superiores na ligação de entidades nos dois domínios explorados (químicos com interesse biológico e fenótipos humanos)

O terceiro objetivo consistiu em extrair automaticamente relações entre entidades, podendo assim ser usadas para criar uma rede de informação. Devido à pouca disponibilidade de conjuntos de dados anotados disponíveis para esta tarefa, foi usada supervisão distante para treinar um classificador de relações com um conjunto de documentos e base de conhecimento. As vantagens desta abordagem residem nos seguintes fatores: não é necessária a anotação manual de documentos com relações; e pode ser adaptada a vários domínios usando documentos e uma base de conhecimento apropriada. A aplicabilidade desta abordagem foi demonstrada por aplicação à extração de relações entre genes e microRNAs, e entre células e citocinas. No primeiro caso, foi gerado um grafo de conhecimento para um conjunto de documentos sobre fibrose quística, obtendo vários microRNAs que regulam a atividade de genes relacionados com essa doença. Foi também gerado um grafo de conhecimento para obter mais informação útil para terapias celulares tolerogênicas. Neste grafo foi comparada a informação obtida sobre células dendríticas e antigênicas com uma base de dados existente, tendo sido possível obter mais citocinas relacionadas com estas células.

Foi também desenvolvido um método de extração de relações entre duas entidades que combina técnicas de aprendizagem profunda com informação de ontologias, de forma a aproveitar o conhecimento de domínio incorporado nas ontologias. Este método tem em conta os ascendentes das duas entidades para classificar a relação, tendo obtido resultados positivos na extração de relações entre fármacos e entre fenótipos e genes.

Esta tese apresenta várias soluções que demonstram a utilidade de métodos de prospeção de texto para biologia de sistemas, através da extração de informação específica para o domínio, usando a literatura. As abordagens apresentadas facilitam a integração de várias áreas de investigação e levam a uma melhor compreensão de sistemas biológicos. No futuro, estas abordagens podem ser combinadas num único sistema, que possa ser usado para analisar a informação obtida sobre várias doenças.

Palavras Chave: Prospeção de texto; Extração de Informação; Biologia de Sistemas; Aprendizagem Automática

De acordo com o disposto no artigo 24º do Regulamento de Estudos de Pós-Graduação da Universidade de Lisboa, Despacho nº 7024/2017, publicado no Diário da República –2ª Série –nº 155 –11 de Agosto de 2017, foram utilizados nesta dissertação resultados incluídos nos seguintes artigos:

- A. Lamurias and F. Couto. 'Text Mining for Bioinformatics using Biomedical Literature'. In: *Reference Module in Life Sciences (Encyclopedia of Bioinformatics and Computational Biology)*. Vol. 1. Oxford: Elsevier, 2019. DOI: https://doi.org/10.1016/B978-0-12-809633-8.20409-3 (Capítulo 2)
- F. Couto and A. Lamurias. 'Semantic similarity definition'. In: *Reference Module in Life Sciences (Encyclopedia of Bioinformatics and Computational Biology)*. Vol. 1. Oxford: Elsevier, 2019. DOI: https://doi.org/10.1016/B978-0-12-809633-8.20401-9 (Capítulo 3)
- Francisco Couto and Andre Lamurias. 'MER: a Shell Script and Annotation Server for Minimal Named Entity Recognition and Linking'. In: *Journal of Cheminformatics* 10.58 (2018). ISSN: 1758-2946. DOI: https://doi.org/10.1186/s13321-018-0312-9 (Capítulo 4)
- A. Lamurias, L. Clarke and F. Couto. 'Extracting MicroRNA-Gene Relations from Biomedical Literature using Distant Supervision'. In: *PLoS ONE* 12.3 (2017). ISSN: 1932-6203. DOI: https://doi.org/10.1371/journal.pone.0171929 (Capítulo 6)
- A. Lamurias et al. 'Generating a Tolerogenic Cell Therapy Knowledge Graph from Literature'. In: *Frontiers in Immunology* 8.1656 (2017). ISSN: 1664-3224. DOI: https://doi.org/10.3389/fimmu.2017.01656 (Capítulo 7)
- Andre Lamurias et al. 'BO-LSTM: Classifying relations via long short-term memory networks along biomedical ontologies'. In: *BMC Bioinformatics* 20.10 (2019). ISSN: 1471-2105. DOI: https://doi.org/10.1186/s12859-018-2584-5 (Capítulo 8)

No cumprimento do disposto da referida deliberação, a autora esclarece serem da sua responsabilidade, exceto quando referido o contrário, a execução das experiências que permitiram a elaboração dos resultados apresentados, assim como da interpretação e discussão dos mesmos. Os resultados obtidos por outros autores foram incluídos com a autorização dos mesmos para facilitar a compreensão dos trabalhos e estão assinalados nas respetivas figuras e metodologias.

Contents

List of Figures

LIST OF FIGURES

List of Tables

List of Abbreviations

ADR Adverse Drug Reactions
ART Average Response Time
CA Common Ancestors
CE Common Entries
ChEBI Chemical Entities of Biological interest
DCA Disjunctive Common Ancestors
DS Distant Supervision
DDI Drug-drug Interactions
EL Entity Linking
F$_D$ Frequency in a external dataset D
HPO Human Phenotype Ontology
IC Information Content
IC$_{dshared}$ Disjoint Shared Information Content
IC$_{shared}$ Shared Information Content
IE Information Extraction
KOS Knowledge Organization Systems
LSTM Long Short-Term Memory
MAD Mean Annotations per Document
MICA Most Informative Common Ancestors
ML Machine Learning
MTBF Mean Time Between Failures
MTDV Mean Time per Document Volume
MTSA Mean Time Seek Annotations
MTTR Mean Time To Repair
NEL Named Entity Linking
NER Named Entity Recognition
NLP Natural Language Processing
OWL Web Ontology Language
POS Part-of-speech
PPR Personalized PageRank
RE Relationship Extraction
RNN Recurrent Neural Networks
SB Semantic-Base
SSM Semantic Similarity Measure
SDP Shortest Dependency Paths

TIPS Technical Interoperability and Performance of annotation Servers
UMLS Unified Medical Language System
URI Unique Resource Identifier
VM Virtual Machine

1

Introduction

This chapter provides the necessary motivation and background to understand the objectives of this thesis. The main focus of this chapter is on showing the importance of text mining to systems biology and disease networks. Systems biology can benefit greatly from text mining due to the high number of studies that have to be assimilated to understand a particular system. In this chapter, the main hypothesis of the thesis is explained, as well as its objectives and the methods used to achieve each one.

The study of biological systems is a challenging task due to the complexity of their components and mechanisms. Systems biology studies the components of a biological system and how they interact as a whole, rather than focusing on each particular component [1]. Understanding biological systems can lead to the development of predictive models of metabolic processes, for instance, to improve drug discovery and personalized medicine. Human diseases are complex biological systems of major interest to the scientific community [2]. By combining knowledge about various human diseases, it is possible to generate a disease network. These networks are useful to understand the molecular mechanisms that are common in multiple diseases. For example, Goh et al. [3] constructed a human disease network,

based on the Online Mendelian Inheritance in Man (OMIM) database, a compendium of human disease-causing genes and phenotypes. The authors identified that disease genes, in contrast to essential genes, do not tend to be important in the interactome and are expressed only in certain tissues. These networks can also be constructed with the associations between diseases and other entities, such as symptoms [4], pathways [5], promoter regions [6] and microRNAs (miRNAs) [7]. The development of disease networks requires the integration of various data sources and studies, as well as extensive literature review.

Many databases exist for diverse types of information, for example, to store omics data, such as genomics and proteomics, as well as clinical information or molecular structures. Ontologies are used to organize the knowledge about a given domain. An ontology can be defined as a vocabulary of terms and a formal specification of their meaning [8], and can be represented as directed acyclic graphs, where each node represents a concept and each edge represents a relation between two concepts. A commonly used relation type is subsumption (*is-a*), meaning that a concept is a subclass of another concept. Ontologies are commonly used in the biomedical informatics community, where the Gene Ontology is an example of a successful effort at organizing the nomenclature of molecular functions, biological processes and cellular locations [9, 10]. More recently, linked data has become more predominant, leading to the popularization of knowledge graphs [11]. The main purpose of knowledge graphs is to improve the organization of knowledge by connecting diverse sources of information and establishing meaningful associations that can be used to answer user queries [12]. Systems biology can benefit from knowledge graphs because biological systems often involve many concepts and mechanisms, and linking knowledge about different systems is beneficial to better understand disease networks.

Text documents, such as research articles, technical reports, and patents, are the preferred method of communication by researchers. Researchers use documents to express new ideas, theories, hypotheses, methods, approaches, and experimental results with other researchers and interested parties. Therefore, there is a lot of effort put into scientific communication,

as it is an essential aspect of scientific research; the impact of a research work depends on the way it is presented to the community. Sharing these documents is crucial to scientific research, as new studies can use the knowledge generated by previous studies to improve results, develop new methods, and save time and money that would have to be spent to arrive at the same findings.

A researcher working on a biological system should consider all of the existing knowledge about that system, as well as other similar systems that may have analogous properties. Systems described by large quantities of published documents require a considerable effort to organize the information distributed across them, while less studied systems often require an extra effort to find all relevant documents, since directly related information is scarce. Modern document repositories store large quantities of documents, providing a simple way to find information about a specific domain. One of the largest sources of biomedical literature is the MEDLINE database, created in 1965. This database contains over 28 million references to journal articles in Life and Health sciences, with an increasing number of references being added every year (Figure 1.1). Other document repositories also contain information relevant to Life and Health sciences. For example, the World Intellectual Property Organization (WIPO), an agency of the United Nations, reports an increasing number of patents registered every year (Figure 1.2). A patent application contains the background information necessary to understand the invention, as well as a detailed description of the invention. Another example of a biomedical text repository is ClinicalTrials.Gov, which stores information about clinical trials (Figure 1.3). Observing the increasing growth rates shown in Figures 1.1, 1.2 and 1.3, it is clear that to find useful information in these large-scale text repositories, automatic and efficient methods are necessary.

The documents in these repositories are written in natural language since they are created by humans, to be understandable by other humans. However, computers are better suited to process structured information, hence specific techniques are required to process natural language text. Automatic methods for Information Retrieval and Information Extraction

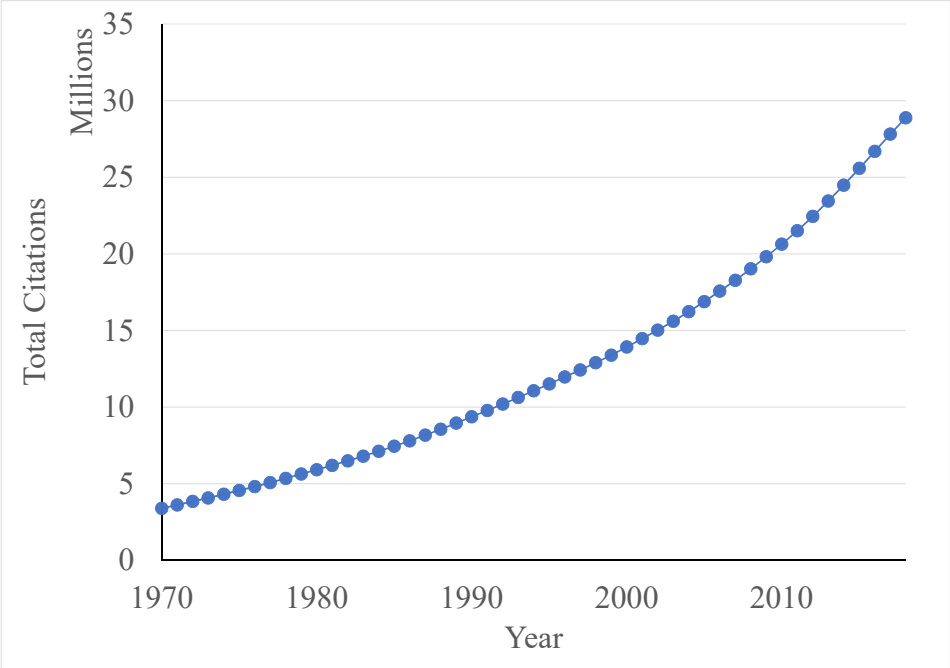


Figure 1.1: Growth of the total number of citations in the MEDLINE database. Source: https://www.nlm.nih.gov/bsd/index_stats_comp.html

aim at obtaining relevant information from large datasets, where manual methods would be infeasible. When applied to literature, this task is known as text mining [13]. Text mining techniques have been successfully applied to biomedical documents, for example, to identify disease names [14] and protein-protein interactions [15].

Since text mining can be approached from many different angles, the text mining community has defined tasks with specific objectives, for which methods can be proposed and improved. This way, each task can be evaluated individually, and a pipeline that performs multiple tasks can be assembled. One common text mining task is Named Entity Recognition (NER). The objective of this task is to identify relevant entities mentioned in any given document. Here, an entity can refer to any type of concept relevant to satisfy our information need. For example, it may be relevant to identify genes, diseases, phenotypes, and chemical compounds mentioned in documents. This task is challenging since the computer does not have any previous knowledge of what to expect from a document. While some documents

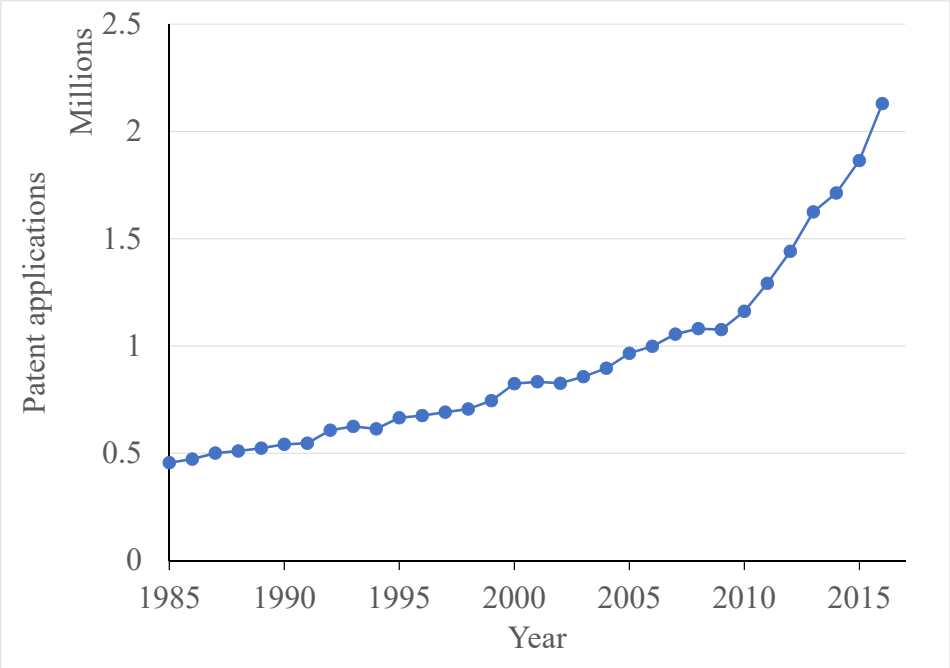


Figure 1.2: Number of worldwide patent applications submitted to patent offices by year. Source: World Intellectual Property Organization.

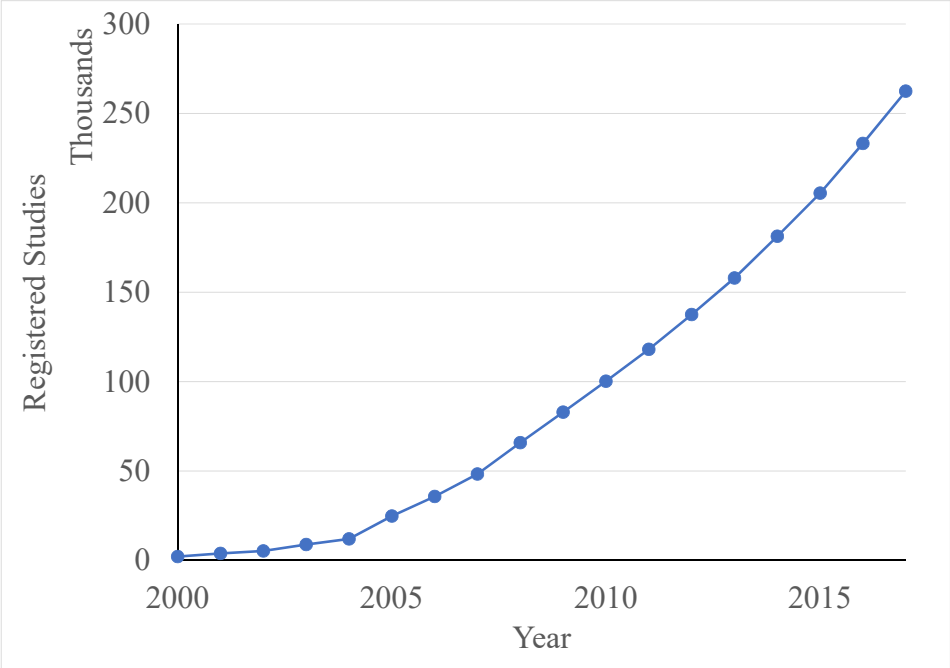


Figure 1.3: Number of registered studies in ClinicalTrials.gov by year. Source: https://clinicaltrials.gov/ct2/resources/trends

may refer to a restricted group of entities, others may contain entities that may not be expected, but could be quite relevant to a researcher. Identifying entities mentioned in documents is useful to index them or for downstream applications.

The entities found in documents may also be linked to an external reference database, as part of a task known as Entity Linking (EL)[1]. Several entity types have reference knowledge bases that aim at cataloging the variety of instances of that entity, as well as unifying variations of the nomenclature, assigning a unique identifier to each entity instance. For example, ontologies such as Chemical Entities of Biological Interest (ChEBI) and Human Phenotype Ontology (HPO) serve this purpose for biological molecules and human phenotypes, respectively. Other databases also serve this purpose, such as UniProt for proteins and SNOMED for clinical terms. By integrating the information extracted from text repositories with established knowledge bases, we can improve the usefulness of text mining methods.

[1] Variations of this task are also known as Entity Disambiguation, Harmonization or Normalization

For example, we can link studies that use different nomenclatures for the same concepts, or we can even use text mining methods to identify concepts that are missing from a particular knowledge base.

It is expected that text documents describe not only entities and their properties, but also the relations between them. A biological system is constituted not only by its elements but also by their associations and the ways they interact. Text mining can also be used to identify these relations, a task known as Relation Extraction (RE). This task is of major importance to disease networks since the edges of the network can be defined directly from text evidence. The ambiguity of scientific language is one of the challenges of this task, as well as the complexity of the mechanisms described, which may involve several entities and multiple types of relations.

The standard strategy to evaluate text mining solutions in a particular task is to perform the task manually on a set of documents (known as the gold standard) and compare manual and automatic annotations. In the case of NER, the segments of text identified as entities should be the same, or partially the same. Likewise, in EL, the database records matched with the entities are compared with the gold standard, and in RE, the sets of entities classified as being associated are compared. Evaluation measures such as precision, recall, and F1-score are then computed to assess the quality of the information extracted. Precision measures the quality of the results, corresponding to the proportion of false positives in the extracted results, while recall is the proportion of false negatives in the total information that could have been extracted. F1-score is the harmonic average of precision and recall, which is important to balance these two measures.

The success of text mining methods is dependent on the domain to which they are applied to. Comparing with domains such as news articles, biomedical literature is more complex to perform text mining due to the variety of terms and complexity of the subject matter. Therefore, specific approaches and solutions for this domain are necessary in order to obtain good results.

1.1 Objectives

The hypothesis of this thesis was that biological systems can be elucidated by the development of text mining solutions that can automatically extract information from documents. To test this hypothesis, I established three specific objectives. Figure 1.4 shows how these specific objectives relate to each other, to accomplish the main objective of the thesis. These are the three objectives of this thesis:

Objective 1 - **Named Entity Recognition (NER):** Identifying the entities mentioned in a given set of documents is the first step to generate a network about a biological system. This objective consists in developing text mining solutions to recognize biomedical entities in text. In Figure 1.4, we can see that the output of this task is the entities associated with each document.

Objective 2 - **Entity linking (EL)** The information extracted from a set of documents should be linked to other knowledge bases in order to enhance its quality. This is done by linking each entity to an entry of a reference knowledge base. This objective consists in developing an automatic approach to biomedical EL. Figure 1.4 shows that the entities found in text are converted to concepts from a reference knowledge base.

Objective 3 - **Relation Extraction (RE):** To generate a network, we need to know the relations that may exist between the concepts. Hence, we need to identify these relations in a set of documents; This objective consists in automatically extracting relations between entities, as shown in Figure 1.4.

1.2 Methodology

To achieve each of the previously established objectives, specific methodologies were explored. Each objective targets a text mining task, to develop new approaches that can

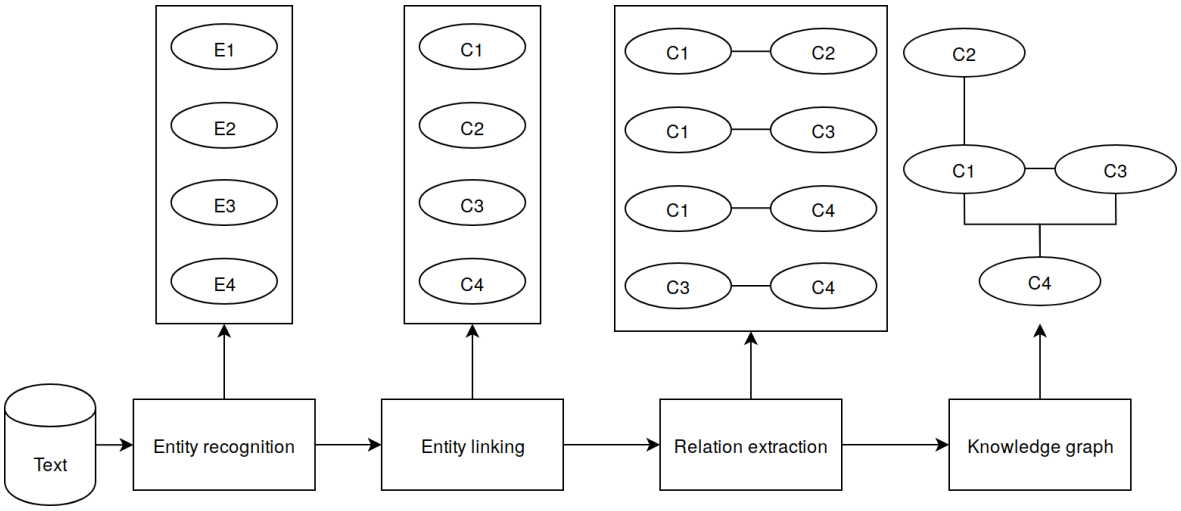


Figure 1.4: Visual representation of the objectives of this thesis, how they are associated with each other and their expected output. E1-4 refers to entities that were mentioned in a text, while C1-4 refers to entries of a knowledge base that represent those entities.

improve the results of those tasks. Using these approaches, text mining solutions were developed for specific case-studies. The methodology followed for each objective will now be described.

1.2.1 Named Entity Recognition

The main challenge of Objective 1 - Named Entity Recognition was to develop a solution that can be applied to various types of entities. Each nomenclature has its rules, exceptions, and some may be more consistent than others. An effective solution to NER should recognize entity names in text with high quality (i.e. minimal false positives and false negatives) at a reasonable computational speed and be easily adapted to different problems.

For this objective, a rule-based solution was developed, matching the names of the concepts of an ontology with the text of the documents. This solution processes the text in a specific way so that variations of the names were also identified while taking into account computational performance. Additionally, a machine learning solution was developed,

which can be trained for entity types using an annotated dataset. This solution requires more computational resources but can obtain more accurate results than the rule-based solution. Additionally, a crowdsourcing solution was explored to generate annotated datasets. Using this solution, multiple people read and annotate the documents, and the most consensual annotations were validated, reducing the cost associated with hiring expert annotators. Each of these solutions can be used to annotate a corpus of documents with entities, while each has its advantages and disadvantages.

1.2.2 Entity Linking

Objective 2 - Entity Linking consisted of matching the entities found in documents with a reference knowledge base. Documents originating from different sources may use different nomenclatures, formatting, and spellings. Even in the same document, a drug can be referred to by its full name and later by its abbreviation, for example. Therefore, a text mining system should harmonize the extracted information and link to a reference knowledge base. The vocabularies of biomedical concepts are constantly being updated as new discoveries are made and as the community agrees on specific nomenclatures. A concept may have multiple synonyms, and sometimes the same words may correspond to different concepts, depending on the context, leading to a many-to-many relationship between concepts and entities.

The most straightforward solution to link entities found in documents is to match the entity text with the labels of the knowledge base. In this thesis, different solutions were explored, where candidate matches are picked from an ontology, and a set of concepts is selected for each document. Then, the semantic similarity between each candidate match of two different sets is calculated and used to generate a graph. Based on a previous work [16], the Personalized PageRank algorithm was adapted to this graph to determine the best candidate match of each set. A measure of coherence was formulated, using semantic similarity to pick the most coherent set of concepts for that document, by maximizing the overall coherence. The method has the advantage of requiring only an ontology with "is-a" rela-

tionships, which are publicly available for several biomedical domains. The structure of the ontology itself is explored to determine the contribution of each candidate match to the overall coherence, and as such no training is necessary.

1.2.3 Relation Extraction

Extracting relations between biomedical entities is necessary to obtain more useful information from a large corpus of documents, hence Objective 3 of this thesis. Considering that multiple entities may be mentioned in the same text and the complexity of scientific writing, this is a challenging task to automate. Furthermore, it is necessary to properly define the relations to be studied and account for ambiguous cases when the text is not explicit. These difficulties also limit the number of datasets available to develop and evaluate systems for this task. We consider that a relation between two entities mentioned in the same sentence (a candidate pair) is true if the sentence describes a relation between them, or false otherwise.

Supervised machine learning solutions require an annotated dataset to train a classifier for a RE task. Since it is not feasible to manually create an annotated dataset for every biological process, there has been an increasing interest in semi-supervised and unsupervised approaches to RE. The common principle of these solutions is that a large unlabeled corpus could still be used to extract relations from text, reducing the cost of manually annotating text. If a pair of entities occurs in the same text window and the knowledge base stores a relation between those two entities, then that text must establish a relation between the two entities. However, this assumption, named distant supervision, does not always apply since the text could have a different meaning. Furthermore, there is no guarantee that the text will describe only relations contained in the knowledge base. Multi-instance learning [17] addresses the former issue, by relaxing the distant supervision assumption. With this type of model, the candidate pairs are grouped into bags where at least one of the pairs is true, but it is unknown if all pairs of the same bag are true. The bags are then classified according to the properties of their respective pairs.

A multi-instance learning solution was developed to train a relation classifier on a corpus of documents and a knowledge base containing relations between entities found in the corpus. This solution can be used for several biomedical domains, and its applicability to two different case studies was demonstrated.

Another solution to RE that was explored was supervised machine learning using recurrent neural networks. Deep learning is a set of machine learning methods, such as recurrent neural networks, that focus on learning a representation of the data instead of its features. Long short-term memory units [18] are a type of recurrent neural networks that, due to their properties, have been successfully applied to RE tasks. Existing solutions train these networks with the information that is established in the sentence that may contain a relation. However, these solutions are focused on the local contextual information and dismiss the background knowledge that a reader may already have. To account for this issue, a solution that combined recurrent neural networks with ancestor information from an ontology was developed. The assumption of this solution is that the ancestors of a concept characterize each of the entities of a candidate pair and can be used in conjunction with the sentence information to determine if the relation is true.

1.3 Contributions

The main contributions of this thesis are the text mining solutions developed and tested on biological case-studies. I developed these solutions according to the objectives previously established, using the methods previously described. The main chapters of this thesis consist of published and submitted papers that were written over the course of my doctoral work. Chapters 2 and 3 consist of review papers that I co-wrote with my supervisor, about text mining applications and semantic similarity in ontologies:

- A. Lamurias and F. Couto. 'Text Mining for Bioinformatics using Biomedical Literature'. In: *Reference Module in Life Sciences (Encyclopedia of Bioinformatics and*

Computational Biology). Vol. 1. Oxford: Elsevier, 2019. DOI: https://doi.org/10.1016/B978-0-12-809633-8.20409-3

- F. Couto and A. Lamurias. 'Semantic similarity definition'. In: *Reference Module in Life Sciences (Encyclopedia of Bioinformatics and Computational Biology)*. Vol. 1. Oxford: Elsevier, 2019. DOI: https://doi.org/10.1016/B978-0-12-809633-8.20401-9

1.3.1 Objective 1

The main contributions of Objective 1 were the MER (Minimal Entity Recognizer) and IBEnt (Identifying Biomedical Entities) components. This objective resulted in a research paper published in a Q1 journal according to the Scimago Journal Rank. I contributed to the development, evaluation, comparison of results, and writing of this paper, and for this reason, it was used as Chapter 3. This paper describes the MER software tool, which simplifies the process of identifying concepts of a specific ontology in documents and obtains better results than other rule-based tools and a lower processing time, with an average of 2.9 seconds per document. The source code of MER is publicly available and it was deployed on a cloud infrastructure so that it can be easily integrated with other components.

- Francisco Couto and Andre Lamurias. 'MER: a Shell Script and Annotation Server for Minimal Named Entity Recognition and Linking'. In: *Journal of Cheminformatics* 10.58 (2018). ISSN: 1758-2946. DOI: https://doi.org/10.1186/s13321-018-0312-9

- Code and data available at: https://github.com/lasigeBioTM/MER

I also contributed to a paper that used IBEnt to identify concepts of the Human Phenotype Ontology in text:

- M. Lobo, A. Lamurias and F. Couto. 'Identifying Human Phenotype Terms by Combining Machine Learning and Validation Rules'. In: *BioMed Research International* 2017 (2017). ISSN: 2314-6133. DOI: https://doi.org/10.1155/2017/8565739

- Code and data available at: https://github.com/lasigeBioTM/IHP

The IBEnt component, based on machine learning algorithms, was used to participate in the BioCreative V.5 challenge, obtaining competitive results in terms of processing time and quality of the annotations. The MER component was also used in this competition, as an annotation server for various types of entities. IBEnt also participated in the BioCreative V.5 challenge, as well as the 2016 and 2017 SemEval challenges:

- F. Couto, L. Campos and A. Lamurias. 'MER: a Minimal Named-Entity Recognition Tagger and Annotation Server'. In: *BioCreative V.5 Challenge Evaluation*. 2017. URL: http://www.biocreative.org/media/store/files/2017/BioCreative_V5_paper18.pdf

- A. Lamurias, L. Campos and F. Couto. 'IBEnt: Chemical Entity Mentions in Patents using ChEBI'. in: *BioCreative V.5 Challenge Evaluation*. 2017. URL: http://www.biocreative.org/media/store/files/2017/BioCreative_V5_paper12.pdf

- A. Lamurias et al. 'ULISBOA at SemEval-2017 Task 12: Extraction and classification of temporal expressions and events'. In: *10th International Workshop on Semantic Evaluation (SemEval)*. 2017. URL: http://nlp.arizona.edu/SemEval-2017/pdf/SemEval179.pdf

- M. Barros et al. 'ULISBOA at SemEval-2016 Task 12: Extraction of temporal expressions, clinical events and relations using IBEnt'. In: *9th International Workshop*

on Semantic Evaluation (SemEval). 2016. URL: http://www.aclweb.org/anthology/S/S16/S16-1196.pdf

Finally, a crowdsourcing solution to create datasets for NER was presented at the INForum 2017 symposium and at the ICBO 2015 conference early careers track:

- L. Campos, A. Lamurias and F. Couto. 'Can the Wisdom of the Crowd Be Used to Improve the Creation of Gold-standard for Text Mining applications?' In: *INForum - Simpósio de Informática*. 2017. URL: https://www.researchgate.net/publication/318751152_Can_the_Wisdom_of_the_Crowd_Be_Used_to_Improve_the_Creation_of_Gold-standard_for_Text_Mining_applications

- A. Lamurias et al. 'Annotating biomedical terms in electronic health records using crowd-sourcing'. In: *International Conference on Biomedical Ontologies (ICBO), Early Career*. 2015. URL: http://ceur-ws.org/Vol-1515/early1.pdf

1.3.2 Objective 2

For the EL objective, I developed a component that matches a set of entities found in documents to an ontology, by maximizing the global coherence of the concepts. This component was applied to two case-studies and a research article was written detailing the methods and results obtained, and submitted to a Core A conference. This manuscript was used as Chapter 8 of this thesis. The method used for this component was superior to a string matching baseline, obtaining an accuracy of 0.8039 for one of the case-studies. Furthermore, the previously mentioned MER component also performs EL when used with the vocabulary of an ontology.

1.3.3 Objective 3

Regarding Objective 3, the main contribution consisted of two different components to biomedical RE: the first based on distant supervision (IBRel) and the second using deep learning (BO-LSTM). Each component explores a different strategy to minimize the dependency on annotated datasets, by taking advantage of existing resources such as databases and ontologies. Two journal papers were published (both Q1 Scimago Journal Rank) describing the results of IBRel on two case studies, which were used for Chapters 5 and 6, respectively. The first paper focused on miRNA-gene relations, by applying the component on a set of documents about Cystic Fibrosis, obtaining a network of 27 relations from recently published papers. The second paper adapted the component to identify relations between cells and cytokines, obtaining a network of 647 relations and an F-score of 0.789.

- A. Lamurias, L. Clarke and F. Couto. 'Extracting MicroRNA-Gene Relations from Biomedical Literature using Distant Supervision'. In: *PLoS ONE* 12.3 (2017). ISSN: 1932-6203. DOI: https://doi.org/10.1371/journal.pone.0171929

- A. Lamurias et al. 'Generating a Tolerogenic Cell Therapy Knowledge Graph from Literature'. In: *Frontiers in Immunology* 8.1656 (2017). ISSN: 1664-3224. DOI: https://doi.org/10.3389/fimmu.2017.01656

- Code and data available at: https://github.com/lasigeBioTM/ICRel

Furthermore, this component was used for the BioNLP 2016 challenge, where it was tested on several relations types:

- A. Lamurias et al. 'Extraction of Regulatory Events using Kernel-based Classifiers and Distant Supervision'. In: *ACL proceedings of the 4th BioNLP Shared Task Workshop*. 2016. URL: https://aclweb.org/anthology/W/W16/W16-3011.pdf

The deep learning component resulted in a published manuscript (Q1 Scimago Journal Rank) which was used as Chapter 7 of this thesis. This component obtained a F-score of 0.751,

which is comparable to the state-of-the-art for this task.

- Andre Lamurias et al. 'BO-LSTM: Classifying relations via long short-term memory networks along biomedical ontologies'. In: *BMC Bioinformatics* 20.10 (2019). ISSN: 1471-2105. DOI: https://doi.org/10.1186/s12859-018-2584-5
- Code and data available at: https://github.com/lasigeBioTM/BOLSTM

1.4 Overview

The following chapters consist of journal publications written and published through my doctoral studies. Chapters 2 and 3 provide a literature review of text mining tools for biomedicine and semantic similarity methods for ontologies, respectively.

Chapter 4 presents the MER component, developed for Objective 1 - Named Entity Recognition. This component was implemented as a web service and compared to a machine learning approach in terms of performance and quality of the results.

Chapter 5 presents a solution to Objective 2 - Entity Linking, which combines the Personalized PageRank algorithm with the ontology structure to identify the concept that is the best match for each entity.

Chapters 6 and 7 present a solution for Objective 3 - Relation Extraction, using a reference database of associations to train a machine learning classifier. Each chapter describes how this approach was adapted to different case-studies; first to miRNA-gene relations and then to cell-cytokine relations in tolerogenic cell therapies.

Chapter 8 presents an alternative RE solution, using a deep learning algorithm. This solution explores the relations between the concepts of an ontology to better identify candidate relations in the text.

Finally, Chapter 9 presents a general discussion of my doctoral project, its main conclusions, and future work.

References

[1] Hiroaki Kitano. 'Systems biology: a brief overview'. In: *Science* 295.5560 (2002), pp. 1662–1664.

[2] Pramod Rajaram Somvanshi and KV Venkatesh. 'A conceptual review on systems biology in health and diseases: from biological networks to modern therapeutics'. In: *Systems and synthetic biology* 8.1 (2014), pp. 99–116.

[3] Kwang-il Goh et al. 'The human disease network'. In: *Proceedings of the National Academy of Sciences* 104.21 (2007), pp. 8685–8690.

[4] XueZhong Zhou et al. 'Human symptoms–disease network'. In: *Nature Communications* 5.May (2014). ISSN: 2041-1723. DOI: 10.1038/ncomms5212. URL: http://www.nature.com/doifinder/10.1038/ncomms5212.

[5] D S Lee et al. 'The implications of human metabolic network topology for disease comorbidity'. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.29 (2008), pp. 9880–9885. ISSN: 1091-6490. DOI: 10.1073/pnas.0802208105. URL: http://www.pnas.org/content/105/29/9880.

[6] Rajesh Chowdhary et al. 'A database of annotated promoters of genes associated with common respiratory and related diseases'. In: *American Journal of Respiratory Cell and Molecular Biology* 47.1 (2012), pp. 112–119. ISSN: 10441549. DOI: 10.1165/rcmb.2011-0419OC.

[7] Ming Lu et al. 'An analysis of human microRNA and disease associations'. In: *PLoS ONE* 3.10 (2008), pp. 1–5. ISSN: 19326203. DOI: 10.1371/journal.pone.0003420.

[8] Mike Uschold and Michael Gruninger. 'Ontologies: Principles, methods and applications'. In: *The knowledge engineering review* 11.2 (1996), pp. 93–136.

[9] M. Ashburner et al. 'Gene Ontology: tool for the unification of biology'. In: *Nature genetics* 25.1 (2000), pp. 25–29.

[10] Gene Ontology Consortium. 'Expansion of the Gene Ontology knowledgebase and resources'. In: *Nucleic acids research* 45.D1 (2016), pp. D331–D338.

[11] Heiko Paulheim. 'Knowledge graph refinement: A survey of approaches and evaluation methods'. In: *Semantic web* 8.3 (2017), pp. 489–508.

[12] Sören Auer et al. 'Towards a Knowledge Graph for Science'. In: *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*. WIMS '18. Novi Sad, Serbia: ACM, 2018, 1:1–1:6. ISBN: 978-1-4503-5489-9. DOI: 10.1145/3227609.3227689. URL: http://doi.acm.org/10.1145/3227609.3227689.

[13] Juliane Fluck and Martin Hofmann-Apitius. 'Text mining for systems biology'. In: *Drug Discovery Today* 19.2 (2014), pp. 140–144. ISSN: 18785832. DOI: 10.1016/j.drudis.2013.09.012. URL: http://dx.doi.org/10.1016/j.drudis.2013.09.012.

[14] Balu Bhasuran et al. 'Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases'. In: *Journal of Biomedical Informatics* 64 (2016), pp. 1–9. ISSN: 15320464. DOI: 10.1016/j.jbi.2016.09.009. URL: http://dx.doi.org/10.1016/j.jbi.2016.09.009.

[15] Martin Krallinger et al. 'Overview of the protein-protein interaction annotation extraction task of BioCreative II.' In: *Genome biology* 9 Suppl 2.Suppl 2 (2008), S4. ISSN: 1465-6914. DOI: 10.1186/gb-2008-9-s2-s4. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2559988%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[16] Maria Pershina, Yifan He and Ralph Grishman. 'Personalized Page Rank for Named Entity Disambiguation'. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Section 4. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015, pp. 238–243. ISBN: 9781941643495. DOI: 10.3115/v1/N15-1026. URL: http://aclweb.org/anthology/N15-1026.

[17] Thomas G Dietterich, Richard H Lathrop and Tomás Lozano-Pérez. 'Solving the multiple instance problem with axis-parallel rectangles'. In: *Artificial intelligence* 89.1 (1997), pp. 31–71.

[18] Sepp Hochreiter and Jürgen Schmidhuber. 'Long short-term memory'. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[19] A. Lamurias and F. Couto. 'Text Mining for Bioinformatics using Biomedical Literature'. In: *Reference Module in Life Sciences (Encyclopedia of Bioinformatics and Computational Biology)*. Vol. 1. Oxford: Elsevier, 2019. DOI: https://doi.org/10.1016/B978-0-12-809633-8.20409-3.

[20] F. Couto and A. Lamurias. 'Semantic similarity definition'. In: *Reference Module in Life Sciences (Encyclopedia of Bioinformatics and Computational Biology)*. Vol. 1. Oxford: Elsevier, 2019. DOI: https://doi.org/10.1016/B978-0-12-809633-8.20401-9.

[21] Francisco Couto and Andre Lamurias. 'MER: a Shell Script and Annotation Server for Minimal Named Entity Recognition and Linking'. In: *Journal of Cheminformatics* 10.58 (2018). ISSN: 1758-2946. DOI: https://doi.org/10.1186/s13321-018-0312-9.

[22] M. Lobo, A. Lamurias and F. Couto. 'Identifying Human Phenotype Terms by Combining Machine Learning and Validation Rules'. In: *BioMed Research International*

2017 (2017). ISSN: 2314-6133. DOI: https://doi.org/10.1155/2017/8565739.

[23] F. Couto, L. Campos and A. Lamurias. 'MER: a Minimal Named-Entity Recognition Tagger and Annotation Server'. In: *BioCreative V.5 Challenge Evaluation*. 2017. URL: http://www.biocreative.org/media/store/files/2017/BioCreative_V5_paper18.pdf.

[24] A. Lamurias, L. Campos and F. Couto. 'IBEnt: Chemical Entity Mentions in Patents using ChEBI'. In: *BioCreative V.5 Challenge Evaluation*. 2017. URL: http://www.biocreative.org/media/store/files/2017/BioCreative_V5_paper12.pdf.

[25] A. Lamurias et al. 'ULISBOA at SemEval-2017 Task 12: Extraction and classification of temporal expressions and events'. In: *10th International Workshop on Semantic Evaluation (SemEval)*. 2017. URL: http://nlp.arizona.edu/SemEval-2017/pdf/SemEval179.pdf.

[26] M. Barros et al. 'ULISBOA at SemEval-2016 Task 12: Extraction of temporal expressions, clinical events and relations using IBEnt'. In: *9th International Workshop on Semantic Evaluation (SemEval)*. 2016. URL: http://www.aclweb.org/anthology/S/S16/S16-1196.pdf.

[27] L. Campos, A. Lamurias and F. Couto. 'Can the Wisdom of the Crowd Be Used to Improve the Creation of Gold-standard for Text Mining applications?' In: *INForum - Simpósio de Informática*. 2017. URL: https://www.researchgate.net/publication/318751152_Can_the_Wisdom_of_the_Crowd_Be_Used_to_Improve_the_Creation_of_Gold-standard_for_Text_Mining_applications.

[28] A. Lamurias et al. 'Annotating biomedical terms in electronic health records using crowd-sourcing'. In: *International Conference on Biomedical Ontologies (ICBO), Early Career*. 2015. URL: http://ceur-ws.org/Vol-1515/early1.pdf.

[29] A. Lamurias, L. Clarke and F. Couto. 'Extracting MicroRNA-Gene Relations from Biomedical Literature using Distant Supervision'. In: *PLoS ONE* 12.3 (2017). ISSN: 1932-6203. DOI: https://doi.org/10.1371/journal.pone.0171929.

[30] A. Lamurias et al. 'Generating a Tolerogenic Cell Therapy Knowledge Graph from Literature'. In: *Frontiers in Immunology* 8.1656 (2017). ISSN: 1664-3224. DOI: https://doi.org/10.3389/fimmu.2017.01656.

[31] A. Lamurias et al. 'Extraction of Regulatory Events using Kernel-based Classifiers and Distant Supervision'. In: *ACL proceedings of the 4th BioNLP Shared Task Workshop*. 2016. URL: https://aclweb.org/anthology/W/W16/W16-3011.pdf.

[32] Andre Lamurias et al. 'BO-LSTM: Classifying relations via long short-term memory networks along biomedical ontologies'. In: *BMC Bioinformatics* 20.10 (2019). ISSN: 1471-2105. DOI: https://doi.org/10.1186/s12859-018-2584-5.

2

Text Mining for Bioinformatics using Biomedical Literature

ANDRE LAMURIAS AND FRANCISCO M. COUTO

Abstract

Biomedical literature has become a rich source of information for various applications. Automatic text mining methods can make the processing of extracting information from a large set of documents more efficient. However, since natural language is not easily processed by computer programs, it is necessary to develop algorithms to transform text into a structured representation. Scientific texts present a challenge to text mining methods since the language used is formal and highly specialized. This article presents an overview of the current biomedical text mining tools and bioinformatics applications using them.

2.1 Introduction

Biomedical literature is one of the major sources of current biomedical knowledge. It is still the standard method researchers use to share their findings, in the form of articles, patents and other types of written reports [1]. However, it is essential that a research group working on a given topic is aware of the work that has been done on the same topic by other groups. This task requires manual effort and may take a long time to complete, due to the large quantity of published literature. One of the largest sources of biomedical literature is the MEDLINE database, created in 1965 and accessible through PubMed. This database contains over 23 million references to journal articles in the life sciences, and more than 860,000 entries were added in 2016[1]. There are also other document repositories relevant to biomedicine, such as the European Patent Office[2], and ClinicalTrials.gov.

Automatic methods for Information Extraction (IE) aim at obtaining useful information from large datasets, where manual methods would be unfeasible. Text mining aims at using IE methods to process text documents. The main challenge of text mining is in developing algorithms that can be applied to unstructured text to obtain useful structured information. Biomedical literature is particularly challenging to text mining algorithms for several reasons. The writing style differs from other types of literature since it is more formal and complex. Furthermore, different types of documents have different styles, depending on whether the document is a journal paper, patent or clinical report [2]. Finally, there are a wide variety of terms that can be used, referring to genes, species, procedures, and techniques and, within each specific term, it is also common to have multiple spellings, abbreviations and database identifiers. These issues make biomedical text mining an interesting field for which to develop tools, due to the challenges that it presents [3].

The interactions found in the biomedical literature can be used to validate the results of new research or even to formulate new hypotheses to be tested experimentally. One of the

[1] https://www.nlm.nih.gov/bsd/index_stats_comp.html
[2] https://www.epo.org/searching-for-patents.html

first demonstrations of the hidden knowledge contained in a large literature was Swanson's ABC model [4], who found that dietary fish oils might benefit patients with Raynaud's syndrome, by connecting the information present in two different sets of articles that did not cite each other. This inference has been independently confirmed by others in clinical trials [5]. In the same study, Swanson provided two other examples of inferences that could not be drawn from a single article, but only by combining the information of multiple articles. Considering that, since that study, the number of articles available has grown immensely, it is intuitive that many new chemical interactions might be extracted from this source of information.

More recently, bioinformatics databases have adopted text mining tools to more efficiently identify new entries. MirTarBase [6] is a database of experimentally validated miRNA-target interactions published in journal papers. The curators of this database use a text mining system to identify new candidate entries for the database, which are then manually validated. This system was necessary due to the important role miRNAs have been found to play in human diseases over the last decade, leading to a high number of papers published about this subject. The introduction of the system as part of the workflow has led to a 7-fold increase in the number of interactions added to the database.

Text mining has generated much interest in the bioinformatics community in recent years. As such, several tools and applications have been developed, based on adaptations of text mining techniques to diverse problems and domains. This paper provides a survey of biomedical text mining tools and applications that demonstrate the usefulness of text mining techniques. The rest of the paper consists of the following: Section 2.2 provides the basic concepts of text mining relevant to this article, Section 2.3 describes some toolkits that can be used to develop text mining tools, Section 2.4 describes the most used text mining tools, and Section 2.5 describes applications built using those tools that have been distributed to the general public. Section 2.6 provides a summary of the community challenges organized to evaluate biomedical text mining tools. Finally, Section 2.7 suggests future directions for

biomedical text mining tools and applications, and Section 2.8 summarizes the main conclusions of the article.

2.2 Background/Fundamentals

When developing and using text mining tools, it is necessary to first define what type of information should be extracted. This decision will then influence the datasets to be considered, which text mining tasks will be explored, and which tools will be used. The objective of this section is to provide an overview of the options available to someone interested in developing a new text mining tool or using text mining for their work. The concepts presented are simple to understand and applicable to various problems.

2.2.1 NLP Concepts

Natural Language Processing (NLP) has been the focus of many researchers since the 1950's [7]. The main difference between NLP and text mining is the objective of the tasks. While NLP techniques aim at making sense of the text, for example, determining its structure or sentiment, the objective of text mining tasks is to obtain concrete structured knowledge from text. However, there is overlap between the two fields, and text mining tools usually make use of NLP concepts and tasks.

The following list defines NLP concepts relevant to text mining.

Token: a sequence of characters with some meaning, such as a word, number or symbol. The NLP task of identifying the tokens of a text is known as tokenization. It is of particular importance to text mining since most algorithms will not consider elements smaller than tokens.

Part-of-speech (POS): the lexical category of each token, for example, noun, adjective, or punctuation. The category imparts additional semantics to the tokens. Part-of-speech

tagging is an NLP task that consists in classifying each token automatically.

Lemma and stem: the base form of a word. The lemma represents the canonical form of the word, corresponding to a real word. The stem does not always correspond to a real word, but only to the fragment of a word that never changes. For example, the lemma of the word ”induces” is ”induce” while the stem is ”induc-”.

Sentence splitting: the NLP task consisting of identifying the sentence boundaries of a text. The methods used to accomplish this task should consider the difference between a period at the end of a sentence, and at the end of an acronym or abbreviation. It is desirable to break a document into sentences because they represent unique ideas. Although the context of the whole document is also important, extracting the knowledge of each sentence independently can provide useful results.

Entity: a segment of text with relevance to a specific domain. An entity may be composed of one or more tokens. Entity types relevant to biomedicine include genes, proteins, chemicals, cell lines, species, and biological processes.

2.2.2 Text Mining Tasks

Text mining tools focus on one or more text mining tasks. It is necessary to define these tasks properly so that it is possible to choose the type of tools that should be used for a given problem. Furthermore, these tasks are used to evaluate the performance of a tool on community challenges. The text mining tasks presented here are common to all domains and sources of text, although the performance of the methods on different domains may differ, i.e., a method that has a good performance on patent documents may not perform as well on clinical reports, due to the different characteristics of the text. The common final objective of these tasks, as to all text mining, is to extract useful knowledge from a high volume of documents, while the extracted knowledge can be useful for several applications, which will be described in section 2.5.

Topic modeling: the classification of documents according to their topics or themes. The objective of this task is to organize a set of documents to identify which documents are more relevant to a given topic [8]. Related tasks include document triage [9] and document clustering.

Named Entity Recognition (NER): consists of identifying entities that are mentioned in the text. In most cases, the exact location of each entity in the text is required, given by the offset of its first and last character. In some cases, discontinuous entities may be considered, therefore requiring multiple offset pairs. The classification of entity properties such as its type (e.g., protein, cell line, chemical) can be included in this task [10].

Normalization: consists of matching each entity to an identifier belonging to a knowledge base that unequivocally represents its concept. For example, a protein may be mentioned by its full name or by an acronym; in this case, the normalization process should assign the same identifier to both occurrences. The identifiers can be provided by an external database or ontology [11]. Related tasks include named entity disambiguation [12], entity linking, and harmonization.

Relationship Extraction (RE): the identification of entities that participate in a relationship described in the text. Most tools consider relations between two entities in the same sentence. Biomedical relations commonly extracted are protein-protein and drug-drug interactions, see [13], for example.

Event extraction: can be considered an extension of the relationship extraction task, where the label of the relationship and role of each participant is specified. The events extracted should represent the mechanisms described in the text [14]. Related task: slot-filling.

2.2.3 Text Mining Approaches

To accomplish the tasks described above, text mining tools employ diverse approaches. They may focus on one specific approach, or combine several approaches according to their respective advantages, the latter being more common. Most approaches can also be adapted for performing multiple tasks.

Classic approaches: approaches based on statistics that can be calculated on a large corpus of documents [15]. Some of the most popular approaches are term frequency - inverse document frequency for topic modeling, and co-occurrence for relationship extraction. These approaches preceded the popularization of machine learning algorithms, although most current approaches still have a statistical background.

Rule-based methods: consist of defining a set of rules to extract the desired information. These rules can be a list of terms, regular expressions or sentence constructions. Due to the manual effort necessary to develop these rules, text mining tools based on this approach have limited applicability.

Machine learning (ML) algorithms: are used for automatically learning various tasks. In the specific case of text mining, it is necessary to convert the text to a numeric representation, which is the expected input of these algorithms. Text mining tools using ML contain models trained on a corpus, that can then be applied to other texts. In some cases, it may be possible to train additional models using other corpora. Several types of ML approaches can be considered, for example, supervised learning, in which the labels of each instance of the training data are known and used to train the classifier, and unsupervised learning, in which the algorithm learns to classify the data without a labeled training set.

Distant supervision (DS): a learning process which heuristically assigns labels to the data according to the information provided by a knowledge base. These annotations are

prone to error, but using ML algorithms adapted to this method, it can provide effective classification models. Distant supervision is sometimes referred to as weak supervision.

2.2.4 Biomedical Corpora

Biomedical corpora are necessary to develop and evaluate text mining tools. The simplest corpora consist of a set of documents associated with a specific topic (e.g., disease, gene, or pathway). For some tasks, such as simple topic modeling tasks, it is enough to know which documents are relevant. However, most ML algorithms require annotated text to train their models. The type of annotations necessary to evaluate a task should be similar to the type of annotations to be extracted by the tools (NER tasks require text annotated with relevant entities, while relationship extraction requires the relations between the entities described in the text to be annotated). The annotations should be manually curated by domain experts according to established guidelines. Inter-annotator agreement measures, such as the kappa statistic [16], can be used to assess the reliability of the annotations. However, text mining tools may also be used to help curators by providing automatic annotations as a baseline[17].

The size of an annotated corpus is limited by the manual effort necessary to annotate the documents. Simpler tasks, such as topic modeling, can be performed more quickly by human annotators, so it is less expensive to develop an annotated corpus for this task. Relationship extraction requires that the annotators first identify the entities mentioned in the text, and then the relationships described between the entities. For this reason, it is more expensive to develop an annotated corpus for this task. Biomedical text mining community challenges have contributed to the release of several annotated gold standards that can be used to evaluate systems. Section 2.6 provides a summary of these challenges. Table 2.1 provides a list of annotated biomedical corpora relevant to various text mining tasks.

Table 2.1: Corpora relevant to biomedical text mining tasks.

Name	Reference	Annotations	Document types
CRAFT	[18]	Biomedical entities	Full-text articles
MedTag	[19]	Biomedical entities	PubMed abstracts
Genia	[20]	Biomedical entities and events	PubMed abstracts
CHEMDNER	[21]	Chemical compounds	PubMed abstracts
CHEMDNER-patents	[22]	Chemical compounds and proteins	Patent abstracts
DDI	[23]	Drug-drug interactions	Drug descriptions and journal abstracts
SeeDev	[24]	Seed development events	Full-text articles
Thyme	[25]	Events and time expressions	Clinical notes
MLEE	[26]	Biological events	PubMed abstracts

2.3 Text Mining Toolkits

Although biomedical text mining requires specialized approaches to deal with the characteristics of the biomedical literature, general text mining tools can be used as a starting point for more specialized approaches. These general tools can be adapted to specific domains, either by using models trained with biomedical datasets or by developing pre- and post-processing rules developed for this type of text. Text mining toolkits are a type of software that can perform various NLP and text mining tasks. The objective of these toolkits is to provide general-purpose methods for performing various text mining tasks, which can be adapted to specific problems. There are several toolkits available, that can be used to pre-process the data, compare the performance of various tools and approaches, and select the best combination for a specific problem. This section provides a survey of well-known text mining toolkits that have been used as frameworks of biomedical text mining tools. In addition to the toolkits presented here, tools can be developed from scratch using programming languages and libraries that implement specific algorithms.

One of the most widely used text mining toolkits is Stanford CoreNLP [27], which aggregates various tools developed by the Stanford NLP team for processing text data. Biomedical text mining tools may use Stanford CoreNLP to pre-process the data (e.g., for sentence splitting, tokenization and co-reference resolution) and to generate features for machine learning classifiers (e.g., for POS tagging, lemmatization, and dependency parsing).

NLTK [28], another NLP toolkit, was implemented as a Python library. This toolkit provides interfaces to various NLP resources, such as WordNet, tokenizers, stopwords lists, and datasets from community challenges. It is often used by developers who are getting started in text mining, due to its well-designed API, and to the availability of various online tutorials for this toolkit. More recently, another Python-based toolkit was released, spaCy[3], which is more focused on computational performance, using state-of-the-art algorithms.

ClearTK [29] is a text mining toolkit based on machine learning and the Apache Unstructured Information Management Architecture (UIMA). This framework provides interfaces to several machine learning libraries and feature extractors.

GATE [30] is one of the few text mining toolkits which has features specially designed for biomedical text mining. This toolkit provides plugins for bioinformatics resources such as Linked Life Data and other ontologies, and specialized biomedical NLP tools. Furthermore, a graphical user interface is available to visualize and edit the data and system architecture.

2.4 Biomedical Text Mining Tools

This section describes text mining tools commonly used in bioinformatics. These tools generally focus on one specific task, presenting novel approaches, and are evaluated on gold standards. We focus on tools described in the literature and freely available to the community. Even though the current trend is to make software available on code repositories such as GitHub and Bitbucket, this has not always been the case, and past works may not be accessible if the source code was not shared with the community. The tools described in this section have been used in community challenges and may require considerable technical skill to apply to specific problems since the results provided by their developers often refer to gold standards and not to real-world use-cases. These tools are usually fine tuned to work with English texts, but automatic translation techniques have been shown to be ef-

[3] https://spacy.io/

fective when using texts in other languages [31]. Table 2.2 provides a list of biomedical text mining tools that are available to the community.

2.4.1 NER and Normalization

Biomedical text mining tools can be organized in terms of the text mining tasks performed. The biomedical community challenges organized in the last decade have motivated several teams to develop tools for bioinformatics and biomedical text mining. The focus of these challenges has been in recognizing genes, proteins and chemical compounds mentioned in texts, and linking those terms to databases. This leads to an imbalance in the quantity and variety of tools available for NER and normalization when compared to other tasks.

BANNER [32] uses Conditional Random Fields [57] to perform NER of chemical compounds and genes. ABNER [33] and LingPipe [34] use similar approaches, each one combining different techniques to improve the results on gold standards, by optimizing the system architecture and feature selection. LingPipe also performs other NLP tasks, such as topic modeling and part-of-speech tagging, while all three provide ways to train models on new data. More recently, other systems have combined machine learning algorithms and manual rules to achieve better results in the biomedical domain [48, 49, 47].

GNormPlus [35] is a modular system for gene NER and normalization, performing mention simplification and abbreviation resolution to match each gene to an identifier, with higher accuracy, even when more than one species is involved. It is part of a set of NER tools developed by NCBI for various entity types, which includes tmChem [37], DNorm [36] and tmVar [38]. These tools are often evaluated in text mining community challenges.

The GENIA project is responsible for various contributions to biomedical text mining, including an annotated corpus [20] and various tools for text mining tasks. GENIA tagger [39] performs NER of several types of entities relevant to biomedicine (protein, DNA, RNA, cell line and cell types), as well as POS tagging. GENIA sentence splitter [40] is an ML-based tool for identifying sentence boundaries in biomedical texts, trained on the GENIA

Table 2.2: Text mining tools for bioinformatics and biomedical literature.

Name	Reference	Tasks	Approaches	GUI
BANNER	[32]	NER	ML	N
ABNER	[33]	NER	ML	N
LingPipe	[34]	General NLP	ML and Rule-based	N
GNormPlus	[35]	NER and Normalization	ML	N
DNorm	[36]	NER and Normalization	ML	N
tmChem	[37]	NER	ML	N
tmVar	[38]	NER	ML	N
GENIA tagger	[39]	NER and POS tagging	ML	N
GENIA sentence splitter	[40]	Sentence splitting	ML	N
Acronime	[41]	Abbreviation resolution	Rule-based	Y
@Note	[42]	NER, document retrieval	ML	Y
MetaMap	[43]	NER and Normalization	Rule-based	Y
LDPMap	[44]	Normalization	Rule-based	N
SimSem	[45]	Normalization	Rule-based and ML	N
MER	[46]	NER	Rule-based	N
IBEnt	[47]	NER and Normalization	Rule-based and ML	N
cTakes	[48]	NER, normalization, and RE	Rule-based	Y
Neji	[49]	NER and Normalization	ML and Rule-based	Y
jSRE	[50]	RE	ML	N
DeepDive	[51]	RE	ML/DS	N
IBRel	[52]	RE	ML/DS	N
TEES	[53]	Event extraction	ML and Rule-based	N
VERSE	[54]	Event extraction	ML	N
EventMine	[55]	Event extraction	ML	Y
Textpresso	[56]	NER and RE	Rule-based	Y

corpus. Acromine [41] is another tool developed by the same team, with the purpose of providing definitions for abbreviations found in MEDLINE abstracts.

Since the vocabulary used in clinical records is quite different from other biomedical texts, tools have been developed specifically for this type of documents. These tools are based on the Unified Medical Language System (UMLS), a collection of vocabularies associated with the clinical domain. cTakes [48] is a Java-based tool for processing clinical text, originally developed at the Mayo clinic, which performs several biomedical text mining tasks. It is possible to use this tool through a graphical user interface. Due to the large size and complex structure of UMLS, tools have been specifically developed just to find UMLS concepts in documents. Such tools include MetaMap [43], and LDPMap [44]. SimSem [45] is a tool for entity normalization, using string matching techniques and machine learning. This tool can match strings to a variety of bioinformatics knowledge bases, such as ChEBI, Gene Ontology, Entrez Gene, and UMLS. [46] introduced a system, MER (Minimal Entity Recognizer), which can be easily adapted to different types of entities. This system requires only a file with one entity per line, and uses a simple matching algorithm to find those entities in text.

2.4.2 Relationship and Event Extraction

For RE, most tools use ML algorithms to classify which pairs of entities mentioned in the text constitute a relationship. In this task, kernel methods and Support Vector Machines are popular. jSRE [50] uses a shallow linguistic kernel which takes into account the tokens, POS, and lemmas around each entity of the pair. It has been used for various problems, including drug-drug interaction extraction [58].

Distant supervision has become particularly relevant to RE tasks because it is more expensive to develop a corpus annotated with relations. [59] developed an approach to gene RE using DeepDive, a general purpose system for training distantly supervised RE models. They applied this approach to a corpus of full-text documents from three journals, using the

BioGRID and Negatome databases as reference. Another DS-based tool, IBRel [52], uses TransmiR, a database of miRNA-gene associations, to extract the same type of relationships from text.

Biomedical event extraction is a complex task, but some tools have been developed. TEES (Turku Event Extraction System) [53] identifies complex events based on trigger words and graph methods. This system has been evaluated on multiple community challenges, on event extraction and RE tasks, such as the BioNLP-ST 2011 event extraction task. In the 2016 edition of BioNLP-ST, [54] presented VERSE, a system for extracting relationships and events from text, and evaluated it on three different subtasks. This system is based on ML algorithms, and has the advantage of being able to extract relationships between entities in different sentences.

Textpresso [56] is a system for biomedical information extraction based on regular expressions and ontologies. This system has been applied to various domains, and a portal to search the results obtained on each domain is provided in the web interface.

2.5 Applications

Even though it is important to develop methods for specific tasks, those methods will only be useful to the community if they can be easily used to help address biomedical problems. Since recent text mining tools have obtained good performance on evaluation corpora, efforts have been made deliver these tools to the general public. In this section, we present a survey of text mining applications that are available in the form of web pages and APIs that focus on the user experience. Table 2.3 provides a summary of these applications.

Some biomedical text mining applications simply provide access to a text mining tool via a web application. The user uploads one or more documents, which are processed by the tool in a server, and the results are delivered to the user. Even though this is an important effort, it assumes that the user already has chosen the documents to be processed, and it depends on

Table 2.3: Bioinformatics applications that either use text mining tools or their results, accessible from the web.

Name	Reference	API
Whatizit	[60]	Y
becas	[61]	Y
PubTator	[62]	Y
SciLite	[63]	Y
BEST	[64]	N
STRING	[65]	Y
STITCH	[66]	Y
FACTA+	[67]	N
PolySearch2	[68]	Y
Evex	[69]	Y
MEDIE	[70]	N

downstream applications to use the results. Whatizit [60] is a text mining application that can be used to identify biomedical entities in text using a web browser or API. This application is based on a rule-based text mining system which annotates the documents submitted by users. The entities correspond to entries in biomedical knowledge bases, such as ChEBI and UniProt. The results are presented as a web page, where each entity type is marked with a different color. A similar application is BeCAS [61], based on the Neji tool. With this application, it is also possible to access the results through a web browser or the API, which can then be exported to various file formats.

Other text mining applications provide pre-processed results, reducing the time necessary to obtain results. For example, PubTator [62] contains every PubMed abstract, annotated with the NCBI NER tools, and it is updated as new abstracts are added to PubMed. Users can search for a list of abstracts or by keyword. It is possible to create a collection of abstracts, manually fix annotation errors, and download the results. PubTator provides access to the results through an API, for integration with other applications. For example, the Mark2Cure crowdsourcing project uses this API to provide a baseline of automatic annotations to its users, while the HuGE navigator knowledge base [71] relies on PubTator

to improve its weekly update process. Another application based on pre-processed results is SciLite, a platform for displaying text mining annotations, which is integrated with the Europe PMC database [63]. This application shows a list of biomedical terms associated with each document, allowing users to endorse and report incorrect annotations to improve the text mining method. Biomedical Entity Search Tool (BEST) [64] uses text mining techniques to retrieve entities relevant to user queries. BEST is updated daily with the abstracts added to PubMed, and 10 types of entities are identified in each document.

The STRING database stores information about protein-protein interaction networks [65]. It contains information obtained through various methods, including text mining. The interactions extracted using text mining methods are obtained from PubMed and a collection of full-text documents. The RE method used is based on co-occurrence of proteins in the same document, and presence of trigger words such as ”binding” and ”phosphorylation by”. A related database, STITCH [66], uses a similar method to identify chemical-protein interactions based on the biomedical literature.

FACTA+ [67] is a text mining application for identifying biomedical events described in PubMed abstracts. It uses both co-occurrence and machine learning approaches to extract relationships from text. The user can perform a keyword search to obtain associated documents and biomedical entities, such as genes, diseases, and drugs. Furthermore, FACTA+ can be used to identify indirect relations between a concept and a type of biomedical entity. For example, it is possible to search for a disease name and obtain genes that are indirectly associated with that disease, through an intermediary disease, ranked by a novelty and reliability score.

PolySearch2 [68] can also identify relationships between biomedical concepts based on co-occurrence at the sentence level. With this application, it is possible to obtain all the entities, of a specific type, associated with the input query. The corpora and databases used by this application are stored locally and updated daily to ensure that the complete information is available to the users.

EVEX [69] is a database of biomolecular events extracted from abstracts and full-text articles using text mining tools such as BANNER and TEES. This database contains more than 40 million associations between genes and proteins, and its data can be downloaded and accessed through an API, although it is not updated regularly. MEDIE [70] contains biomolecular events extracted from MEDLINE, each event being composed of a subject, a verb, and an object. Using MEDIE, it is possible to search by subject, verb or object (or a combination of the three) and obtain all matching events extracted from the abstracts.

2.6 Community Challenges

Text mining challenges are organized regularly, by the community, with the purpose of evaluating the performance of text mining tools. These text mining challenges are open to the community, meaning that any academic or industry team can participate. Each challenge usually comprises several tasks (sometimes referred to as tracks), each with a specific motivation, objective and gold standard. Each team may submit results to one or more tasks. Furthermore, the teams may develop their own tools, or adapt existing tools to the proposed task.

The task organizers announce the objectives of their task on the official websites of the challenge and on mailing lists. Since there are various data file formats used in text mining, a sample of the data may be provided to the participants at the same time as the announcement. This is also the case of datasets that require data use agreements. Afterward, the training set is provided to the participants, consisting of documents and annotations. This training set is used to develop or adapt tools and systems to the task. A development set may also be provided, similar in size to the training set, to further improve the systems. During the final phase of the challenge, a testing set is sent to the teams, without the gold standard annotations. The teams have a time period to submit the annotations obtained with their tools, which are then compared to the gold standard by the organizers. Each task has a

defined set of measures to perform this evaluation and rank the teams. The results are then published on the challenge website and in a task overview paper.

One of the earliest NLP challenges, TREC, mainly focuses on the news domain, but it has included a bioinformatics task in some of its editions (TREC Genomics and TREC Chemistry). In 2003, this challenge had a task for retrieving documents related to gene functions [72], while in later years, more complex tasks have also been proposed [73]. Other NLP challenges, such as KDD Cup [74] and CoNLL [75], also include bioinformatics tasks. SemEval is a series of semantic analysis evaluations organized yearly, and in the most recent editions, there has been at least one task relevant to bioinformatics [76, 77, 78].

Due to increasing interest in biomedical NLP and text mining, community challenges specifically for this domain have been organized. BioCreative was first organized in 2004, and it consisted of the identification of gene mentions and Gene Ontology terms in articles, and of gene name normalization [79]. Since then, five more editions of this challenge have been organized, with a wide variety of tasks. BioNLP-ST has organized various biomedical IE tasks, usually focused on a specific biological system such as seed development [24], epigenetics and post-translational modifications [80], and cancer genetics [81]. Other community challenges relevant to biomedical text mining include JNLPBA [82], BioASQ [83], i2b2 [84], and ShARe/CLEF eHealth [85]. [86] provides an overview of the community challenges organized over a period of 12 years.

2.7 Future Directions

More recent approaches to RE have explored deep learning techniques [87]. Deep learning is an ML approach based on artificial neural networks that has become popular in the last few years due to its performance in fields such as speech recognition, computer vision, and text mining [88]. In the case of text mining, deep learning is associated with word embeddings, which consist of vector representations of word frequencies, that are used as inputs to

the networks. There are still few biomedical text mining systems using deep learning techniques. However, various resources are available for this purpose, such as software toolkits that implement these algorithms, as well as a set of resources generated from biomedical literature [89].

As NER, normalization, and relationship extraction tasks improve in terms of precision and recall, semantic and question answering techniques can be developed to explore the extracted information. Semantic similarity is a metric used to compare concepts, usually based on a text corpus or an ontology [90]. These measures can both improve text mining tools by estimating the coherency of the entities and relations extracted, and be improved by applications that can generate candidate entries that may be missing from the ontology [91]. Furthermore, question-answering systems can use semantic similarity methods to provide answers with more accuracy [92].

2.8 Closing Remarks

There has been a considerable effort by the text mining community to develop and release tools and applications for bioinformatics and biomedical literature. The tools presented in this article use various methods to automate useful tasks, and they can be used by researchers who want to adapt it to their own needs. This article also presents various applications based on text mining results, which demonstrate real-world use-cases of text mining tools. The evolution of biomedical text mining methods has led to more efficient parsing of biomedical literature. These advances should affect how databases are created and maintained, and how documents are indexed by search engines. We expect that future bioinformatics search engines, instead of simply retrieving documents relevant to the query, will be able to directly answer user queries and generate new literature-based hypotheses.

2.9 Acknowledgement

This work was supported by the Fundação para a Ciência e a Tecnologia through the PhD grant PD/BD/106083/2015 and LaSIGE Unit Strategic Project, ref. UID/CEC/00408/2013 (LaSIGE).

References

[1] Marti A Hearst. 'Untangling text data mining'. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics. 1999, pp. 3–10.

[2] Carol Friedman, Pauline Kra and Andrey Rzhetsky. 'Two biomedical sublanguages: a description based on the theories of Zellig Harris'. In: *Journal of biomedical informatics* 35.4 (2002), pp. 222–235.

[3] K Bretonnel Cohen and Lawrence Hunter. 'Natural language processing and systems biology'. In: *Artificial intelligence methods and tools for systems biology*. Springer, 2004, pp. 147–173.

[4] Don R Swanson. 'Medical literature as a potential source of new knowledge.' In: *Bulletin of the Medical Library Association* 78.1 (1990), p. 29.

[5] Ralph A DiGiacomo, Joel M Kremer and Dhiraj M Shah. 'Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study'. In: *The American journal of medicine* 86.2 (1989), pp. 158–164.

[6] Chih-Hung Chou et al. 'miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database'. In: *Nucleic acids research* 44.D1 (2016), pp. D239–D247.

[7] Madeleine Bates. 'Models of natural language understanding'. In: *Proceedings of the National Academy of Sciences* 92.22 (1995), pp. 9977–9982.

[8] David M Blei. 'Probabilistic topic models'. In: *Communications of the ACM* 55.4 (2012), pp. 77–84.

[9] George Buchanan and Fernando Loizides. 'Investigating document triage on paper and electronic media'. In: *Research and Advanced Technology for Digital Libraries* (2007), pp. 416–427.

[10] David Nadeau and Satoshi Sekine. 'A survey of named entity recognition and classification'. In: *Lingvisticae Investigationes* 30.1 (2007), pp. 3–26.

[11] Yoshimasa Tsuruoka, John McNaught and Sophia Ananiadou. 'Normalizing biomedical terms by minimizing ambiguity and variability'. In: *BMC bioinformatics* 9.3 (2008), S2.

[12] Razvan Bunescu and Marius Pasca. 'Using Encyclopedic Knowledge for Named Entity Disambiguation'. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* April (2006), pp. 3–7.

[13] Isabel Segura-Bedmar, Paloma Martínez and María Herrero-Zazo. 'Lessons learnt from the DDIExtraction-2013 shared task'. In: *Journal of biomedical informatics* 51 (2014), pp. 152–164.

[14] Sophia Ananiadou et al. 'Event extraction for systems biology by text mining the literature'. In: *Trends in Biotechnology* 28.7 (2010), pp. 381–390. DOI: 10.1016/j.tibtech.2010.04.005.

[15] Christopher D Manning, Hinrich Schütze et al. *Foundations of statistical natural language processing*. Vol. 999. MIT Press, 1999.

[16] Jean Carletta. 'Assessing agreement on classification tasks: the kappa statistic'. In: *Computational linguistics* 22.2 (1996), pp. 249–254.

[17] Rainer Winnenburg et al. 'Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?' In: *Briefings in Bioinformatics* 9.6 (2008), pp. 466–478. ISSN: 1477-4054. DOI: 10.1093/bib/bbn043. URL: http://www.ncbi.nlm.nih.gov/pubmed/19060303.

[18] Michael Bada et al. 'Concept annotation in the CRAFT corpus'. In: *BMC bioinformatics* 13.1 (2012), p. 161.

[19] Lawrence H Smith et al. 'MedTag: a collection of biomedical annotations'. In: *Proceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases: mining biological semantics*. Association for Computational Linguistics. 2005, pp. 32–37.

[20] J-D Kim et al. 'GENIA corpus—a semantically annotated corpus for bio-textmining'. In: *Bioinformatics* 19.suppl 1 (2003), pp. i180–i182.

[21] Martin Krallinger et al. 'The CHEMDNER corpus of chemicals and drugs and its annotation principles'. In: *Journal of cheminformatics* 7.1 (2015), S2.

[22] Martin Krallinger et al. 'Overview of the CHEMDNER patents task'. In: *Proceedings of the fifth BioCreative challenge evaluation workshop*. 2015, pp. 63–75.

[23] María Herrero-Zazo et al. 'The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions'. In: *Journal of biomedical informatics* 46.5 (2013), pp. 914–920.

[24] Estelle Chaix et al. 'Overview of the regulatory network of plant seed development (seedev) task at the bionlp shared task 2016'. In: *Proceedings of the 4th BioNLP shared task workshop. Berlin: Association for Computational Linguistic*. 2016, pp. 1–11.

[25] William F Styler IV et al. 'Temporal annotation in the clinical domain'. In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 143–154.

[26] Sampo Pyysalo et al. 'Event extraction across multiple levels of biological organization'. In: *Bioinformatics* 28.18 (2012), pp. i575–i581.

[27] Christopher D. Manning et al. 'The Stanford CoreNLP Natural Language Processing Toolkit'. In: *Association for Computational Linguistics (ACL) System Demonstrations*. 2014, pp. 55–60. URL: http://www.aclweb.org/anthology/P/P14/P14-5010.

[28] Steven Bird, Ewan Klein and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[29] Steven Bethard, Philip Ogren and Lee Becker. 'ClearTK 2.0: Design Patterns for Machine Learning in UIMA'. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3289–3293. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/218_Paper.pdf.

[30] Hamish Cunningham et al. 'Getting more out of biomedical documents with GATE's full lifecycle open source text analytics'. In: *PLoS Comput Biol* 9.2 (2013), e1002854.

[31] Luís Campos, Vasco Pedro and Francisco Couto. 'Impact of translation on named-entity recognition in radiology texts'. In: *Database* 2017 (2017).

[32] Robert Leaman, Graciela Gonzalez et al. 'BANNER: an executable survey of advances in biomedical named entity recognition.' In: *Pacific symposium on biocomputing*. Vol. 13. 2008, pp. 652–663.

[33] Burr Settles. 'ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text'. In: *Bioinformatics* 21.14 (2005), pp. 3191–3192.

[34] Bob Carpenter. 'LingPipe for 99.99% recall of gene mentions'. In: *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Vol. 23. 2007, pp. 307–309.

[35] Chih-Hsuan Wei, Hung-Yu Kao and Zhiyong Lu. 'GNormPlus: an integrative approach for tagging genes, gene families, and protein domains'. In: *BioMed research international* 2015 (2015).

[36] Robert Leaman, Rezarta Islamaj Doğan and Zhiyong Lu. 'DNorm: disease name normalization with pairwise learning to rank'. In: *Bioinformatics* 29.22 (2013), pp. 2909–2917.

[37] Robert Leaman, Chih-Hsuan Wei and Zhiyong Lu. 'tmChem: a high performance approach for chemical named entity recognition and normalization'. In: *Journal of cheminformatics* 7.1 (2015), S3.

[38] Chih-Hsuan Wei et al. 'tmVar: a text mining approach for extracting sequence variants in biomedical literature'. In: *Bioinformatics* (2013), btt156.

[39] Yoshimasa Tsuruoka and Jun'ichi Tsujii. 'Bidirectional inference with the easiest-first strategy for tagging sequence data'. In: *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics. 2005, pp. 467–474.

[40] Rune Sætre et al. 'AKANE system: protein-protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask'. In: *Proceedings of the Second BioCreative Challenge Workshop*. Madrid. 2007, pp. 209–212.

[41] Naoaki Okazaki and Sophia Ananiadou. 'Building an abbreviation dictionary using a term recognition approach'. In: *Bioinformatics* 22.24 (2006), pp. 3089–3095.

[42] Anália Lourenço et al. '@Note: a workbench for biomedical text mining'. In: *Journal of biomedical informatics* 42.4 (2009), pp. 710–720.

[43] Alan R Aronson and François-Michel Lang. 'An overview of MetaMap: historical perspective and recent advances'. In: *Journal of the American Medical Informatics Association* 17.3 (2010), pp. 229–236.

[44] Kaiyu Ren et al. 'Effectively processing medical term queries on the UMLS Metathesaurus by layered dynamic programming'. In: *BMC medical genomics* 7.1 (2014), S11.

[45] Pontus Stenetorp, Sampo Pyysalo and Jun'ichi Tsujii. 'SimSem: Fast Approximate String Matching in Relation to Semantic Category Disambiguation'. In: *Proceedings of BioNLP 2011 Workshop*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 136–145. URL: http://www.aclweb.org/anthology/W11-0218.

[46] F. Couto, L. Campos and A. Lamurias. 'MER: a Minimal Named-Entity Recognition Tagger and Annotation Server'. In: *BioCreative V.5 Challenge Evaluation*. 2017. URL: http://www.biocreative.org/media/store/files/2017/BioCreative_V5_paper18.pdf.

[47] M. Lobo, A. Lamurias and F. Couto. 'Identifying Human Phenotype Terms by Combining Machine Learning and Validation Rules'. In: *BioMed Research International* 2017 (2017). ISSN: 2314-6133. DOI: https://doi.org/10.1155/2017/8565739.

[48] Guergana K Savova et al. 'Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications'. In: *Journal of the American Medical Informatics Association : JAMIA* 17.5 (2010), pp. 507–513. ISSN: 1067-5027. DOI: 10.1136/jamia.2009.001560.

[49] David Campos, Sergio Matos and José Luís Oliveira. 'A document processing pipeline for annotating chemical entities in scientific documents.' In: *J. Cheminformatics* 7.S-1 (2015), S7.

[50] Claudio Giuliano, Alberto Lavelli and Lorenza Romano. 'Exploiting shallow linguistic information for relation extraction from biomedical literature'. In: *11th Con-*

ference of the European Chapter of the Association for Computational Linguistics. Vol. 18. 2006. Citeseer. 2006, pp. 401–408.

[51] Ce Zhang. 'DeepDive: a data management system for automatic knowledge base construction'. PhD thesis. The University of Wisconsin-Madison, 2015.

[52] A. Lamurias, L. Clarke and F. Couto. 'Extracting MicroRNA-Gene Relations from Biomedical Literature using Distant Supervision'. In: *PLoS ONE* 12.3 (2017). ISSN: 1932-6203. DOI: https://doi.org/10.1371/journal.pone.0171929.

[53] Jari Björne et al. 'EXTRACTING CONTEXTUALIZED COMPLEX BIOLOGICAL EVENTS WITH RICH GRAPH-BASED FEATURE SETS'. In: *Computational Intelligence* 27.4 (2011), pp. 541–557.

[54] Jake Lever and Steven JM Jones. 'VERSE: Event and relation extraction in the BioNLP 2016 Shared Task'. In: *Proceedings of the 4th BioNLP Shared Task Workshop*. 2016, p. 42.

[55] Makoto Miwa et al. 'Wide coverage biomedical event extraction using multiple partially overlapping corpora'. In: *BMC bioinformatics* 14.1 (2013), p. 175.

[56] Hans-Michael Michael Müller, Eimear E. Kenny and Paul W. Sternberg. 'Textpresso: an ontology-based information retrieval and extraction system for biological literature.' In: *PLoS Biology* 2.11 (2004), e309. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0020309.

[57] Charles Sutton and Andrew McCallum. 'An introduction to conditional random fields for relational learning'. In: *Introduction to statistical relational learning* (2006), pp. 93–128.

[58] Isabel Segura-Bedmar, Paloma Martinez and Cesar de Pablo-Sánchez. 'Using a shallow linguistic kernel for drug–drug interaction extraction'. In: *Journal of biomedical informatics* 44.5 (2011), pp. 789–804.

[59] Emily K Mallory et al. 'Large-scale extraction of gene interactions from full-text literature using DeepDive'. In: *Bioinformatics* 32.1 (2016), pp. 106–113.

[60] Dietrich Rebholz-Schuhmann et al. 'Text processing through Web services: calling Whatizit'. In: *Bioinformatics* 24.2 (2008), pp. 296–298.

[61] Tiago Nunes et al. 'BeCAS: biomedical concept recognition services and visualization'. In: *Bioinformatics* 29.15 (2013), pp. 1915–1916.

[62] Chih-Hsuan Wei, Hung-Yu Kao and Zhiyong Lu. 'PubTator: a web-based text mining tool for assisting biocuration'. In: *Nucleic acids research* (2013), gkt441.

[63] Aravind Venkatesan et al. 'SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data'. In: *Wellcome Open Research* 1 (2016), p. 25. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.10210.1.

[64] Sunwon Lee et al. 'BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature'. In: *PloS ONE* 11.10 (2016), e0164680.

[65] Damian Szklarczyk et al. 'The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible'. In: *Nucleic acids research* 45.D1 (2017), pp. D362–D368.

[66] Damian Szklarczyk et al. 'STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data'. In: *Nucleic acids research* 44.D1 (2016), pp. D380–D384.

[67] Yoshimasa Tsuruoka et al. 'Discovering and visualizing indirect associations between biomedical concepts'. In: *Bioinformatics* 27.13 (2011), pp. 111–119. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr214.

[68] Yifeng Liu, Yongjie Liang and David Wishart. 'PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more'. In: *Nucleic acids research* 43.W1 (2015), W535–W542.

[69] Kai Hakala et al. 'EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction'. In: *Proceedings of the BioNLP Shared Task 2013 Workshop*. Association for Computational Linguistics. 2013, pp. 26–34.

[70] Yusuke Miyao et al. 'Semantic retrieval for the accurate identification of relational concepts in massive textbases'. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 2006, pp. 1017–1024.

[71] Wei Yu et al. 'A navigator for human genome epidemiology.' In: *Nature genetics* 40.2 (2008), pp. 124–125. ISSN: 1061-4036. DOI: 10.1038/ng0208-124.

[72] William R Hersh and Ravi Teja Bhupatiraju. 'TREC genomics track overview.' In: *TREC*. Vol. 2003. 2003, pp. 14–23.

[73] William Hersh and Ellen Voorhees. 'TREC genomics special issue overview'. In: *Information Retrieval* 12.1 (2009), pp. 1–15.

[74] Alexander Yeh, Lynette Hirschman and Alexander Morgan. 'Background and overview for KDD Cup 2002 task 1: Information extraction from biomedical articles'. In: *ACM SIGKDD Explorations Newsletter* 4.2 (2002), pp. 87–89.

[75] Richárd Farkas et al. 'The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text'. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*. Association for Computational Linguistics. 2010, pp. 1–12.

[76] Isabel Segura Bedmar, Paloma Martínez and María Herrero Zazo. 'Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)'. In: *Proceedings of the Seventh International Workshop on Semantic Evaluation*. Association for Computational Linguistics. 2013.

[77] Noémie Elhadad et al. 'SemEval-2015 task 14: Analysis of clinical text'. In: *Proc of Workshop on Semantic Evaluation. Association for Computational Linguistics*. 2015, pp. 303–10.

[78] Steven Bethard et al. 'Semeval-2016 task 12: Clinical tempeval'. In: 2016, pp. 1052–1062.

[79] Lynette Hirschman et al. 'Overview of BioCreAtIvE: critical assessment of information extraction for biology'. In: *BMC Bioinformatics* 6.1 (2005), S1.

[80] Tomoko Ohta, Sampo Pyysalo and Jun'ichi Tsujii. 'Overview of the epigenetics and post-translational modifications (EPI) task of BioNLP shared task 2011'. In: *Proceedings of the BioNLP Shared Task 2011 Workshop*. Association for Computational Linguistics. 2011, pp. 16–25.

[81] Sampo Pyysalo et al. 'Overview of the cancer genetics and pathway curation tasks of BioNLP shared task 2013'. In: *BMC Bioinformatics* 16.10 (2015), S2.

[82] Jin-Dong Kim et al. 'Introduction to the bio-entity recognition task at JNLPBA'. In: *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*. Association for Computational Linguistics. 2004, pp. 70–75.

[83] Anastasia Krithara et al. 'Results of the 4th edition of BioASQ Challenge'. In: *Proceedings of the Fourth BioASQ workshop at the Conference of the Association for Computational Linguistics*. 2016, pp. 1–7.

[84] Weiyi Sun, Anna Rumshisky and Ozlem Uzuner. 'Evaluating temporal relations in clinical text: 2012 i2b2 Challenge'. In: *Journal of the American Medical Informatics Association* 20.5 (2013), pp. 806–813.

[85] Liadh Kelly et al. 'Overview of the share/clef ehealth evaluation lab 2014'. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer. 2014, pp. 172–191.

[86] Chung Chi Huang and Zhiyong Lu. 'Community challenges in biomedical text mining over 10 years: Success, failure and the future'. In: *Briefings in Bioinformatics* 17.1 (2016), pp. 132–144. ISSN: 14774054. DOI: 10.1093/bib/bbv024.

[87] Makoto Miwa and Mohit Bansal. 'End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures'. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1105–1116. DOI: 10.18653/v1/P16-1105. URL: http://aclweb.org/anthology/P16-1105.

[88] Yann LeCun, Yoshua Bengio and Geoffrey Hinton. 'Deep learning'. In: *Nature* 521.7553 (2015), p. 436.

[89] Sampo Pyysalo et al. 'Distributional semantics resources for biomedical text processing'. In: *Proceedings of Languages in Biology and Medicine*. LBM, 2013.

[90] Francisco M Couto and H Sofia Pinto. 'The next generation of similarity measures that fully explore the semantics in biomedical ontologies'. In: *Journal of bioinformatics and computational biology* 11.05 (2013), p. 1371001.

[91] Maria Pershina, Yifan He and Ralph Grishman. 'Personalized Page Rank for Named Entity Disambiguation'. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Section 4. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015, pp. 238–243. ISBN: 9781941643495. DOI: 10.3115/v1/N15-1026. URL: http://aclweb.org/anthology/N15-1026.

[92] Vanessa Lopez, Michele Pasin and Enrico Motta. 'Aqualog: An ontology-portable question answering system for the semantic web'. In: *European Semantic Web Conference*. Springer. 2005, pp. 546–562.

3

Semantic similarity definition

FRANCISCO M. COUTO AND ANDRE LAMURIAS

Abstract

In bioinformatics, semantic similarity has been used to compare different types of biomedical entities, such as proteins, compounds and phenotypes, based on their biological role instead on what they look like. This manuscript presents a definition of semantic similarity between biomedical entities described by a common semantic base (e.g. ontology) following an information-theoretic perspective of semantic similarity. It defines the amount of information content two entries share in a semantic base, and, by extension, how to compare biomedical entities represented outside the semantic base but linked through a set of annotations. Software to check how semantic similarity works in practice is available at: https://github.com/lasigeBioTM/DiShIn/.

3.1 Introduction

The biological role of an entity is considered to be its semantics, which has been increasingly being represented through common vocabularies. The entries in these vocabularies represent biological features, that are often connected with each other by semantic relations, such as subsumption. The availability of these common vocabularies, and their usage to semantically annotate entities enabled the development of computational semantic similarity measures [1]. Before defining semantic similarity, we should start by defining why bioinformatics needs semantic similarity in the first place, then what it is, to finally describe how it can be calculated.

3.1.1 Why?

Biomedical entities, such as proteins or chemical compounds, are frequently compared to each other to find similarities that may able us to transfer knowledge from one another. In the case of proteins, one of the most popular techniques is to calculate sequence similarity by locating short matches between sequences and then generate local alignments [2]. In the case of compounds, one of the most popular techniques is to calculate the number of 2D substructural fragments (molecular fingerprints) that they have in common [3]. The above techniques are popular mainly because they can be implemented by high performance tools, such as BLAST [4], and are based on simple, unambiguous and widely available digital representations. However, these digital representations result from observations of how these biomedical entities look like, and not about their semantics. This means that from these digital representations we cannot have a direct insight about their biological role. Sequence similarity and common fingerprints measure how close two entities are in terms of what they look like, which may differ from their biological role.

There is an association between what an entity looks like and its biological role, i.e. proteins with similar sequence tend to have similar molecular functions, as well as with com-

pounds with similar molecular shapes. However, there are many exceptions. For example, crystallins have a high sequence similarity to several different enzymes due to evolution, but in the eye lens their role is to act as structural proteins, not enzymes [5]. Another example is caffeine and adenosine. These two molecules have a similar shape, so similar that caffeine is able to bind to adenosine receptors [6]. However, adenosine induces sleep and suppresses arousal while caffeine makes you more awake and less tired. Semantic similarity addresses the above exceptions, by comparing biomedical entities based on what they do and not on what they look like. This means that when looking for similar compounds to caffeine, other central nervous system stimulants, such as doxapram, will appear before adenosine that has the opposite effect.

3.1.2 What?

Digital representations of biomedical entities based on structure can normally be expressed using a simple syntax. For example, ASCII strings are used to represent: the nucleotide sequences of genes; the amino acid sequences of proteins, and also the structure of compounds using SMILES. Semantics is however more complex since it may have different interpretations according to a given context. For example, the meaning of a biological role of a given gene may differ from a biological or medical perspective. For humans the easiest way to represent semantics is to use free text due to its flexibility to express any concept. For example, short text comments are usually valuable semantic descriptions to understand the meaning of a piece of information. However, for computers free text is not the most effective form of encoding semantics, making semantic similarity measurement between different text descriptions almost unfeasible.

In recent years, the biomedical community made a substantial effort in representing the semantics of biomedical entities by using common vocabularies, which vary from simple terminologies to highly complex semantic models. These vocabularies are instantiated by Knowledge Organization Systems (KOS) in the form of classification systems, thesauri, lex-

ical databases, gazetteers, and taxonomies, and ontologies[7]. Perhaps the most well-known KOS is the Gene Ontology, which has been extensively used to annotate gene-products with terms describing their molecular functions, biological processes and cellular components, and the source of most semantic similarity studies in bioinformatics. This manuscript will denote a KOS used in a semantic similarity measure as its Semantic-Base (SB). Semantic similarity measures become feasible when a biomedical community accepts a SB as a standard to represent the semantics of the entities in their domain. Semantic similarity is therefore a measure of how close are the semantic representations of different biomedical entities in a given SB. This means that the semantic similarity between two entities depends on their SB representation and also on a similarity measure that calculates how close these representations are in the SB.

3.1.3 How?

We may think that given a SB, we should be able to find the optimal quantitative function to implement semantic similarity. However, the notion of semantic similarity is dependent on what are the objectives of the study. For example, a biologist and a physician may have two different expectations about the semantic similarity between the biological roles of two genes.

In bioinformatics, ontologies have been the standard SB for calculating semantic similarity. The SB provides an unambiguous context on where semantic representations can be interpreted. A semantic representation is sometimes referred as a set of annotations, i.e. a link between the entity and an entry in the SB. Each entity can have multiple annotations. This means that the similarity measure may be applied for multiple entries in the SB. There are also different types of annotations. For example, an annotation can represent a finding with experimental evidence, or just a prediction from a computational method. Semantic similarity can explore the different types of annotations, for example to filter out annotations in which we have lower confidence.

A similarity measure is a quantitative function between entries in the SB, which explores the relations between its entries to measure their closeness in meaning. An entry is normally connected to the other entries by different types of relations represented in the SB. The similarity measure calculates the degree of shared meaning between two entries, resulting in a numerical value. For example, this can be performed by identifying a path between both entries in the SB, and calculating the semantic gap encoded in that path. This means that a semantic similarity measure can be defined by the SB and the quantitative measure used, which will be formulated in the following sections.

3.2 Semantic Base

Definition 1 (Semantic-Base) *A Semantic-Base is a tuple $SB =< E, R >$, such that E is the set of entries, and R is the set of relations between the entries. Each relation is pair of entities (e_1, e_2) with $e_1, e_2 \in E$.*

When using biomedical ontologies, the entries represent the classes, terms or concepts. This definition ignores the type of relations that may be present in the ontology, since semantic similarity measures are normally restricted to subsumption relations (*is-a*). Nevertheless, a measure may use other type of relation, or even use different types of relations. The interpretation of its results should take this into consideration. One of the reasons why subsumption relations are used is because they are transitive, i.e. if $(e_1, e_2) \in R$ and $(e_2, e_3) \in R$ then we can implicitly assume that (e_1, e_3) is also a valid relation. This enables us to define the ancestors and descendants of a given entry.

Definition 2 (Ancestors) *Given a SB represented by the tuple $< E, R >$, and T the transitive closure of R on the set E (i.e. the smallest relation on E that contains R and is transitive), the Ancestors of a given entry $e \in E$ are defined as $Anc(e) = \{a : (e, a) \in T\}$*

Definition 3 (Descendants) *Given a SB represented by the tuple $< E, R >$, and T the*

transitive closure of R on the set E, the Descendants of a given entry $e \in E$ are defined as $Des(e) = \{d : (d, e) \in T\}$

There are multiple successful semantic similarity measures being used in bioinformatics. Many of them are inspired on the contrast model proposed by [8], in the sense that they balance the importance of common features versus the exclusives. Thus, a semantic similarity measure can be categorized by how it defines the common features, and how it calculates the importance of each feature. The first step in most measures is to find the common ancestors in the SB to define the common features.

Definition 4 (Common Ancestors) *Given a SB represented by the tuple $< E, R >$, the Common Ancestors of two entries $e_1, e_2 \in E$ is defined as $CA(e_1, e_2) = Anc(e_1) \cap Anc(e_2)$.*

3.3 Information Content

This manuscript follows an information-theoretic perspective of semantic similarity [9]. To calculate the importance of each entry the measures identify the information content of each entry. [10] defined the information content of an entry based on the notion of the entropy of the random variable X known in information theory [11]. The intuition is to measure the surprise evoked by having an entry $e \in E$ in the semantic representation.

Definition 5 (Information Content) *Given a SB represented by the tuple $< E, R >$, and a probability function $p : E \rightarrow]0, 1]$, the information content of an entry $e \in E$ is defined as $IC(e) = -log(p(e))$.*

The probability function should be defined in a way that bottom-level entries in the SB become more informative than top-level entries, making the $IC(e)$ correlated with the specificity of e in the SB.

The definition of the probability function p can follow two different approaches:

Intrinsic: p is based only on the internal structure of the SB.

Extrinsic: p is based on the frequency of each entry in an external dataset.

Considering the graph represented in Figure3.1 as our SB, and assuming an intrinsic approach $p(e) = \frac{Desc(e)+1}{|E|}$, then we have all the bottom entries with p equal to $\frac{1}{8}$, $p(coinage) = \frac{4}{8}$, $p(precious) = \frac{5}{8}$, and $p(metal) = \frac{8}{8}$. Thus, we have $IC(metal) < IC(precious) < IC(coinage) < IC(platinum) \ldots < IC(copper)$. Note also that the addition of 1 to avoid having a zero probability for the entries without descendants.

Definition 6 (Frequency) *Given a SB represented by the tuple $< E, R >$, and an external dataset D , and a predicate $refer(d, e)$ that is true when a data element $d \in D$ refers the entry $e \in E$, then the frequency of a given entry in that dataset is defined as*

$$F_D(e) = |\{d : refer(e_1, d) \wedge d \in D \wedge e_1 \in Desc(e) \cup \{e\}\}|$$

Note that when using subsumption relations, i.e. an occurrence of an entry, it is also an implicit occurrence of all its ancestors.

Definition 7 (Extrinsic Probability) *Given a SB represented by the tuple $< E, R >$, and a frequency measure F_D the extrinsic probability function of an entry $e \in E$ is defined as*

$$p(e) = \frac{F_D(e) + 1}{max\{F_D(e_1) : e_1 \in E\} + 1}$$

.

Note that top-level entries have high frequency values due the occurrences of their descendants, so their IC is close to zero. Note again the addition of 1 this time in both parts of the fraction to avoid having a zero probability.

Considering again the graph represented in Figure3.1 as our SB, and assume an external dataset D containing exactly one occurrence of each entry, then we have all the bottom

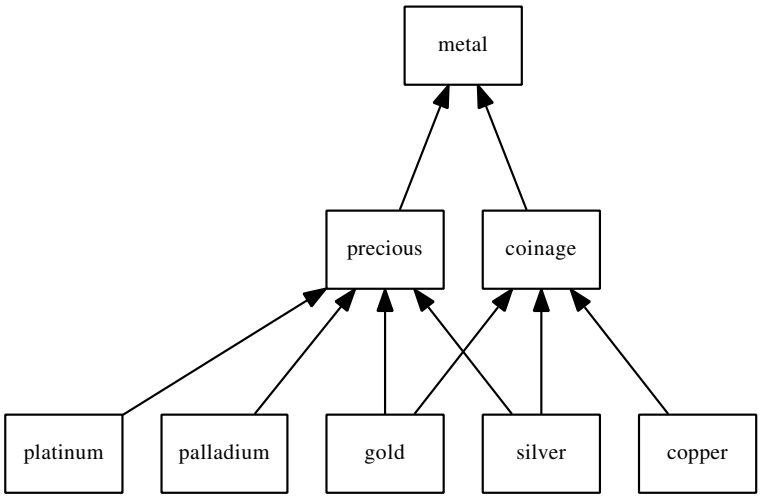


Figure 3.1: This graph represents an example of a classification of metals with multiple inheritance, since *gold* and *silver* are considered both precious and coinage metals.

entries with F_D equal to $\frac{2}{9}$, $F_D(coinage) = \frac{5}{9}$, $F_D(precious) = \frac{6}{9}$, and $F_D(metal) = \frac{9}{9}$. Thus, we again have $IC(metal) < IC(precious) < IC(coinage) < IC(platinum) \ldots < IC(copper)$. We will assume this IC instantiation for the remainder examples in this manuscript.

3.4 Shared Ancestors

Not all ancestors are relevant when calculating semantic similarity since some of them are already subsumed by others and do not represent any new information. So normally the measures select only the most informative ones.

Definition 8 (Most Informative Common Ancestors) *Given a SB represented by the tuple* $< E, R >$*, and an IC measure, the Most Informative Common Ancestors of two entries* $e_1, e_2 \in E$ *is defined as*

$$MICA(e_1, e_2) = \{a : a \in CA(e_1, e_2) \wedge IC(a) = max\{IC(a_1) : a \in CA(e_1, e_2)\}\}$$

Considering again the graph represented in Figure3.1 as our SB, and the extrinsic IC defined above, then we have $MICA(platinum, copper) = \{metal\}$, $MICA(silver, gold) = \{coinage\}$, and $MICA(platinum, gold) = \{precious\}$.

Sometimes the most informative common ancestors are not sufficient, since they may neglect multiple inheritance relations. Thus, instead of $MICA$, the measures can use the disjunctive common ancestors [12].

Definition 9 (Disjunctive Common Ancestors) *Given a SB represented by the tuple $< E, R >$, and an IC measure, and a function $PD : E \times E \times E \to \mathbb{N}$, that calculates the difference between the number of paths from the two entries to one of their comon ancestors, the Disjunctive Common Ancestors of two entries $e_1, e_2 \in E$ is defined as*

$$\begin{aligned} DCA(e_1, e_2) = \{a \ : \\ & a \in CA(e_1, e_2) \wedge \\ & \forall_{a_x \in CA(e_1, e_2)} PD(e_1, e_2, a) = PD(e_1, e_2, a_x) \\ & \Rightarrow IC(a) > IC(a_x)\} \end{aligned}$$

Considering again the graph represented in Figure3.1 as our SB, the extrinsic IC defined above, then we have $DCA(silver, gold) = \{coinage, precious\}$, and $DCA(platinum, gold) = \{precious, metal\}$.

3.5 Shared Information

The importance of common features is defined by the shared IC present in the common ancestors, normally its average.

Definition 10 (Shared Information Content) *Given a SB represented by the tuple $< E, R >$, and an IC measure, the Shared Information Content of two entries $e_1, e_2 \in E$ is defined as* $IC_{shared}(e_1, e_2) = \overline{\{IC(a) : a \in DCA(e_1, e_2)\}}$.

Note that DCA can be replaced by $MICA$, however since all ancestors in $MICA$ have the same IC value by definition only that IC value is used in practice.

Considering again the graph represented in Figure3.1 as our SB, the extrinsic IC defined above, then when using $MICA$ we have $IC_{shared}(platinum, gold) = -log(\frac{6}{9})$. If we use DCA then we have $IC_{shared}(platinum, gold) = (-log(\frac{6}{9}) - log(\frac{9}{9}))/2$.

More recently, [13] proposed the usage of the disjointness axioms in semantic similarity by defining the disjoint shared information content. The idea is that if we know that two entries are disjoint, then we should decrease their amount of shared information.

Definition 11 (Disjoint Shared Information Content) *Given a SB represented by the tuple $< E, R >$, a set of axioms A, and an IC_{shared} measure, the Disjoint Shared Information Content of two entries, $e_1, e_2 \in E$ is defined as $IC_{dshared}(e_1, e_2) = IC_{shared}(e_1, e_2) - k(e_1, e_2)$ with $k : E \times E \rightarrow \mathbb{N}$ satisfying the following conditions: i) $k(e_1, e_2) > 0$ if e_1 and e_2 are disjoint according to A; ii) $k(e_1, e_2) = 0$ if otherwise.*

3.6 Similarity Measure

Definition 12 (Semantic Similarity Measure) *Given a SB represented by the tuple $< E, R >$, a Semantic Similarity Measure is a quantitative function $SSM : E \times E \rightarrow \mathbb{R}$.*

Note that a semantic similarity measure is not expected to be instantiated by the inverse of a metric or distance function, but the following conditions are normally satisfied:

non-negativity: $SSM(e_1, e_2) \geq 0$ with $e_1, e_2 \in E$;

symmetry: $SSM(e_1, e_2) = SSM(e_2, e_1)$ with $e_1, e_2 \in E$.

Many measures are also normalized, i.e. $SSM(e_1, e_2) \in [0..1]$ with $e_1, e_2 \in E$; and $SSM(e, e) = 1$ with $e \in E$.

The seminal work based on Resnik's measure [10] was one of the first measures to be successfully applied to a biomedical ontology, namely the Gene Ontology. [14]. The measure was defined as:

$$SSM_{resnik}(e_1, e_2) = IC_{shared}(e_1, e_2)$$

Another well-known measure, was defined by [15] as:

$$SSM_{lin}(e_1, e_2) = \frac{2 \times IC_{shared}(e_1, e_2)}{IC(e_1) + IC(e_2)}$$

where the denominator represents the exclusive features.

Note that both measures are independent of using $MICA$ or DCA as the common features.

Considering again the graph represented in Figure3.1 as our SB, the extrinsic IC defined above, and $MICA$, then we have $SSM_{resnik}(platinum, gold) = -log(\frac{6}{9})$ and $SSM_{lin}(platinum, gold) = (2 \times -log(\frac{6}{9}))/(-log(\frac{2}{9}) - log(\frac{2}{9}))$.

3.6.1 Entity Similarity

Until now we only defined SSM in terms of entries, but a biomedical entity may not be directly represented in the SB, but instead linked to the SB through annotations. For example in the case of proteins, they are not represented as entries of the Gene Ontology but through annotations. In opposition, chemical compounds are represented as entries of the ontology Chemical Entities of Biological Interest (ChEBI).

Definition 13 (Annotation) *Given a SB represented by the tuple $< E, R >$ and a set of biomedical entities B, a predicate $annotates(b, e)$ that is true when the entity $b \in B$ is annotated with the entry $e \in E$, then the annotation set of a biomedical entity (or concept) $b \in B$ is defined as*

$$AS(b) = \{e : e \in E \wedge annotates(b, e)\}$$

This definition ignores the type of annotation, e.g. with experimental or computational evidence, since the similarity measure calculation is usually independent of this information. It is up to the user to decide which type of annotations to include.

To compare biomedical entities we need to extend the SSM definition so it applies to the two sets of entries of each entity, instead of a single entry for each entity. For readability we will use the same function name SSM, to represent different functions according to the input domain, i.e. two entries or two sets of entries.

There are multiple successful instantiations of entity semantic similarity measures, and most of them use two aggregate functions (e.g. average, maximum) on the results from comparing each pair of entries annotated to each entry.

Definition 14 (Aggregate Measure) *Given a SB represented by the tuple $< E, R >$, a set of biomedical entities B, two aggregate functions f and g, and two biomedical entities $b_1, b_2 \in B$ the Aggregate Similarity Measure is defined as*

$$SSM_{aggregate}(AS(b_1), AS(b_2)) = f(\{g(\{SSM(e_1, e_2) : e_1 \in AS(b_1)\}) : e_2 \in AS(b_2)\})$$

Considering again the graph represented in Figure3.1 as our SB, f as the average function, g as the maximum function, two entities containing metals $B = \{\alpha, \beta\}$, and their respective annotation set $AS(\alpha) = \{platinum, palladium\}$ $AS(\beta) = \{copper, gold\}$, then we have

$$\begin{aligned} SSM_{aggregate}(\{platinum, palladium\}, \{copper, gold\}) = avg\{ \\ max\{SSM(platinum, copper), SSM(platinum, gold)\}, \\ max\{SSM(palladium, copper), SSM(palladium, gold)\}\} \end{aligned}$$

Another popular approach is to apply the Jaccard coefficient to all common entries vs. the exclusive ones.

Definition 15 (Jaccard Measure) *Given a SB represented by the tuple $< E, R >$, a set of biomedical entities B, an annotation set AT, and two biomedical entities $b_1, b_2 \in B$ the similarity measure is defined as*

$$SSM_{jaccard}(AS(b_1), AS(b_2)) = \frac{\sum\{IC(e) : e \in \{Anc(e_1) : e_1 \in AS(b_1)\} \cap \{Anc(e_2) : e_2 \in AS(b_2)\}\}}{\sum\{IC(e) : e \in \{Anc(e_1) : e_1 \in AS(b_1)\} \cup \{Anc(e_2) : e_2 \in AS(b_2)\}\}}$$

Considering the example above of α and β when using Jaccard we will have

$$SSM_{jaccard}(\{platinum, palladium\}, \{copper, gold\}) = \frac{IC(precious) + IC(metal)}{IC(coinage) + IC(precious) + IC(metal)}$$

3.7 Future Directions

This manuscript is focused on defining semantic similarity using a single KOS, however a large amount of biomedical resources use multiple KOS describing a single domain from different perspectives or even distinct domains. Calculating semantic similarity using multiple KOS as SB is a complex problem, and only a few works have addressed it [16]. Thus, a future formulation of multiple domain semantic similarity is much required.

Another issue is about the incompleteness of KOS. They normally represent work in progress, being updated as our knowledge of the domain becomes more sound and comprehensive. Keeping a KOS up-to-date is also a daunting task in terms of human effort, especially in large KOS, so we should always expect to have a delay until new knowledge is incorporated. This means that the common features identified in a KOS may be incomplete, and the exclusives features may not even be exclusive in the future. If a biomedical entity is not annotated with a specific feature, that does not mean that the entity does not have that feature, it only means that we do not know if it has or not. Thus, a future formulation of semantic similarity that takes in account the incompleteness of KOS is also much required.

3.8 Closing Remarks

This manuscript presented a definition of semantic similarity following an information-theoretic perspective that covers a large number of the measures currently being used in bioinformatics. It defined the amount of information content two entries share in a SB, and how it can be extended to compare biomedical entities represented outside the SB but linked through a set of annotations.

The manuscript aims at providing a generic and inclusive formulation that can be helpful to understand the fundamentals of semantic similarity and at the same time be used as a guideline to distinguish between different approaches. The formulation did not aim at providing a one size fits all definition, i.e. trying to represent all measures being proposed.

The manuscript presented well-known measures in bioinformatics, Resnik, Lin and Jaccard coefficient, according to the proposed definitions. It also presented their results when applied to simple example of a classification of metals, which is used along the text to clarify the definitions being presented. Finally, a software repository [1] is available to test and learn more on how semantic similarity works in practice.

3.9 Acknowledgement

This work was supported by FCT through the PhD grant PD/BD/106083/2015 and LaSIGE Research Unit, ref. UID/CEC/00408/2013 (LaSIGE).

References

[1] Montserrat Batet et al. 'An information theoretic approach to improve semantic similarity assessments across multiple ontologies'. In: *Information Sciences* 283 (2014), pp. 197–210.

[1] https://github.com/lasigeBioTM/DiShIn/

[2] Temple F Smith and Michael S Waterman. 'Identification of common molecular subsequences'. In: *Journal of molecular biology* 147.1 (1981), pp. 195–197.

[3] Peter Willett. 'Similarity searching using 2D structural fingerprints'. In: *Chemoinformatics and computational chemical biology* (2011), pp. 133–158.

[4] S.F. Altschul et al. 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs'. In: *Nucleic acids research* 25.17 (1997), pp. 3389–3402.

[5] Gregory A Petsko and Dagmar Ringe. 'Protein structure and function'. In: New Science Press, 2004.

[6] Bhupendra S Gupta and Uma Gupta. *Caffeine and Behavior: Current Views & Research Trends: Current Views and Research Trends*. CRC Press, 1999.

[7] Marcia Barros and Francisco M Couto. 'Knowledge Representation and Management: a linked data perspective'. In: *IMIA Yearbook* (2016), pp. 178–183.

[8] A. Tversky. 'Features of similarity.' In: *Psychological review* 84.4 (1977), p. 327.

[9] David Sánchez and Montserrat Batet. 'Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective'. In: *Journal of biomedical informatics* 44.5 (2011), pp. 749–759.

[10] P. Resnik. 'Using information content to evaluate semantic similarity in a taxonomy'. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (1995).

[11] Sheldon Ross. *A First Course in Probability 8th Edition*. Pearson, 2009.

[12] F.M. Couto and M.J. Silva. 'Disjunctive shared information between ontology concepts: application to Gene Ontology'. In: *Journal of Biomedical Semantics* 2 (2011), p. 5.

[13] João D Ferreira, Janna Hastings and Francisco M Couto. 'Exploiting disjointness axioms to improve semantic similarity measures'. In: *Bioinformatics* 29.21 (2013), pp. 2781–2787.

[14] P.W. Lord et al. 'Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation'. In: *Bioinformatics* 19.10 (2003), pp. 1275–1283.

[15] Dekang Lin. 'An Information-Theoretic Definition of Similarity'. In: *Proceedings of the Fifteenth International Conference on Machine Learning*. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 296–304. ISBN: 1-55860-556-8. URL: http://dl.acm.org/citation.cfm?id=645527.657297.

[16] Albert Solé-Ribalta et al. 'Towards the estimation of feature-based semantic similarity using multiple ontologies'. In: *Knowledge-Based Systems* 55 (2014), pp. 101–113.

4

MER: a Shell Script and Annotation Server for Minimal Named Entity Recognition and Linking

FRANCISCO M. COUTO AND ANDRE LAMURIAS

Abstract

Named-Entity Recognition aims at identifying the fragments of a given text that mention a given entity of interest, that afterwards could be linked to a knowledge base where that entity is described.

This manuscript presents our Minimal Named-Entity Recognition and Linking tool (MER), designed with flexibility, autonomy and efficiency in mind. To annotate a given text, MER only requires: i) a lexicon (text file) with the list of terms representing the entities of interest; ii) optionally a tab-separated values file with a link for each term; iii) and a Unix shell. Al-

ternatively, the user can provide an ontology from where MER will automatically generate the lexicon and links files. The efficiency of MER derives from exploring the high performance and reliability of the text processing command-line tools `grep` and `awk`, and a novel inverted recognition technique.

MER was deployed in a cloud infrastructure using multiple Virtual Machines to work as an annotation server and participate in the Technical Interoperability and Performance of annotation Servers (TIPS) task of BioCreative V.5. The results show that our solution processed each document (text retrieval and annotation) in less than 3 seconds on average without using any type of cache. MER was also compared to a state-of-the-art dictionary lookup solution obtaining competitive results not only in computational performance but also in precision and recall.

MER is publicly available in a GitHub repository (`https://github.com/lasigeBioTM/MER`) and through a RESTful Web service (`http://labs.fc.ul.pt/mer/`).

4.1 Introduction

Text has been and continues to be for humans the traditional and natural mean of representing and sharing knowledge. However, the information encoded in free text is not easily attainable by computer applications. Usually, the first step to untangle this information is to perform Named-Entity Recognition (NER), a text mining task for identifying mentions of entities in a given text [1, 2]. The second step is linking these mentions to the most appropriate entry in a knowledge base. This last step is usually referred to as the Named-Entity Linking (NEL) task but is also referred to as entity disambiguation, resolution, mapping, matching or even grounding [3].

State-of-the-art NER and NEL solutions are mostly based on machine learning techniques, such as Conditional Random Fields and/or Deep Learning [4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. These solutions usually require as input a training corpus, which consists of a set

of texts and the entities mentioned on them, including their exact location (annotations), and the entries in a knowledge base that represent these entities [14]. The training corpus is used to generate a model, which will then be used to recognize and link entities in new texts. Their effectiveness strongly depends on the availability of a large training corpus with an accurate and comprehensive set of annotations, which is usually arduous to create, maintain and extend. On the other hand, dictionary lookup solutions usually only require as input a lexicon consisting in a list of terms within some domain [15, 16, 17, 18, 19, 20], for example, a list of names of chemical compounds. The input text is then matched against the terms in the lexicon mainly using string matching techniques. A comprehensive lexicon is normally much easier to find or to create and update than a training corpus, however, dictionary lookup solutions are generally less effective than machine learning solutions.

Searching, filtering and recognizing relevant information in the vast amount of literature being published is an almost daily task for researches working in Life and Health Sciences [21]. Most of them use web tools, such as PubMed [22], but many times to perform repetitive tasks that could be automatized. However, these repetitive tasks are sometimes sporadic and highly specific, depending on the project the researcher is currently working on. Therefore, in these cases, researchers are reluctant to spend resources creating a large training corpus or learning how to adapt highly complex text mining systems. They are not interested in getting the most accurate solution, just one good enough tool that they can use, understand and adapt with minimal effort. Dictionary lookup solutions are normally less complex than machine learning solutions, and a specialized lexicon is usually easier to find than an appropriate training corpus. Moreover, dictionary lookup solutions are still competitive when the problem is limited to a set of well-known entities. For these reasons, dictionary lookup solutions are usually the appropriate option when good enough is what the user requires.

This manuscript proposes a novel dictionary lookup solution, dubbed as Minimal named-Entity Recognizer (MER), which was designed with flexibility, autonomy, and efficiency in

mind. MER only requires as input a lexicon in the form of a text file, in which each line contains a term representing a named-entity to recognize. If the user also wants to perform entity linking, a text file containing the terms and their respective Unique Resource Identifiers (URIs) can also be given as input. Therefore, adding a new lexicon to MER could not be easier than this. MER also accepts as input an ontology in Web Ontology Language (OWL) format, which it converts to a lexicon.

MER is not only minimal in terms of the input but also in its implementation, which was reduced to a minimal set of components and software dependencies. MER is then composed of just two components, one to process the lexicon (offline) and another to produce the annotations (on-the-fly). Both were implemented as a Unix shell script [23], mainly for two reasons: i) efficiency, due to its direct access to high-performance text and file processing tools, such as `grep` and `awk`, and a novel inverted recognition technique; and ii) portability, since terminal applications that execute Unix shell scripts are nowadays available in most computers using Linux, macOS or Windows operating systems. MER was tested using the Bourne-Again shell (bash) [24] since it is the most widely available. However, we expect MER to work in other Unix shells with minimal or even without any modifications.

We deployed MER in a cloud infrastructure to work as an annotation server and participate in the Technical Interoperability and Performance of annotation Servers (TIPS) task of BioCreative V.5 [25]. This participation allowed us to assess the flexibility, autonomy, and efficiency of MER in a realistic scenario. Our annotation server responded to the maximum number of requests (319k documents) and generated the second highest number of total predictions (7130k annotations), with an average of 2.9 seconds per request.

To analyze the statistical accuracy of MER's results we compared it against a popular dictionary lookup solution, the Bioportal annotator [26], using a Human Phenotype Ontology (HPO) gold-standard corpus [27]. MER obtained the highest precision in both NER and NEL tasks, the highest recall in NER, and a lower processing time. We also compared MER to Aho-corasick [28], a well-known string search algorithm, where it obtained a lower

processing time and better evaluation scores on the same corpus.

MER is publicly available in a GitHub repository [29], along with the code used to run the comparisons to other systems. The repository contains a small tutorial to help the user start using the program and test it. The remainder of this article will detail the components of MER, and how it was incorporated in the annotation server. We end by analyzing and discussing the evaluation results and present future directions.

4.2 MER

4.2.1 Input

Before being able to annotate any text, MER requires as input a lexicon containing the list of terms to match. The user can provide the lexicon as text file (.txt) where each line represents a term to be recognized. Additionally, to perform NEL a tab-separated values file (.tsv) is required. This links file have to contain two data elements per line: the term and the link. Alternatively, the user can provide an ontology (.owl) and MER will automatically parse it to create the lexicon and links files. So if, for example, we want to recognize terms that are present in ChEBI [30], the user can provide the whole ontology (*chebi.owl*) or just collect the relevant labels and store them in a text file, one label per line. Figure 4.1 presents an example where four ChEBI compounds are represented by a list of terms based on their ChEBI's name.

If the user provides an ontology, MER starts by retrieving all the values of the tags *rdfs:label*, *oboInOwl:hasRelatedSynonym* and *oboInOwl:hasExactSynonym* inside each top-level *owl:Class*. The values are then stored in two files: a regular lexicon with a label (term) per line; and a tab-separated values file with a pair term and respective identifier (URI) per line. The links file is then sorted and will be used by MER to perform NEL. Figures 4.2, 4.3 and 4.4 show a snippet of the links files generated for ChEBI ontology [31], HPO [32, 33], and Human Disease Ontology (DOID) [34, 35], respectively.

```
α-maltose
nicotinic acid
nicotinic acid D-ribonucleotide
nicotinic acid-adenine dinucleotide phosphate
```

Figure 4.1: Example of the contents of a lexicon file representing four compounds.

```
zygadenine                          http://purl.obolibrary.org/obo/CHEBI_10130
zymosterol                          http://purl.obolibrary.org/obo/CHEBI_18252
zymosterol ester                    http://purl.obolibrary.org/obo/CHEBI_52322
zymosterol intermediate 1a          http://purl.obolibrary.org/obo/CHEBI_52388
zymosterol intermediate 1b          http://purl.obolibrary.org/obo/CHEBI_52615
```

Figure 4.2: A snippet of the contents of the links file generated with ChEBI

```
yellow nails            http://purl.obolibrary.org/obo/HP_0011367
yellow nodule           http://purl.obolibrary.org/obo/HP_0025554
yellow papule           http://purl.obolibrary.org/obo/HP_0025507
yellow skin             http://purl.obolibrary.org/obo/HP_0000952
yellow skin plaque      http://purl.obolibrary.org/obo/HP_0031360
```

Figure 4.3: A snippet of the contents of the links file generated with the Human Phenotype Ontology

```
zebrafish allergy                   http://purl.obolibrary.org/obo/DOID_0060517
zellweger syndrome                  http://purl.obolibrary.org/obo/DOID_905
zika fever                          http://purl.obolibrary.org/obo/DOID_0060478
zika virus congenital syndrome      http://purl.obolibrary.org/obo/DOID_0080180
zika virus disease                  http://purl.obolibrary.org/obo/DOID_0060478
```

Figure 4.4: A snippet of the contents of the links file generated with the Disease Ontology

The links file can also be created manually for a specific lexicon not generated from an ontology. Figure 4.5 presents the links file created for the lexicon file of Figure 4.1.

```
α-maltose
http://purl.obolibrary.org/obo/CHEBI_18167
nicotinic acid
http://purl.obolibrary.org/obo/CHEBI_15940
nicotinic acid d-ribonucleotide
http://purl.obolibrary.org/obo/CHEBI_15763
nicotinic acid-adenine dinucleotide phosphate
http://purl.obolibrary.org/obo/CHEBI_76072
```

Figure 4.5: Example of the contents of the links file representing compounds CHEBI:18167, CHEBI:15940, CHEBI:15763 and CHEBI:76072.

```
== one-word (...word1.txt) ====================
α.maltose
== two-word (...word2.txt) ====================
nicotinic acid
== more-words (...words.txt) ==================
nicotinic acid d.ribonucleotide
nicotinic acid.adenine dinucleotide phosphate
== first-two-words (...words2.txt) ============
nicotinic acid
nicotinic acid.adenine
```

Figure 4.6: Each block represents the content of each of the four files created after pre-processing the input file shown in Figure 4.1.

4.2.2 Inverted Recognition

To recognize the terms, a standard solution would be to apply `grep` directly to the input text. However, the execution time is proportional to the size of the lexicon, since each term of the lexicon will correspond to an independent pattern to match. To optimize the execution time, we developed the inverted recognition technique. The inverted recognition uses the words in the processed input text as patterns to be matched against the lexicon file. Since the number of words in the input text is much smaller than the number of terms in the lexicon, `grep` has much fewer patterns to match. For example, finding the pattern *nicotinic acid* in the two-word chemical lexicon created for TIPS is more than 100 times faster than using the

standard solution.

To perform the inverted recognition technique, MER splits the lexicon into three files containing the terms composed by one (one-word), two (two-word) and three or more words (more-words). The second step creates a fourth file containing the first two words (first-two-words) of all the terms in the more-words file. During the above steps, MER makes the following minor modifications to the terms: convert all text to lowercase; contiguous white spaces are replaced by one white space; full stops are removed; leading and trailing white spaces are removed; and all special characters are replaced by a full stop. Since some special characters may cause matching problems, MER assumes that all the special characters (characters that are not alphanumeric or a whitespace, for example, hyphens) can be matched by any other character, so these characters are replaced by a full stop, like in regular expressions. Figure 4.6 presents the contents of each of the four files created using the terms shown in Figure 4.1. Note that the word *acid-adenine* was replaced by *acid.adenine*, and the last file presents the first two words of each entry in the third file. Note also that all the above steps are performed offline and only once per lexicon.

The on-the-fly module of MER starts when the user provides a new input text to be annotated with a lexicon already pre-processed. The goal is to identify which terms of the lexicon are mentioned in the text. The first step of MER is to apply the same minor modifications to the input text as described above, but also remove stop-words, and words with less than 3 characters. This will result in a processed input text derived from the original one. Note that MER only recognizes direct matches, if lexical variations of the terms are needed, then they have to be added in the lexicon, for example by using a stemming algorithm. MER will then create two alternation patterns: i) one-word pattern, with all the words in the input text; and ii) two-word pattern, with all the consecutive pairs of words in the input text. Figure 4.7 shows an example of these two patterns.

Next, MER creates three background jobs to match the terms composed of: i) one word, ii) two words, and iii) three or more words. The one-word job uses the one-word pattern

```
α-maltose and nicotinic acid was found, but not
    nicotinic acid D-ribonucleotide

α.maltose|nicotinic|acid|found|nicotinic|acid|d.ribonucleotide

α.maltose nicotinic|acid found|nicotinic acid
    |nicotinic acid|found nicotinic|acid d.ribonucleotide
```

Figure 4.7: Example of a given sentence to be annotated (first line), and its one-word and two-word patterns created by MER.

```
0       9       α-maltose
14      28      nicotinic acid
48      62      nicotinic acid
48      79      nicotinic acid D-ribonucleotide
```

Figure 4.8: Output example of MER for the sentence in Figure 4.7 and the lexicon in Figure 4.1 without any links file

```
0       9       α-maltose
http://purl.obolibrary.org/obo/CHEBI_18167
14      28      nicotinic acid
http://purl.obolibrary.org/obo/CHEBI_15940
48      62      nicotinic acid
http://purl.obolibrary.org/obo/CHEBI_15940
48      79      nicotinic acid D-ribonucleotide http://purl.obolibrary.org/obo/CHEBI_15763
```

Figure 4.9: Output example of MER for the sentence in Figure 4.7, the lexicon in Figure 4.1, and the links file of Figure 4.5

to find matching terms in the one-word file. Similarly, for the two-word job, that uses the two-word pattern and file. The last job uses the two-word pattern to find matches in the two-first-word file, and the resulting matches are then used as a pattern to find terms in the more-words file. The last job is less efficient since it executes `grep` twice, however, the resulting list of matches with the two-first-word file is usually small, so the second execution is negligible. In the end, each job will create a list of matching terms that are mentioned in the input text.

```
acne
http://purl.obolibrary.org/obo/DOID_6543
asthma
http://purl.obolibrary.org/obo/DOID_2841
bronchitis
http://purl.obolibrary.org/obo/DOID_6132
chronic obstructive pulmonary disease
http://purl.obolibrary.org/obo/DOID_3083
COPD
http://purl.obolibrary.org/obo/DOID_3083
disease
http://purl.obolibrary.org/obo/DOID_4
gastroenteritis
http://purl.obolibrary.org/obo/DOID_2326
impetigo
http://purl.obolibrary.org/obo/DOID_8504
otitis media
http://purl.obolibrary.org/obo/DOID_10754
urinary tract infection
http://purl.obolibrary.org/obo/DOID_13148
```

Figure 4.10: Output example of MER for the abstracts with PubMed identifiers: 29490421 and 29490060, and the Human Disease Ontology

Since the processed input text cannot be used to find the exact position of the term, MER uses the list of matching terms to find their exact position in the original input text. MER uses `awk` to find the multiple instances of each term in the original input text. The `awk` tool has the advantage of working well with UTF-8 characters that use more than one byte, in opposition to `grep` that just counts the bytes to find the position of a match. MER provides partial overlaps, i.e. a shorter term may occur at the same position as a longer one, but not full overlapping matches (same term in the same position). We also developed a test suite to refactor the algorithm with more confidence that nothing is being done incorrectly. The test suite is available in the GitHub repository branch dedicated to development [36].

Figure 4.8 shows the output of MER when using as input text the sentence in Figure 4.7, and the lexicon of Figure 4.1. Note that *nicotinic acid* appears twice at position 14 and 65, as expected, without affecting the match of *nicotinic acid D-ribonucleotide*.

4.2.3 Linking

If the links file is provided, then MER will try to find the recognized term in that file. This step is basically a `grep` at the beginning of each line in the file, and only returns the first exact match of each term. Figure 4.9 shows the output of MER when using the links file of Figure 4.5 that was missing in Figure 4.8. Figure 4.10 shows the output of MER for two abstracts using the Human Disease Ontology. Note that this functionality was implemented after our TIPS participation [37].

4.3 Annotation Server

TIPS is a novel task in BioCreative aiming at the evaluation of the performance of NER web servers, based on reliability and performance metrics. The entities to be recognized in TIPS were not restricted to a particular domain.

The web servers had to respond to single document annotation requests. The servers had to be able to retrieve the text from documents in the patent server, the abstract server and PubMed, without using any kind of cache for the text or for the annotations. The annotations had to be provided in, at least, one of the following formats: BeCalm JSON, BeCalm TSV, BioC XML or BioC JSON.

4.3.1 Lexicons

The first step to participate in TIPS was to select the data sources from which we could collect terms related with the following accepted categories: Cell line and cell type: Cellosaurus [38]; Chemical: HMDB [39], ChEBI [31] and ChEMBL [40]; Disease: Human Disease Ontology [34]; miRNA: miRBase [41]; Protein: Protein Ontology [42]; Subcellular structure: cellular component aspect of Gene Ontology [43]; Tissue and organ: tissue and organ subsets of UBERON [44].

```
  #terms     #words      #char #filename
  116616     137702    1027369 CELL_LINE_AND_CELL_TYPE.txt
  332167     446423   10397574 CHEMICAL.txt
   26216      92688     808366 DISEASE.txt
   73954      73954     991012 MIRNA.txt
  597867    1372326   11863642 PROTEIN.txt
    8146      26117     228167 SUBCELLULAR_STRUCTURE.txt
    5238      16283     126024 TISSUE_AND_ORGAN.txt
 1160204    2165493   25442154 total
```

Figure 4.11: Number of terms, words, and characters in the lexicons used in TIPS, obtained by using the following shell command: `wc -lmw *.txt`.

A post-extraction processing was applied to these data files, which consisted in lower-casing all terms, deleting leading and trailing white spaces and removing repeated terms. Since repeated annotations of different types were not allowed, we created another lexicon containing terms that appeared on more than one of the other lexicons. The terms matched to this lexicon were categorized as Unknown, as suggested by the organization. The software to extract the list of terms from the above data sources can be found in the GitHub repository branch dedicated to TIPS [36].

Figure 4.11 shows the number of terms, number of words, and number of characters of each lexicon created. Our Annotation Server was then able to recognize more than 1M terms composed of more than 2M words and more than 25M characters. All lexicons are available for reuse as a zip file in the TIPS branch of our repository [36].

4.3.2 Input and Output

We adapted MER to provide the annotations in the BeCalm TSV format. Thus, besides the input text and the lexicon, MER had also to receive the document identifier and the section as input. In Figure 4.12, the document identifier is 1 and section is A. The score column is calculated by $1 - 1/\ln(nc)$, where nc represents the number of characters of the recognized term. This assumes that longer terms are less ambiguous, and in that case, the

```
1 A 0   9  0.54488   α−maltose                          lexicon 1
1 A 14  28 0.621077  nicotinic acid                     lexicon 1
1 A 48  62 0.621077  nicotinic acid                     lexicon 1
1 A 48  79 0.708793  nicotinic acid D−ribonucleotide    lexicon 1
```

Figure 4.12: Output example of MER using BeCalm TSV format for the sentence in Figure 4.7 and the lexicon in Figure 4.1

match should have a higher confidence score. Note that MER only recognizes terms with three or more characters, so the minimum score is 0.08976 and the score is always lower than 1.

We used `jq` [45] a command-line JSON processor to parse the requests. The retrieval of each document was implemented using the popular `curl` tool, and we developed a specific parser for each data source to extract the text to be annotated. The parsers are also available at the TIPS branch [36].

4.3.3 Infrastructure

Our annotation server was deployed in a cloud infrastructure composed of three Virtual Machines (VM). Each VM had 8GB of RAM and 4 Intel Core CPUs @ 1.7 GHz, using CentOS Linux release 7.3.1611 as the operating system. We selected one VM (primary) to process the requests, distribute the jobs, and execute MER. The other two VMs (secondary) just execute MER. We installed the NGINX HTTP server running CGI scripts given its high performance when compared with other web servers [46]. We also used the Task Spooler [47] tool to manage and distribute within the VMs the jobs to be processed.

The server is configured to receive the REST API requests defined in the BeCalm API documentation. Each request is distributed to one of the three VMs according to the least-connected method of NGINX. When a *getAnnotations* request is received, the server first downloads the documents from the respective sources and then processes the title and abstract of each document in the same VM. Two jobs are spawned in background, correspond-

ing to the title and abstract. Each annotation server handles all the entity types mentioned in Figure 4.11, spawning a separate job for each entity type. The name of the entity type is added as another column to the output of Figure 4.8. These jobs can run in parallel since they are independent from each other and the output of each job can be easily merged into a final TSV output file. When a job finishes processing, a script checks if the other jobs associated with the same requests have also finished processing. If that is the case, then the results of every job are concatenated and sent back to BeCalm using the *saveAnnotations* method.

To test MER outside of the scope of the TIPS competition, we implemented a different REST API which accepts as input raw text and the name of a lexicon. This way, the document does not have to be retrieved from external sources, and we can evaluate the performance of MER independently. This alternative API can be accessed, along with a simple user interface [48] (Figure 4.13).

4.4 Results and Discussion

4.4.1 Computational Performance

Table 4.1: Official evaluation results of the TIPS task (time values are in seconds).

	MER	Best
# Requests	3.19E+05	3.19E+05
# Predictions	7.13E+06	2.74E+07
Mean time seek annotations (MTSA)	1.29E-01 s	1.37E-02 s
Mean time per document volume (MTDV)	2.38E-03 bytes/s	8.58E-04 bytes/s
Mean annotations per document (MAD)	2.25E+01	1.01E+02
Average response time (ART)	2.90E+00 s	1.07E+00 s
Mean time between failures (MTBF)	4.58E+06 s	4.58E+06 s
Mean time to repair (MTTR)	0.00E+00 s	0.00E+00 s

Table 4.1 shows the official TIPS evaluation data of our system [49]. These results refer to the whole period of the competition, from February, 5th 2017 to March, 30th 2017. The evaluation process and metrics used are described in the workshop article [25]. Each request

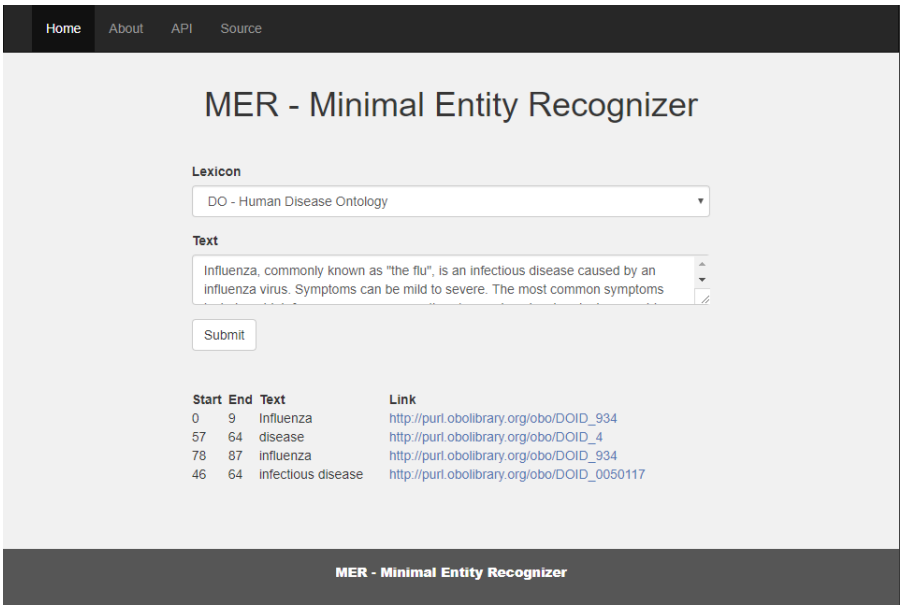


Figure 4.13: Screenshot of the MER web graphical user interface.

consisted of one document that the server had to retrieve either from PubMed or a repository hosted by the organization. Our server was able to handle all 319k requests received during the evaluation period, generating a total of 7.13M annotations (second best) with an average of 22.5 predictions per document (MAD) (third best). In average, each prediction has been generated in 0.129 s (MTSA). Our average processing time value (ART) was 2.9 s, and the processing time per document volume (MTDV) was 0.00238 bytes/s. The Mean time between failures (MTBF) and Mean time to repair (MTTR) metrics were associated with the reliability of server, and our team obtained the maximum scores on those metrics.

MER was able to efficiently process the documents by taking less than 3 seconds on average without using any type of cache. We note that all documents, irrespectively of the source, were annotated using all the entity types presented in the previous Lexicons section. Furthermore, the time to process each document is affected by external sources used to retrieve the document text. If the text is provided with the request, then the processing time should be considerably shorter. Another factor is the latency between our server and the TIPS server. As we were not able to measure this latency, it is difficult to measure the impact

on the response times, and it was not taken into consideration for the evaluation metrics.

We compared the time necessary to process the same sentence on the same hardware using MER and a more complex machine learning system, IBEnt [10], using the sentence of Figure 4.7. While IBEnt took 8.25 seconds to process the sentence, MER took only 0.098 seconds. Although IBEnt is optimized for batch processing, therefore reducing the time per document as the number of documents increases, MER is still 84 times faster than IBEnt in this experiment. Thus, besides being easy to install and configure, MER is also a highly efficient and scalable NER and NEL tool.

Part of the optimization of MER is due to 4 files that are generated during a pre-processing step. This files are generated from the lexicon file, which contains one entity per line. For NEL, there is another step necessary, which is to convert an owl file to a lexicon. This process took around 15 minutes for each ontology. However, processing a lexicon file is quite faster, taking 0.746 seconds for the HPO and 3.671 seconds for the ChEBI ontology.

4.4.2 Precision and Recall

Table 4.2: Comparison between MER and BioPortal on the HPO gold-standard corpus.

	NER			NER+NEL				
	P	R	F	P	R	F	ART	MTSA
BioPortal	0.6862	0.4463	0.5408	0.6118	0.3979	0.4822	1.15E+00 s	1.45E-01 s
MER	0.7184	0.4514	0.5544	0.6155	0.3868	0.4751	7.32E-01 s	9.59E-02 s

We compared the performance of MER with the BioPortal annotator, which is a popular dictionary lookup NER solution. To perform this comparison, we adapted our server to directly receive as input free text, instead of requiring another request to retrieve the documents. We used the HPO corpus to compare the two tools. This corpus is composed by 228 scientific abstracts annotated with human phenotypes, associated with the HPO. We used an updated version of this corpus, which aimed at improving the consistency of the annotations [50]. A total of 2773 textual named entities were annotated in this corpus, corresponding to 2170

unique entity mentions. We compared the quality of the results produced by each tool using the standard precision, recall and F1-score measures, as well as the time necessary to process each document on average (ART) and time per annotation (MTSA).

Table 4.2 shows the results of this comparison, where NER refers to matching the offsets of the automatic annotations with the gold standard, and NEL refers to matching the URI annotated automatically with the gold standard. As expected, combining both tasks (NER+NEL) results in lower scores than performing only NER. Using MER, the F1-score obtained was 0.5544, while BioPortal obtained an F1-score of 0.5408 on the NER task. Considering the NEL task too, BioPortal obtained a better F1-score than MER, indicating that some entities were linked to incorrect URIs. Bioportal annotator employs a semantic expansion technique that could lead to more accurate URI, using the relations defined in the ontology [51].

However, MER obtained lower response times than BioPortal, in terms of time per document and per annotation. To account for the difference in latency between the two servers, we used the `ping` tool to calculate the round-trip time of each server, averaged over 10 packets. MER obtained a round-trip time of 6.72E-03 s while BioPortal obtained 1.86E-01 s, representing a difference of 1.79E-01 s. This means that MER had a better connection to the machine we used to run the experiments, but this had minimal impact when comparing to a difference of 4.18E-01 s in both response times (ART).

We also compared MER with a well-known string search algorithm, Aho-corasick [28], on the HPO corpus. In this case, we did not attempt to match entities to ontology concepts as this would require additional enhancements to the algorithm. We used the same vocabulary with both methods, as well as the same documents, and, unlike the comparison to BioPortal, run a local installation of MER. We used the Makefast tool [52], which implements the Aho-corasick algorithm. MER obtained higher precision, recall and F1-score, as well as a lower processing time per document and per annotation (Table 4.3). MER obtained better evaluation scores since it was developed specifically for NER, while Aho-corasick is

a generic string search algorithm. The processing time was also shorter, possibly due to the pre-processing that is done by MER to the lexicon file. This pre-processing is quick (3.671 s for the HPO ontology) and only has to be done once.

Table 4.3: Comparison between MER and Aho-corasick on the HPO gold-standard corpus.

	NER				
	P	R	F	ART	MTSA
Aho-corasick	0.2282	0.2665	0.2459	0.8596	0.0786
MER	0.7184	0.4514	0.5544	0.5088	0.0667

4.5 Conclusions

We presented MER, a minimal named entity recognition and linking tool that was developed with the concepts of flexibility, autonomy, and efficiency in mind. MER is flexible since it can be extended with any lexicon composed of a simple list of terms and its identifiers (if available). MER is autonomous since it only requires a Unix shell with `awk` and `grep` command-line tools, which are nowadays available in all mainstream operating systems. MER is efficient since it takes advantage of the high-performance capacity of `grep` as a file pattern matcher, and by proposing a novel inverted recognition technique.

MER was integrated in an annotation server deployed in a cloud infrastructure for participating in the TIPS task of BioCreative V.5. Our server was fully developed in-house with minimal software dependencies and is open-source. Without using any kind of cache, our server was able to process each document in less than 3 seconds on average. Performance and quality results show that MER is competitive with state-of-the-art dictionary lookup solutions.

Availability of data and materials

The data and code used for this study are available at https://github.com/lasigeBioTM/MER.

List of abbreviations

NER - Named Entity Recognition

NEL - Named Entity Linking

URI - Unique Resource Identifier

OWL - Web Ontology Language

CHEBI - Chemical Entities with Biological interest

HPO - Human Phenotype Ontology

TIPS - Technical Interoperability and Performance of annotation Servers

MTSA - Mean Time Seek Annotations

MTDV - Mean Time per Document Volume

MAD - Mean Annotations per Document

ART - Average Response Time

MTBF - Mean Time Between Failures

MTTR - Mean Time To Repair

VM - Virtual Machine

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Conceptualization and methodology: FC and AL. Funding acquisition, project administration, and supervision: FC. Investigation, validation, writing, review, and editing: FC and AL. Software: FC and AL. Visualization and writing original draft: FC and AL.

Acknowledgements

This work was supported by FCT through funding of the LaSIGE Research Unit, ref. UID/CEC/00408/2013 and BioISI, ref. ID/MULTI/04046/2013. AL is recipient of a fellowship from BioSys PhD programme (ref PD/BD/106083/2015) from FCT (Portugal). This work was produced with the support of the Portuguese National Distributed Computing Infrastructure (http://www.incd.pt).

References

[1] David Nadeau and Satoshi Sekine. 'A survey of named entity recognition and classification'. In: *Lingvisticae Investigationes* 30.1 (2007), pp. 3–26.

[2] Martin Krallinger et al. 'Information retrieval and text mining technologies for chemistry'. In: *Chemical reviews* 117.12 (2017), pp. 7673–7761.

[3] Maryellen C MacDonald, Neal J Pearlmutter and Mark S Seidenberg. 'The lexical nature of syntactic ambiguity resolution'. In: *Psychological review* 101.4 (1994), p. 676.

[4] Chen-Kai Wang et al. 'An Ensemble Algorithm for Sequential Labelling: A Case Study in Chemical Named Entity Recognition'. In: *Proceedings of the BioCreative V. 5 Challenge Evaluation Workshop*. 2017.

[5] Cristóbal Colón-Ruiz, Isabel Segura-Bedmar and Paloma Martınez. 'Combining the BANNER tool with the DINTO ontology for the CEMP task of BioCreative V. 5'. In: *Proceedings of the BioCreative V. 5 Challenge Evaluation Workshop*. 2017.

[6] Robert Leaman and Zhiyong Lu. 'Towards Robust Chemical Recognition with TaggerOne at the BioCreative V. 5 CEMP Task'. In: *Proceedings of the BioCreative V. 5 Challenge Evaluation Workshop*. 2017.

[7] Yuankai Guo et al. 'Recognition of Chemical Entity Mention in Patents Using Feature-rich CRF'. In: *Proceedings of the BioCreative V. 5 Challenge Evaluation Workshop*. 2017.

[8] André Santos and Sérgio Matos. 'Neji: Recognition of Chemical and Gene Mentions in Patent Texts'. In: *Proceedings of the Biocreative V. 5 Challenge Evaluation Workshop*. 2017.

[9] Zengjian Liu et al. 'HITextracter System for Chemical and Gene/Protein Entity Mention Recognition in Patents'. In: *Proceedings of the Biocreative V. 5 Challenge Evaluation Workshop*. 2017.

[10] A. Lamurias, L. Campos and F. Couto. 'IBEnt: Chemical Entity Mentions in Patents using ChEBI'. In: *BioCreative V.5 Challenge Evaluation*. 2017. URL: http://www.biocreative.org/media/store/files/2017/BioCreative_V5_paper12.pdf.

[11] Ling Luo et al. 'DUTIR at the BioCreative V. 5. BeCalm tasks: A BLSTM-CRF approach for biomedical entity recognition in patents'. In: *Proceedings of the Biocreative V. 5 Challenge Evaluation Workshop*. 2017.

[12] P Corbett and J Boyle. 'Chemlistem-chemical named entity recognition using recurrent neural networks'. In: *Proceedings of the BioCreative V. 5 Challenge Evaluation Workshop*. 2017.

[13] Hong-Jie Dai et al. 'Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization'. In: *Journal of cheminformatics* 7.S1 (2015), S14.

[14] Martin Krallinger et al. 'The CHEMDNER corpus of chemicals and drugs and its annotation principles'. In: *Journal of cheminformatics* 7.1 (2015), S2.

[15] Evangelos Pafilis et al. 'EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation'. In: *Database* 2016 (2016).

[16] Johannes Kirschnick and Philippe Thomas. 'SIA: Scalable Interoperable Annotation Server'. In: (2017).

[17] Jitendra Jonnagaddala et al. 'Improving the dictionary lookup approach for disease normalization using enhanced dictionary and query expansion'. In: *Database* 2016 (2016).

[18] Milena Kraus et al. 'Olelo: a web application for intuitive exploration of biomedical literature'. In: *Nucleic acids research* 45.W1 (2017), W478–W483.

[19] Fabio Rinaldi et al. 'OntoGene web services for biomedical text mining'. In: *BMC Bioinformatics* 15.14 (2014), S6.

[20] Andrew MacKinlay and Karin Verspoor. 'A web service annotation framework for CTD using the UIMA concept mapper'. In: *BioCreative challenge evaluation workshop vol*. Vol. 1. 2013.

[21] Carol Tenopir and Donald W King. 'Reading behaviour and electronic journals'. In: *Learned Publishing* 15.4 (2002), pp. 259–265.

[22] David L Wheeler et al. 'Database resources of the national center for biotechnology information'. In: *Nucleic acids research* 35.suppl_1 (2006), pp. D5–D12.

[23] Cameron Newham and Bill Rosenblatt. *Learning the bash shell: Unix shell programming*. " O'Reilly Media, Inc.", 2005.

[24] *Bash download page*. https://ftp.gnu.org/gnu/bash/. (Accessed on 11/06/2018).

[25] Martin Pérez Perez et al. 'Benchmarking biomedical text mining web servers at BioCreative V.5: the technical Interoperability and Performance of annotation Servers - TIPS track'. In: *Proceedings of the BioCreative V.5 Challenge Evaluation Workshop*. 2017.

[26] Patricia L Whetzel et al. 'BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications'. In: *Nucleic acids research* 39.suppl_2 (2011), W541–W545.

[27] Tudor Groza et al. 'Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora'. In: *Database* 2015 (2015), pp. 1–13. ISSN: 17580463. DOI: 10.1093/database/bav005.

[28] Alfred V. Aho and Margaret J. Corasick. 'Efficient String Matching: An Aid to Bibliographic Search'. In: *Commun. ACM* 18.6 (June 1975), pp. 333–340. ISSN: 0001-0782. DOI: 10.1145/360825.360855. URL: http://doi.acm.org/10.1145/360825.360855.

[29] *MER Source code*. https://github.com/lasigeBioTM/MER. (Accessed on 11/06/2018).

[30] Kirill Degtyarenko et al. 'ChEBI: a database and ontology for chemical entities of biological interest'. In: *Nucleic acids research* 36.suppl_1 (2007), pp. D344–D350.

[31] *ChEBI ontology*. ftp://ftp.ebi.ac.uk/pub/databases/chebi/ontology/chebi_lite.owl. (Accessed on 11/06/2018).

[32] *Human Phenotype Ontology*. https://raw.githubusercontent.com/obophenotype/human-phenotype-ontology/master/hp.owl. (Accessed on 11/06/2018).

[33] Sebastian Köhler et al. 'The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data'. In: *Nucleic acids research* 42.D1 (2013), pp. D966–D974.

[34] *Disease Ontology*. https://raw.githubusercontent.com/DiseaseOntology/HumanDiseaseOntology/master/src/ontology/doid.owl. (Accessed on 11/06/2018).

[35] Warren A Kibbe et al. 'Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data'. In: *Nucleic acids research* 43.D1 (2014), pp. D1071–D1078.

[36] *MER Source code for BioCreative V.5 2017*. https://github.com/lasigeBioTM/MER/tree/biocreative2017. (Accessed on 11/06/2018).

[37] Francisco M Couto, Luis F Campos and Andre Lamurias. 'MER: a minimal named-entity recognition tagger and annotation server'. In: (2017).

[38] *ExPASy - Cellosaurus*. https://web.expasy.org/cellosaurus/. (Accessed on 11/06/2018).

[39] David S Wishart et al. 'HMDB 4.0: the human metabolome database for 2018'. In: *Nucleic Acids Research* 46.D1 (2018), pp. D608–D617. DOI: 10.1093/nar/gkx1089. URL: http://dx.doi.org/10.1093/nar/gkx1089.

[40] Anna Gaulton et al. 'The ChEMBL database in 2017'. In: *Nucleic Acids Research* 45.D1 (2017), pp. D945–D954. DOI: 10.1093/nar/gkw1074. URL: http://dx.doi.org/10.1093/nar/gkw1074.

[41] Ana Kozomara and Sam Griffiths-Jones. 'miRBase: annotating high confidence microRNAs using deep sequencing data'. In: *Nucleic Acids Research* 42.D1 (2014), pp. D68–D73. DOI: 10.1093/nar/gkt1181. URL: http://dx.doi.org/10.1093/nar/gkt1181.

[42] *PRotein Ontology (PRO)*. http://www.obofoundry.org/ontology/pr.html. (Accessed on 11/06/2018).

[43] Gene Ontology Consortium. 'Expansion of the Gene Ontology knowledgebase and resources'. In: *Nucleic acids research* 45.D1 (2016), pp. D331–D338.

[44] Melissa A Haendel et al. 'Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon'. In: *Journal of biomedical semantics* 5.1 (2014), p. 21.

[45] *[jq*. https://stedolan.github.io/jq/. (Accessed on 11/06/2018).

[46] Will Reese. 'Nginx: the high-performance web server and reverse proxy'. In: *Linux Journal* 2008.173 (2008), p. 2.

[47] Lluís Batlle i Rossell. *Task Spooler - batch is back!* http://vicerveza.homeunix.net/~viric/soft/ts/. (Accessed on 11/06/2018).

[48] *MER*. http://labs.rd.ciencias.ulisboa.pt/mer/. (Accessed on 11/06/2018).

[49] Martin Pérez-Pérez et al. 'Next generation community assessment of biomedical entity recognition web servers: metrics, performance, interoperability aspects of BeCalm'. In: *Journal of Cheminformatics* (2018).

[50] M. Lobo, A. Lamurias and F. Couto. 'Identifying Human Phenotype Terms by Combining Machine Learning and Validation Rules'. In: *BioMed Research International* 2017 (2017). ISSN: 2314-6133. DOI: https://doi.org/10.1155/2017/8565739.

[51] Nigam H Shah et al. 'Comparison of concept recognizers for building the Open Biomedical Annotator'. In: *BMC bioinformatics*. Vol. 10. 9. BioMed Central. 2009, S14.

[52] *MultiFast 2.0.0*. http://multifast.sourceforge.net/. (Accessed on 11/06/2018).

5

PPR-SSM: Personalized PageRank using Semantic Similarity Measures for Entity Linking

ANDRE LAMURIAS, LUKA A CLARKE, FRANCISCO M COUTO

Abstract

Entity linking is a text mining task that aims at linking entities mentioned in documents to concepts in a knowledge base. Existing approaches focus on the local similarity of each entity and the global coherence of all entities, but do not take into account the semantics of the domain. In this manuscript, we propose a method to link entities found in documents to concepts from domain-specific ontologies. Our method is based on Personalized PageRank (PPR), using the relations of the ontology to generate a graph of candidate concepts for the entities of the text. We demonstrate how the knowledge encoded in a domain-specific

ontology can be used to calculate the coherence of a set of candidate concepts. We show how this method can be used to effectively link named entities to biomedical ontologies, outperforming a string matching method and other PPR-based methods.

5.1 Introduction

Entity linking matches each entity mention in a document to an entry of a knowledge base (KB) that unequivocally represents that concept [1, 2]. This task is a fundamental component of information extraction systems, in order to integrate the information described in the literature across multiple documents [3]. While several biomedical Named Entity Recognition (NER) approaches have been developed to recognize, for example, genes, drugs and diseases entities in documents [4, 5], fewer approaches exist to link these entities to a KB, given its complexity.

In biomedicine, ontologies are commonly used to organize knowledge about a specific domain, providing a formal representation of concepts and their relations according to the domain. As such, they can be used as reference KBs for text mining tasks such as entity linking [6, 7]. For example, an ontology enables us to calculate the semantic similarity between two concepts and compare which concepts have more in common. Therefore, this source of information can be incorporated into entity linking approaches to improve their performance.

Entity linking is a challenging task for biomedical literature when compared to other domains. For example, while there is no exact match for "iron chloride" in ChEBI, a database of chemical entities with biological interest [8], there are 157 abstracts on PubMed that match that exact string at the time we were writing this manuscript. These cases are problematic to automatic approaches because the entity string itself is ambiguous, requiring more advanced approaches to resolve this ambiguity. According to the Human Phenotype Ontology (HPO), dyschromatopsia and color-blindness refer to the same phenotype. Therefore a search for

one of those names should retrieve documents that also mention the other one. Another example, a protein may be mentioned by its full name or by an acronym; in this case, the normalization process should assign the same identifier to both occurrences. To properly perform biomedical entity linking, it is necessary to account for these issues, as well as with the constant flow of newly published information.

PageRank is a graph-based algorithm initially developed to rank web pages for search results [9]. An adaptation of this algorithm, Personalized PageRank (PPR) [10], has been successfully applied to Word Sense Disambiguation [11] and Named Entity Disambiguation [12]. The PPR algorithm, which we make use of in this work, is based on random walks along the graph, with a given probability of jumping to a specific source node.

Our main contribution is a novel domain-specific ontology-based entity linking method for documents annotated with named entities that can be applied to various domains. Our method uses the PPR algorithm on a graph obtained from the relations established in the ontology, exploring the semantic similarity between the candidate matches of each entity to maximize the global coherence. We applied this method to two gold standards: i) one annotated with chemical entities and ii) another annotated with human phenotypes. We used the ChEBI and HPO ontologies as a reference in the chemical and phenotype gold standards, respectively. This method outperformed string matching and other PPR approaches. We also studied the effect of different semantic similarity measures in the results. We provide the code used in the experiments as well as the results on two gold standards[1].

The rest of the paper is organized as follows: a survey of related works is summarized in Section 5.2; the details of the proposed model are described in Section 5.3; Section 5.4 presents the experimental results and error analysis; Section 5.5 concludes the study.

[1]Link provided upon acceptance.

5.2 Related work

Previous studies follow mainly two types of approaches: local similarity approaches, where the similarity between the entity text and candidate match is explored, and global approaches, which attempt at selecting the set of candidate matches that best represents the entities of a document [13, 14]. One of the most commonly used KBs for entity linking is the Wikipedia, which contains information about a great variety of topics. For this reason, it can be used to map entities of different domains to a KB. This variety of topics also increases the difficulty of the task, since the same expression can have different meanings according to its context. The disambiguation pages show the diverse meanings that an expression may have. For example, "New York" can refer to the state, the city in the state of New York, cities in other states, works of art, sports teams and ship names.

Bunesco et al. presented a method based on Support Vector Machines, using a dictionary generated from the Wikipedia to detect and link entities [15] . Other authors aimed at maximizing the global coherence between the linked entities [13, 16, 17]. Pershina et al. presented a graph-based method based on the Personalized PageRank (PPR) algorithm to this task, incorporating both local and global coherence [18]. They assumed that the probability of each node is related to how likely it is to fit with the other highest scoring nodes. More recently, Radhakrishnan et al. presented a method that improved entity similarity by training embedding vectors on a densified KB [14]. Since the majority of entity linking gold standards are based on the Wikipedia, these systems are designed to that specific KB, and rarely focus on generalization to other KBs.

5.2.1 Graph-based approaches

Several graph-based approaches have been proposed for entity linking. [19] developed a graph-based framework to rank the entries of a database according to their relevance to a query. [20] proposed a method to rank the concepts and relations of an ontology according

to their importance to the domain. Although this method is helpful to understand a domain better using ontologies, the authors did not explore its utility for other text mining tasks. [21] explored Markov networks for entity linking, applied to citation databases. These types of approaches require training data, which is not always available, particularly in some biomedical domains. Unlike other authors that explored graph-based methods for entity linking, we propose a method that takes advantage of the semantic relations described in the ontology.

5.2.2 Biomedical entity linking

The Wikipedia as a KB for entity linking has two properties that are useful for this task: redirect pages, which account for synonyms and lexical variations; and disambiguation pages, which account for strings with multiple meanings. While biomedical ontologies can incorporate synonyms, there is no equivalent to disambiguation pages. When such ambiguity arises, it is necessary to understand the context of the sentence to determine the correct definition.

The gene normalization task of BioCreative consisted in determining the unique identifiers of genes and proteins mentioned in scientific articles [22, 23]. The objective of this task, as with the other BioCreative tasks, was to promote the development of new text mining methods specifically for biomedical text. The organizers selected and manually annotated articles with gene names, using Entrez Gene as reference. Three editions of this task were organized, each edition increasing the difficulty, with the final edition requiring the full-text annotation and being species non-specific. The gold standards developed for this task were made available and can then be used to benchmark new methods. Tsuruoka et al. [24] presented a method to develop heuristic rules for biomedical entity linking automatically. Their method obtained better computational performance than string matching while requiring minimal expert knowledge in the development of the rules.

A domain-specific ontology can be defined as a directed acyclic graph where each node is a concept of the domain and the edges represent known relations between these con-

cepts [25]. This definition is the traditional representation of existing biomedical ontologies, which are nowadays a mainstream approach to formalize knowledge about entities, such as genes, chemicals, phenotypes, and disorders. Biomedical ontologies are usually publicly available and cover a large variety of topics of Life and Health Sciences. The success of exploring a given biomedical ontology for performing a specific task can be easily extended to other topics due to the standard structure of biomedical ontologies. For example, the same measures of metadata quality have been successfully applied to resources annotated with different biomedical ontologies [26]. Our method combines the advantages of PPR-based methods that do not require training data, with domain knowledge from biomedical ontologies. Therefore, it can be adapted for other domains, as long as there is an exhaustive and domain-specific ontology available.

5.3 Methods

5.3.1 Problem definition

We now define the concepts necessary to understand the entity linking problem and our proposed solution. We consider the problem setting where a corpus of documents is annotated with entity mentions, and each entity mention has a set of KB candidate matches. The objective of entity linking is to link each entity mention to an entry of a KB. We can define a KB as a tuple $< C, R >$, where C is the set of concepts about a particular subject, and R the set of relations between the concepts, where each relation is a pair of concepts (c_1, c_2) with $c_1, c_2 \in C$[2]. We consider a candidate list $CL(e) = \{c_e^1, ..., c_e^i\}$ for each entity $e \in E$, where E is the set of named entities mentioned in a document. We want to find the $c_e \in CL(e)$ that best represents each e.

[2]We use typewriter to indicate a concept of a KB , italics to indicate relations between concepts, and quotes to indicate entity mentions.

For each document, we can construct a graph G consisting of the edges defined by:

$$G = \{(e, c_e)|e \in E, c_e \in CL(e)\}$$

where e corresponds to each named entity of a document and c_e to each candidate match of that entity. This graph is based only on the relations established in a document. Our objective is to define a function $disambiguate$ such that

$$disambiguate(e) = \arg\max_{c_e}\{score(e, c_e)\}$$

where $score$ is a scoring function that evaluates how likely the candidate is to be the correct choice for entity e.

5.3.2 Ontology-based Personalized PageRank

We assume that a measure of global coherence among the candidate concepts could be used as a scoring function. This idea has been explored by other authors, who suggest random walks methods such as Personalized PageRank (PPR) to rank the importance of each node in a graph. Nodes with greater weight would be more relevant to the results. The weights are determined by simulating random walks on the graph, with a certain probability of jumping to a random node. The PPR algorithm is a variant of PageRank where the jump is always made to the same node. Using the graph previously described, we apply the PPR algorithm to calculate the weights of each node in relation to each other, which we use as a coherence score.

Note that in our graph model, each node represents a candidate concept associated with a named entity. Therefore, we consider only edges between nodes associated with different entities, since only one element of each candidate list can be correct. Our approach to entity linking explores the structure of the ontology to generate the graph. If a node is within distance d of another node, we consider that they are linked. To calculate this distance, we

do not take into consideration the directionality of the relations of the ontology. Therefore, any two nodes of the same document can form an edge as long as there is a path with length equal to or shorter than d between them, and they are associated with different entities.

Figure 5.1 shows an example of the graph generated by a set of named entities from an abstract annotated with HPO concepts. To simplify the figure, we show only three entities and the two highest scoring candidates of each entity. We considered $d = 6$ for this example. Due to its spelling similarity, `tremor` is a candidate match to the entity "tumour", when in fact the correct match should be `neoplasm`. Note that `neoplasm` is a candidate for that entity because HPO has `tumour` as a synonym of `neoplasm`. The candidate `tremor` is linked only to one other candidate, while `tumour` is linked to candidates from both entities. Hence, `neoplasm` is more likely to maximize the global coherence. Likewise, `Abnormality of the nervous system` is linked only to one candidate, so it will have a negative contribution to the global coherence. Both candidates of the entity "neurofibromatosis" are linked to the same candidates. In these cases, we adopt a conservative approach and pick the candidate with more descendants in the ontology, since it represents a more generic concept. Therefore, `neurofibromas` would be the chosen candidate for that entity.

The PPR algorithm is used to calculate the coherence of each node in relation to another node, which can also be interpreted as the PageRank score. To accomplish this, we personalized the graph to each node, referred to as the source node. We estimate the coherence of node n to source node s, given by $PPR(s \rightarrow n)$, corresponding to the weight of n when personalizing to s. We multiply the PPR score by the normalized information content (IC) value of the concept associated with node n, in order to account for the different degrees of specificity of the concepts of an ontology. Therefore we calculate the coherence of node n relative to node s as

$$coherence_s(n) = PPR(s \rightarrow n) \cdot IC(n) \tag{5.1}$$

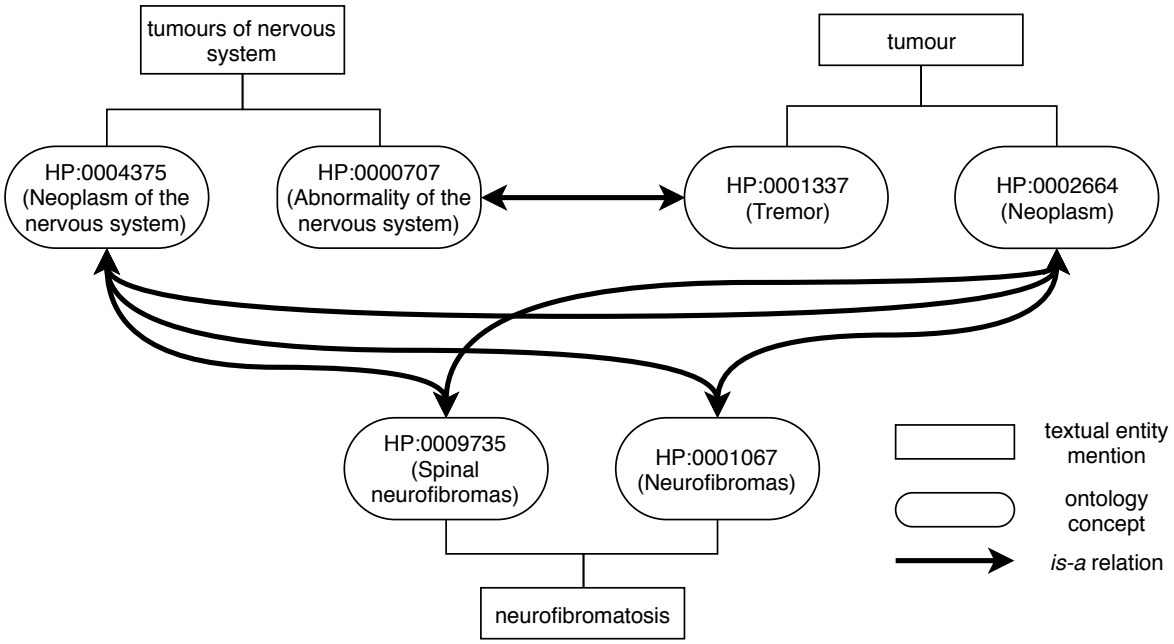


Figure 5.1: Example of the graph generated from abstract PMID2888021 using HPO.

We estimate IC of a node n as:

$$IC(n) = -\log(p(n)) \tag{5.2}$$

where $p(n)$ is the probability of that node appearing on a corpus [27].

Finally, we sum the coherence score of node n to each source node s to estimate its global coherence:

$$coherence(n) = \sum_{s \in G} coherence_s(n) \tag{5.3}$$

5.3.3 Semantic similarity

Semantic similarity measures (SSMs) estimate the similarity between concepts using the relations defined by an ontology [28]. The semantic similarity between concepts can improve the graph model previously described by adjusting the contribution of each node to another

node. If two nodes share more semantics, they should have a greater contribution to each other's global coherence score.

SSMs are normally restricted to subsumption relations (*is-a* or *subClassOf*), which are transitive, meaning that if R is the set of relations between concepts, $(c_1, c_2) \in R$, and $(c_2, c_3) \in R$, then $(c_1, c_3) \in R$. Therefore, the ancestors of c are given by

$$Anc(c) = \{a : (c, a) \in T\}$$

where T is the smallest relation set on C that contains R and is transitive.

Many SSMs explore the ancestors exclusive to each concept, as well as their common ancestors. We can define the common ancestors CA between two concepts as

$$CA(c_1, c_2) = Anc(c_1) \cap Anc(c_2)$$

Some SSMs use only the most informative common ancestors (MICA), which can be considered the most relevant to compare entities:

$$MICA(c_1, c_2) = \{a : a \in CA(c_1, c_2) \wedge IC(a) = \max\{IC(a) : a \in CA(c_1, c_2)\}\} \quad (5.4)$$

Alternatively, SSMs can consider multiple inheritance relations, which we refer to disjunctive common ancestors (DCA):

$$DCA(c_1, c_2) = \{a : a \in CA(c_1, c_2) \wedge \forall_{a_x \in CA(c_1, c_2)} PD(c_1, c_2, a) = PD(c_1, c_2, a_x) \Rightarrow IC(a) > IC(a_x)\} \quad (5.5)$$

where PD is a function that calculates the difference between the number of paths of c_1 and c_2 to their common ancestors.

SSMs can use the IC of the concepts to estimate its similarity. Several measures have

been proposed, one of the most commonly used being the measure proposed by Resnik [27]:

$$SSM_{Resnik}(c_1, c_2) = IC_{shared}(c_1, c_2) \tag{5.6}$$

where IC_{shared} is the average of the information content of the MICA or DCA.

Another SSM was proposed by Lin et al. [29], which balances the IC of the common ancestors with the IC of the concepts themselves:

$$SSM_{Lin}(c_1, c_2) = \frac{2 \times IC_{shared}(c_1, c_2)}{IC(c_1) + IC(c_2)} \tag{5.7}$$

Finally, Jiang and Conrath [30] proposed a measure of distance between concepts of an ontology, given by

$$dist_{jc}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times IC_{shared}(c_1, c_2) \tag{5.8}$$

As an SSM should be inversely proportional to the distance (i.e. less distance, more similar), we can use this distance to calculate a semantic similarity score:

$$SSM_{jc}(c_1, c_2) = \begin{cases} \frac{1}{dist(c1,c2)}, \text{if } dist > 0 \\ 1, \text{otherwise} \end{cases} \tag{5.9}$$

Each of the presented measures uses the IC of the common ancestors between the two concepts. As such, they can use either MICA or DCA to calculate the IC_{shared} factor. We adapted the coherence score of node e according to source node s as:

$$coherence_e = PPR(s \rightarrow e) \cdot SSM(s, e) \cdot IC(e) \tag{5.10}$$

where SSM corresponds to one of the three SSM previously described.

5.3.4 Models

We studied the effect of SSM as a factor on the scoring function, and how it affects the accuracy of entity linking results. We first applied a baseline approach that consisted in selecting the ontology concept label most similar to the textual entity mention. This was implemented using the Levenshtein distance to obtain the label with the shortest distance to the text. This approach compares only the lexical form of the label, ignoring any context and semantics.

Then, we applied the PPR algorithm, using an approach similar to [18], but adapted to biomedical ontologies, which we refer to as the PPR model. As shown in (5.1), we can adjust the PPR score of each node with its IC. We refer to this model as PPR-IC. As previously explained, our adaptation of this approach has a distance parameter, corresponding to the maximum ontology distance between concepts. We studied the effect of this parameter on the PPR algorithm, to find the best value to use for further experiments.

We can then further adjust the contribution of each node to another node in the graph with the semantic similarity between them. As opposed to the model proposed by Pershina et al., we use all the candidates associated with the other entity mentions and not just the top scoring. This SSM factor will increase the weight of similar concepts, most likely to be coherent with the source node, and reduce the contribution of concepts less related to the source node. We refer to this model as PPR-SSM and study the effect of three SSMs on the accuracy of entity linking, evaluated on two gold standards. Furthermore, we compare two versions of each SSM: one using the IC of the MICA (5.4) and another using the DCA (5.5).

5.4 Results and discussion

5.4.1 Data

We evaluated our method on two gold standards, consisting of biomedical documents manually annotated with ontology concepts. Table 5.1 presents a comparison between the two gold standards. The ChEBI-patents corpus consists of 40 patent documents annotated with chemical entities, using the ChEBI ontology as reference. This gold standard was developed by a team of curators from ChEBI and the European Patent Office. The documents were selected to be representative of the universe of chemical patent documents. Whenever possible, the curators added the ChEBI concept identifier to the entity annotations. Since we were interested only in linking entity mentions to concept identifiers, we discarded entity mentions that were not assigned an identifier. There were 8407 textual entity mentions annotated with ChEBI identifiers in this corpus, corresponding to 2081 unique entity mentions. The ChEBI team provides an API that can be used to retrieve a list of concepts associated with a text search, which we used to obtain the candidate list for each entity. Since the annotation process was performed in 2009, we also used the ChEBI API to update concept identifiers that have changed since then automatically.

Table 5.1: Summary of the gold standards used for evaluation.

Gold standard	ChEBI-patents	HPO-GSC
Documents	40	228
Total entities	18061	2773
w/ ID	8407	2773
w/ candidates	6607	1890
Entities/doc	210.2	12.2

Additionally, we evaluated our method on a gold standard corpus of 228 scientific abstracts annotated with human phenotypes, associated with the Human Phenotype Ontology (HPO), which we refer to as HPO-GSC. We used an updated version of this corpus, which

aimed at improving the consistency of the annotations [31]. A total of 2773 textual named entities were annotated in this corpus, corresponding to 2170 unique entity mentions. We found that phenotype entities were more varied regarding nomenclature due to the existence of more synonyms for the same phenotype when compared to chemical entities. Comparing with the ChEBI-patents corpus, we can see that this corpus has fewer entities per document (ChEBI-patents: 210 entities/document; HPO-GSC: 12 entities/document). This ratio is relevant for our method because it aims at maximizing the coherence between entities, and documents with fewer entities are more prone to errors. We obtain a list of candidates for each entity through fuzzy string matching with the labels and synonyms of the HPO.

While in some cases the label of the concept matches the textual mention, in other cases there are some differences. Acronyms are common to both phenotypes and chemical entities. The HPO-GSC gold standard contains some overlapping entities, which could be mapped to different ontology concepts. While "neurofibromatosis" and "neurofibromas" were mapped to different concept identifiers, the current version of HPO merged those two concepts. As with the ChEBI-patents gold standard, we retrieved the most current identifier of each concept annotated on each gold standard. Other challenges consist in dealing with plurals (both the entity text and concept label can be plural or singular) and abbreviations and acronyms (the ontology may have some of these synonyms but not all).

We used the April 2018 release of ChEBI and the March 2018 release of HPO. The version of the ChEBI ontology that was used has about 54k manually verified chemical compounds. This ontology are curated by experts and updated monthly, while various sources are used to keep it as complete as possible, including user submissions. The HPO contains about 13k phenotypes and is focused on medically relevant phenotypes, and associating those phenotypes with diseases. This ontology is used various applications that deal with clinical data. Both ontologies tackle specific and complex areas of knowledge that benefit greatly from information retrieval methods.

5.4.2 Evaluation setup

We evaluated each model on both datasets considering the entities that were manually mapped to an ontology concept and for which the correct solution was in the set of candidates. Using the matching methods presented in the Methods section, we obtain a list of candidates for each entity. Table 5.1 shows that on the datasets used, 6607 (78.59%) and 1890 (68.16%) entities of the ChEBI-patents and HPO-GSC had its solution in the respective candidate list. We applied our PPR-SSM method for entity disambiguation to both datasets.

We found that the majority of concepts were not directly linked to each other in the ontology, meaning that the graph of each document would not have enough edges to apply PPR. For this reason, we studied the effect of considering the transitivity of subsumption relations, with a maximum distance between 0 and 8. For example, if 5 is the maximum distance, we would consider edges between concepts that have a path of 5 or fewer concepts between them.

We use the scoring functions described in (5.3) and (5.10) to rank the candidate list of each entity mention. In case of a tie, we pick the candidate with more subclasses. We considered only candidates with a matching score higher than 0.7, which was determined empirically to be the best threshold value. We then compared the predicted candidate with the gold standard to calculate the accuracy score.

The PPR algorithm was computed using the Monte Carlo approach presented by Fogaras and Racz [10]. We executed 2,000 iterations for each source node, performing five steps of PPR, with a probability of jumping to source node equal to 0.2. These values were suggested by Pershina et al. [18], which we kept since we saw no major improvements with a different number of iterations, steps or jump probability.

5.4.3 Experiments

Table 5.2 compares the accuracy of the proposed method with a string matching baseline and two other versions of the PPR algorithm: the first consisting of the PPR-based approach proposed by Pershina et al. adapted to biomedical ontologies and the second adding a weight to the contribution of each node based on its IC (5.1). We performed a baseline evaluation, which consisted in picking the top candidate with highest string matching similarity. We can see that although string matching obtains decent results on both gold standards, the PPR approach improves the accuracy on ChEBI-patents, while the PPR-IC approach improves both gold standards when compared to the baseline. Adding semantic similarity as a factor in the contribution of each candidate has a positive effect in both gold standards, obtaining a higher accuracy than the other approaches.

Table 5.2: Accuracy of PPR-SSM compared with a baseline and PPR model, on the ChEBI-patents and HPO-GSC gold standards.

Method	ChEBI-patents	HPO-GSC
Top match	0.527	0.638
PPR	0.665	0.554
PPR-IC	0.803	0.656
PPR-SSM	**0.804**	**0.683**

We compared paths of length 0, where no nodes are linked, to length 8, meaning that if there is a path shorter or equal to 8 between the concepts in the ontology, an edge is created in the graph. A maximum distance of 1 means that two concepts would be linked only if there was a direct relation between them. Since there is no difference between the accuracy of using maximum distance 0 and 1 on both gold standards, we can assume that direct relations solely are not enough to estimate coherence with the PPR algorithm. Each gold standard has a different optimal distance, with ChEBI-patents obtaining its best accuracy with distance 3 and HPO-GSC with distance 6. According to our experiments, the concepts linked by distances greater than those values do not contribute positively to the estimation

of coherence within candidates. We used those distance values when comparing different PPR-based approaches (Table 5.2) and SSMs (Table 5.3).

Table 5.3: Comparison of different semantic similarity measures for PPR-based entity linking.

SSM	IC_{shared}	ChEBI-patents	HPO-GSC
Resnik	MICA	0.7916	0.6306
	DCA	0.7916	0.634
Lin	MICA	0.7965	**0.6825**
	DCA	0.7965	0.6775
JC	MICA	0.8014	0.6775
	DCA	**0.8039**	0.6633

Comparing the results obtained with each SSM, we can see that different measures obtain the best results on each dataset. While JC-DCA obtains the best accuracy on the ChEBI-patents, Lin-MICA obtains the best accuracy on HPO-GSC. In both cases, the Resnik measure obtains lower scores than the PPR-IC model. The main difference between Resnik and the other measures is that it does not take into account the individual IC of the two concepts. On ChEBI-patents, none of the measures had a noticeable effect on the performance, and in most cases, it decreases the accuracy. However, the PPR-IC model leads to considerable improvement, so there would be fewer and more difficult cases for the PPR-SSM model to resolve. As the effect of the PPR-IC model on HPO-GSC was not as high, both Lin and JC measures improved the results.

5.4.4 Error analysis

We manually analyzed the errors of the PPR-SSM model on each gold standard, in order to understand the limitations of our approach. On the ChEBI-patents corpus, some errors were due to the same words being used to refer to a family of compounds and a type of chemical compound. For example, "polyamine" can refer to CHEBI:51349 (`polyamine`

macromolecule) and CHEBI:88061 (polyamine). Other errors were caused by the lack of edges between candidates, which happened in some documents. In these cases, the PPR algorithm cannot be applied, and the candidate with the highest number of descendants is chosen, which is not always the correct choice and does not take into account the global coherence. Another common error in linking chemical entities is with chemical compounds that have different charges, for example, biliverdin and biliverdin(2-). These two concepts are linked by *is conjugate acid of* and *is conjugate base of* relations. However, they have a different set of *is-a* ancestors, having only organic molecular entity and its respective ancestors in common. Context information from the text could help understand the specific molecule that is being mentioned. Many entities of the gold standard were not annotated with ChEBI identifiers (Table 5.1). These missing identifiers could improve the results of our method on this gold standard since the graph of each document would be more complete, and the global coherence score would take into account the complete set of entities.

On the HPO-GSC corpus, some errors were due to concepts with similar meanings. For example, "microretrognathia" and "micrognathia" are both facial deformations related to the development of the fetal mandible, and their respective HPO concepts have the same edges. Another common error was when dealing with child and parent concepts. For example, HP:0009588 refers to Vestibular Schwannoma and HP:0009589 to Bilateral vestibular Schwannoma and both appear in the candidate list for Bilateral vestibular Schwannoma. The parent term, Vestibular Schwannoma, obtained a higher score, resulting in an error. The parent term is closer to the top concepts, and therefore it will have paths to more concepts. The HPO has several instances where related concepts have similar labels, with a difference of just one word. Even though we try to account for this issue by giving more weight to concepts with higher information content, sometimes this weight is not enough and concepts that have more links are ranked higher than the correct answer.

5.5 Conclusion

Entity linking is an essential task to information extraction systems so that the information extracted can be linked to existing resources. However few approaches take advantage of the knowledge encoded in domain-specific ontologies. We proposed a method that combined existing entity linking approaches based on PPR with information from ontologies to calculate a global coherence score. Using this score, we ranked candidate matches to a named entity. Our method outperformed string matching and PPR-based methods in two case-studies, obtaining an accuracy of 0.8039 on the ChEBI-patents gold standard and 0.6825 on HPO-GSC. These results show the potential of the proposed method to be adapted to other domains. The code used to implement the method is publicly available.

The same performance may not occur in generic ontologies, such as DBpedia and YAGO, because these do not contain as many details as domain ontologies. Furthermore, unlike the ontologies used in this manuscript, the DBpedia ontology does not organize the entities, but only the classes of entities. Thus, our method could be adapted to generic ontologies, but the focus of our method is in domain-specific ontologies, which are rich in detail about a particular subject.

References

[1] W. Shen, J. Wang and J. Han. 'Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions'. In: *IEEE Transactions on Knowledge and Data Engineering* 27.2 (Feb. 2015), pp. 443–460. ISSN: 1041-4347. DOI: 10.1109/TKDE.2014.2327028.

[2] Delip Rao, Paul McNamee and Mark Dredze. 'Entity linking: Finding extracted entities in a knowledge base'. In: *Multi-source, multilingual information extraction and summarization*. Springer, 2013, pp. 93–115.

[3] Peter Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[4] S. K. Chan, W. Lam and X. Yu. 'A Cascaded Approach to Biomedical Named Entity Recognition Using a Unified Model'. In: *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. Oct. 2007, pp. 93–102. DOI: 10.1109/ICDM.2007.20.

[5] Martin Krallinger et al. 'Information retrieval and text mining technologies for chemistry'. In: *Chemical reviews* 117.12 (2017), pp. 7673–7761.

[6] Raul Rodriguez-Esteban. 'Biomedical text mining and its applications'. In: *PLoS Computational Biology* 5.12 (2009), pp. 1–5. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000597.

[7] A. C. B. Garcia, I. N. Ferraz and F. Pinto. 'The Role of Domain Ontology in Text Mining Applications: The ADDMiner Project'. In: *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*. Dec. 2006, pp. 34–38. DOI: 10.1109/ICDMW.2006.157.

[8] Janna Hastings et al. 'The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013'. In: *Nucleic Acids Research* 41.D1 (2013), pp. 456–463. ISSN: 03051048. DOI: 10.1093/nar/gks1146.

[9] Sergey Brin and Lawrence Page. 'The anatomy of a large-scale hypertextual web search engine'. In: *Computer networks and ISDN systems* 30.1-7 (1998), pp. 107–117.

[10] Dániel Fogaras and Balázs Rácz. 'Towards Scaling Fully Personalized PageRank'. In: *Algorithms and Models for the Web-Graph*. Ed. by Stefano Leonardi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 105–117. ISBN: 978-3-540-30216-2.

[11] Ravi Sinha and Rada Mihalcea. 'Unsupervised graph-based word sense disambiguation using measures of word semantic similarity'. In: *Semantic Computing, 2007. ICSC 2007. International Conference on*. IEEE. 2007, pp. 363–369.

[12] Ayman Alhelbawy and Robert Gaizauskas. 'Graph ranking for collective named entity disambiguation'. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2014, pp. 75–80.

[13] Lev Ratinov et al. 'Local and Global Algorithms for Disambiguation to Wikipedia'. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11. Portland, Oregon: Association for Computational Linguistics, 2011, pp. 1375–1384. ISBN: 978-1-932432-87-9. URL: http://dl.acm.org/citation.cfm?id=2002472.2002642.

[14] Priya Radhakrishnan, Partha Talukdar and Vasudeva Varma. 'ELDEN: Improved Entity Linking Using Densified Knowledge Graphs'. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (2018), pp. 1844–1853. DOI: 10.18653/v1/N18-1167. URL: http://aclweb.org/anthology/N18-1167.

[15] Razvan Bunescu and Marius Pasca. 'Using Encyclopedic Knowledge for Named Entity Disambiguation'. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* April (2006), pp. 3–7.

[16] Johannes Hoffart et al. 'Robust disambiguation of named entities in text'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2011, pp. 782–792.

[17] Xiao Cheng and Dan Roth. 'Relational inference for wikification'. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, pp. 1787–1796.

[18] Maria Pershina, Yifan He and Ralph Grishman. 'Personalized Page Rank for Named Entity Disambiguation'. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Section 4. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015, pp. 238–243. ISBN: 9781941643495. DOI: 10.3115/v1/N15-1026. URL: http://aclweb.org/anthology/N15-1026.

[19] Andrey Balmin, Vagelis Hristidis and Yannis Papakonstantinou. 'Objectrank: Authority-based keyword search in databases'. In: *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment. 2004, pp. 564–575.

[20] Gang Wu et al. 'Identifying potentially important concepts and relations in an ontology'. In: *International Semantic Web Conference*. Springer. 2008, pp. 33–49.

[21] P. Singla and P. Domingos. 'Entity Resolution with Markov Logic'. In: *Sixth International Conference on Data Mining (ICDM'06)*. Dec. 2006, pp. 572–582. DOI: 10.1109/ICDM.2006.65.

[22] A Morgan et al. 'Overview of BioCreative II gene normalization'. In: *Genome Biology* 9.Suppl 2 (2008), S3.

[23] Zhiyong Lu et al. 'The gene normalization task in BioCreative III'. In: *BMC bioinformatics* 12.8 (2011), S2.

[24] Yoshimasa Tsuruoka et al. 'Discovering and visualizing indirect associations between biomedical concepts'. In: *Bioinformatics* 27.13 (2011), pp. 111–119. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr214.

[25] Larry Smith et al. 'Overview of BioCreative II gene mention recognition.' In: *Genome biology* 9 Suppl 2.Suppl 2 (2008), S2. ISSN: 1474-760X. DOI: 10.1186/gb-2008-9-s2-s2. URL: http://genomebiology.com/2008/9/S2/S2.

[26] João D Ferreira et al. 'Assessing Public Metabolomics Metadata, Towards Improving Quality'. In: *Journal of integrative bioinformatics* 14.4 (2017).

[27] P Resnik. 'Using information content to evaluate semantic similarity in a taxonomy'. In: *International Joint Conference on Artificial Intelligence*. Vol. 14. Citeseer. 1995, pp. 448–453.

[28] F. Couto and A. Lamurias. 'Semantic similarity definition'. In: *Reference Module in Life Sciences (Encyclopedia of Bioinformatics and Computational Biology)*. Vol. 1. Oxford: Elsevier, 2019. DOI: https://doi.org/10.1016/B978-0-12-809633-8.20401-9.

[29] Dekang Lin. 'An Information-Theoretic Definition of Similarity'. In: *Proceedings of the Fifteenth International Conference on Machine Learning*. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 296–304. ISBN: 1-55860-556-8. URL: http://dl.acm.org/citation.cfm?id=645527.657297.

[30] Jay J. Jiang and David W. Conrath. 'Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy'. In: *Proceedings of the 10th Research on Computational Linguistics International Conference*. Taipei, Taiwan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), 1997, pp. 19–33. URL: http://www.aclweb.org/anthology/O97-1002.

[31] M. Lobo, A. Lamurias and F. Couto. 'Identifying Human Phenotype Terms by Combining Machine Learning and Validation Rules'. In: *BioMed Research International* 2017 (2017). ISSN: 2314-6133. DOI: https://doi.org/10.1155/2017/8565739.

6

Extracting MicroRNA-Gene Relations from Biomedical Literature using Distant Supervision

ANDRE LAMURIAS, LUKA A CLARKE AND FRANCISCO M COUTO

Abstract

Many biomedical relation extraction approaches are based on supervised machine learning, requiring an annotated corpus. Distant supervision aims at training a classifier by combining a knowledge base with a corpus, reducing the amount of manual effort necessary. This is particularly useful for biomedicine because many databases and ontologies have been made available for many biological processes, while the availability of annotated corpora is still limited. We studied the extraction of microRNA-gene relations from text. MicroRNA regulation is an important biological process due to its close association

with human diseases. The proposed method, IBRel, is based on distantly supervised multi-instance learning. We evaluated IBRel on three datasets, and the results were compared with a co-occurrence approach as well as a supervised machine learning algorithm. While supervised learning outperformed on two of those datasets, IBRel obtained an F-score 28.3 percentage points higher on the dataset for which there was no training set developed specifically. To demonstrate the applicability of IBRel, we used it to extract 27 miRNA-gene relations from recently published papers about cystic fibrosis. Our results demonstrate that our method can be successfully used to extract relations from literature about a biological process without an annotated corpus. The source code and data used in this study are available at https://github.com/AndreLamurias/IBRel.

6.1 Introduction

One of the major sources of current scientific knowledge is scientific literature, in the form of articles, patents and other types of written reports. This is still the standard method researchers use to share their findings. However, it is essential that a research group working on a certain topic is aware of the work that has been done on the same topic by other groups. This task requires manual effort and may take a long time to complete, due to the amount of published literature. One of the largest sources of biomedical literature is the MEDLINE database, created in 1965. This database contains over 26 million references to journal articles in life sciences, while more than 800,000 were added in 2015.

Automatic methods for Information Retrieval and Information Extraction aim at obtaining relevant information from large datasets, where manual methods would be infeasible. When applied to literature, this task is known as text mining. Named entity recognition is a text mining task that aims at identifying the segments of text that refer to an entity or term of interest. Another task is normalization, which consists of assigning an ontology concept identifier to each recognized entity. Finally, the relations described between the

identified entities can be extracted, which is known as relation extraction. The language used for scientific communication is formal, but the names of the biomedical entities may not be consistent across different papers. Nonetheless, text mining has been applied successfully to biomedical documents, for example, to identify drugs [1] and protein-protein interactions [2]. Supervised machine learning can be used to train a relation classifier. This approach requires an annotated corpus so that the algorithm can learn to predict the label of new instances. The algorithms that have been used for this task are, for example, conditional random fields [3] and kernel methods [4], based on shallow linguistic information [5] and parse trees [6].

In some domains, such as microRNA regulation, there is a limited amount of annotated corpora to train systems due to the cost of manually annotating text. MicroRNAs, or miRNAs, are small endogenous sequences of nucleotides used by animals, plants, and viruses to downregulate gene expression by targeting messenger RNA for cleavage or translational repression [7]. Since they were discovered, these molecules have been found to be associated with several biological processes, including various developmental and physiological processes. For this reason, their dysfunction might contribute to human diseases [8, 9]. The expression of each miRNA is regulated by transcription factors. Therefore, these regulatory relations provide an interesting case study of complex biological processes, where miRNAs are regulated upstream by transcription factors, while miRNAs target specific genes, and each miRNA-gene pair may be associated with one or more diseases. miRNAs are considered potential diagnostic and therapeutic targets for complex diseases [10]. As of September 2016, a "miRNA" keyword search on PubMed retrieves 52144 citations, of which 39568 were published in the last 5 years. The knowledge contained in these documents is of great importance to researchers working on a specific disease since it could lead to the formulation of new hypothesis.

Several databases have been created to improve the quality of the current miRNA knowledge. One of these databases, miRBase, indexes the reference names, sequences, and an-

notations of newly discovered miRNAs [11]. This initiative is particularly important in order to keep the nomenclature of all miRNAs consistent and unambiguous.

The Human MicroRNA Disease Database stores associations between miRNAs and diseases supported experimentally [12]. Another database, miRTarBase [13], provides information about miRNA-target relations, based on experimental data published in papers. Furthermore, this database provides a user interface with several features, such as visualization of miRNA-target networks, and an error report system. The authors update this database regularly, using natural language processing tools to choose which papers should be integrated. The reduce the curators' workload, the developers of this database added a text mining module on its latest release, contributing to an increase in the number of relations by 7-fold, comparing to the previous version. Chowdhary et al. [14] proposed a database for respiratory and related diseases, where the promoter regions of genes associated with these diseases are annotated with TFs and TF binding sites. With this database, it is also possible to compare genes, TFs, GO terms and miRNAs associated with selected diseases.

This increased interest in miRNA regulation has led to the development of computational methods to extract evidence based miRNA associations with genes, targets and diseases [15]. Computational methods provide various advantages over experimental methods, such as higher reproducibility and lower costs. The main techniques used to develop these methods are semantic similarity, network analysis and machine learning [16]. TFmiR is a web tool to analyze relations between miRNAs, transcription factors and genes of a specific disease, exploring the information contained in various knowledge bases [17]. This tool takes as input a list of miRNAs and genes and performs network analysis according to user input scenarios. The authors were able to identify core regulators and TF-miRNA regulatory motifs that were confirmed to be described in the literature. Liu et. al. [18] identified potential miRNA-disease relations by combining a disease network with a miRNA network based on miRNA-disease associations known from the Human MicroRNA Disease Database. Using miR-isomiRExp, it is possible to cluster miRNA isoforms according to their expression

pattern [19]. This type of analysis can be advantageous to understand miRNA maturation, processing mechanisms, and functional characteristics.

Recently, text mining approaches have been used to extract information about miRNA regulation. Murray et al. [20] extracted miRNA-target relations from PubMed using a list of verbal phrases, chosen to extract regulatory and functional interactions. Their method identified (miRNA, verb, gene) triplets, which were then manually validated, to reduce the number of false positives. The authors were able to identify 1165 miRNA-gene relations, which they used to generate a network. By aggregating relations described in multiple papers, they obtained a snapshot of the miRNAome and linked miRNAs to biological processes and diseases based on their corresponding genes. However, they did not evaluate the extraction process against a gold standard, and hence we were not able to compare their results to other works in terms of precision and recall.

miRSel is a database of miRNA-gene relations which uses text mining methods to automatically update its entries [21]. The authors extracted miRNA entities using regular expressions and gene entities based on a dictionary compiled from several databases. Similarly to Murray et. al., they also compiled a list of 70 expressions used to describe miRNA-gene relations and extracted the instances where a miRNA, gene, and expression co-occurred. They evaluated their method on a set of 89 sentences from PubMed abstracts, obtaining an F-score of 0.83.

The developers of OMIT (Ontology for MicroRNA Target) explored automatic methods to find new miRNA terms to add to the ontology [22]. They obtained abstracts related to miRNA through keyword search on PubMed and filtered out the terms that were already mapped to the ontology. Then, nouns and noun phrases that did not match with existing ontology concepts were considered candidate terms. The most frequent candidate terms were then reviewed by domain experts and added to the ontology when appropriate. Starting with 49,447 abstracts and 488,576 nouns and noun phrases, the authors were able to add 117 new terms to the ontology. This type of approach can be enhanced by using a more advanced

term extraction method, in order to present only high confidence candidates to the domain experts, requiring less manual effort.

Bagewadi et. al [23] compared various approaches to miRNA-gene relation extraction, including co-occurrence and machine learning algorithms. To evaluate these approaches, they manually annotated a corpus of 301 abstracts with various types of entities, including miRNAs and genes/proteins, and with the relations described in each sentence. Using the supervised machine learning approach, their best F-score was 0.76, while using a co-occurrence approach their best F-score was 0.73.

Li et. al [24] developed miRTex, which extracts miRNA-gene relations based on a set lexico-syntactic rules. They developed an annotated corpus of 150 abstracts to evaluate their system, which obtained the F-score of 0.94 for miRNA-gene relation extraction. Then, they applied their system to a set of 13M abstracts and 1M full-text documents and released a database containing those results. The authors obtained an F-score of 0.87 on Bagewadi et al.'s corpus. However, their method was based on hand-crafted rules and lists of keywords which are difficult to generalize and require costly manual curation. Although they obtained high F-score values for miRNA-gene relation extraction, it is not clear how their methods could be efficiently applied to other datasets. This is a common issue of rule-based and supervised learning approaches.

It is not feasible to develop an annotated corpus for every domain since it is a time-consuming process and the annotations are likely to be biased to that particular corpus. Consequentially, there has been an increasing interest in semi-supervised and unsupervised approaches to perform relation extraction. Fully unsupervised approaches explore clustering algorithms to identify patterns in the text that could indicate the presence of a relation. For example, Rosenfeld et al. [25] clustered pairs of entities using context features related both to the pair and to each entity, obtaining high precision levels. Alternatively, other authors have developed bootstrapping methods based on a small set of relations [26].

Distant supervision (sometimes referred to as weak supervision) combines advantages of

both supervised and unsupervised learning [27]. This technique assumes that any sentence that mentions a pair of entities corresponding to a knowledge base entry is likely to describe a relation between those entities. For example, any sentence mentioning "Nikola Tesla" and "New York" would be identified as a positive example of a "lived in" relation. This would include sentences such as "Nikola Tesla lived in New York from 1933 to 1943" but also "Nikola Tesla planned the Wardenclyffe Tower facility in New York", which does not in fact represent a "lived in " relation. However, the fact that a corpus of any size can be used as training data is an advantage over supervised learning, which is limited by the amount of documents manually annotated by experts. The pseudo-relations inferred using this technique can then be used to train a classifier using machine learning algorithms.

Multi-instance learning [28] addresses some limitations of distant supervision, by considering that not every co-occurrence will correspond to a relation mention. With this type of model, the pairs are grouped into bags where at least one of the pairs is true, but it is unknown if all pairs of the same bag are true. Riedel et al. [29] used this technique to extract Freebase relations from newspaper articles, obtaining a lower error rate than other distant supervision approaches. Min et al. [30] proposed an approach to reduce the number of incorrectly labeled relations, by considering only positive and unlabeled pairs. They found out that many of the pairs classified as negative from two distant supervision datasets were actually false negatives. These false negatives will have a significant impact on the performance of a classifier trained on those datasets.

Biomedicine is a challenging domain for text mining, due to the complexity of the studied processes. It is often necessary to train classifiers with a dataset annotated by domain experts with specific entities and relations due to the specialized terminology used to describe some processes. Distant supervision can overcome this necessity, by combining a set of documents with an existing knowledge base. These knowledge bases can be in the form of databases and ontologies, which already exist for many biological processes. Craven and Kumlien [31] have previously explored biomedical databases to generate training data for a relation

extractor. They retrieved 924 abstracts that were referenced in the entries of the Yeast Protein Database and selected sentences that mentioned two entities corresponding to a database entry. Using these sentences, they trained a sentence classifier to extract subcellular locations of proteins. Their results demonstrated that weakly labeled data can be advantageous for relation extraction. Other authors have also explored this type of approach in the context of the biomedical domain [32, 33].

In this paper, we describe our method, IBRel - Identifying Biomedical Relations, which does not require a manually annotated corpus. To the best of our knowledge, this is the first biomedical relation extraction method based on multi-instance learning. Our method was based on the sparse multi-instance learning algorithm, used to train on an automatically generated corpus of 4,000 documents related to miRNAs. We evaluated IBRel on three datasets, comparing multi-instance learning with co-occurrence and supervised learning. IBRel was superior to supervised learning on one of three datasets, for which there was no specific training set available. To demonstrate how this method can be applied to a specific subject, we used it to extract relations from abstracts related to miRNA regulation and cystic fibrosis (CF). Recently, the role of miRNAs as therapeutic targets and in regulating cystic fibrosis transmembrane conductance regulator (CFTR) expression has been a topic of increasing interest to the CF research community [34, 35]. We were able to extract several miRNA-gene relations relevant to CF, highlighting how this work can lead to the improvement of our knowledge about human diseases.

6.2 Materials and Methods

6.2.1 Corpora

Table 6.1 provides details about the corpora used for this work, both to develop (Dev) and evaluate (Eval) the system. Our objective was to perform a robust evaluation of our miRNA-gene relation extraction method. As such, the corpora used represented various

annotations methodologies. TransmiR and miRTex were annotated only with document-level relations, while Bagewadi contained mention-level relations. Document-level relation annotations consist of a list of relations associated with each document, with no specific text span associated with each relation. When a corpus is annotated with mention-level relations, the location in the text of each annotated relation is known. The algorithms we used required mention-level relations for training, so both the miRTex and TransmiR corpus could not be used to develop the relation extraction system, but only for its evaluation. The IBRel-miRNA corpus contains more documents and entities than the others because it was automatically generated. The purpose of this corpus was to develop an approach based on distant supervision. We applied our method on a corpus of abstracts about cystic fibrosis and miRNAs, in order to demonstrate how IBRel can be used to obtain knowledge about a specific disease.

We used the training set of Bagewadi's corpus [23] to train a miRNA-gene relation classifier as well as classifiers for miRNA and gene entity recognition. Furthermore, we used the respective test set to evaluate miRNA and gene entity recognition and miRNA-gene mention-level relation extraction. Bagewadi's corpus consisted of MEDLINE abstracts, selected using the keyword "miRNA". The authors annotated 301 documents with specific and non-specific miRNAs, Gene/Proteins, Diseases, Species, and Relations Triggers, as well as undirected

Table 6.1: Corpora used to develop and evaluate the system. Each line refers to a corpus, how it was used (Dev: development; Eval: evaluation; NER: Named Entity Recognition; RE: Relation extraction), the total number of relevant entities and relations annotated, and the number of documents.

	NER		RE				
Corpus	Dev	Eval	Dev	Eval	Entities	Relations	Documents
Bagewadi's	X	X	X	X	1963	318	301
miRTex	X	X		X	1245	771	350
TransmiR		X		X	1145	547	243
IBRel-miRNA			X		52970	NA	4000
IBRel-CF				X	612	NA	51

relations between entities mentioned in the same sentence. The inter-annotator agreement score was 0.916 for specific miRNAs and 0.752 for Gene/Proteins.

We used the miRTex corpus [24] to evaluate miRNA and gene entity recognition, as well as document-level miRNA-gene relation extraction. This corpus was annotated with miRNA and genes entities, and with three types of relations: miRNA-gene, miRNA-target, and gene-miRNA. A relation was classified as miRNA-target if a direct interaction between a miRNA and gene was described. In this corpus, the relations were annotated only at document-level, i.e. no specific text span associated with each relation. The inter-annotator agreement for the relations was 0.86, determined for a set of 20 abstracts.

TransmiR is a database that stores transcription factor-miRNA regulatory relationships found in the literature [36]. In this study, we created the TransmiR corpus, based on the document abstracts associated with the entries related to humans of this database. The abstracts were retrieved from PubMed using the identifiers provided with each entry. Since one of the fields of each entry of this database was "organism", we used every entry that had something other than "human" as the knowledge base for distant supervision. There were three abstracts that were not available on PubMed (PMIDs 17972953, 20046097 and 18818704), resulting in a total of 243 abstracts. Each abstract was annotated with the miRNA-gene relations that exist in the database. Our objective was to determine if we can obtain the same relations using our method.

Distant supervision requires a large corpus and a knowledge base containing the type of relations to be extracted. Regarding the knowledge base, we used the non-human entries of the TransmiR database to avoid overlap with the TransmiR corpus. Furthermore, we obtained 4,000 documents about miRNAs from PubMed, using the MeSH term "miRNA", ordered by publication date. We refer to this corpus as IBRel-miRNA corpus. This corpus consisted uniquely of these documents, without any type of annotation. However, to use it for distant supervision, we classified the text with named entity recognition classifiers, in order to obtain miRNA and gene named entities. This process is described in more detail in the "Biomedical

Named Entity Recognition" section. Entities found were matched to the knowledge base to generate training data for the distant supervision model.

To demonstrate the usefulness of this technique to a particular real-world problem, we retrieved a corpus of 51 documents from PubMed, using the keywords "cystic fibrosis" and "miRNA" (IBRel-CF corpus). Similarly to the IBRel-miRNA corpus, we classified each document with named entity recognition classifiers, in order to obtain miRNA and gene named entities. Afterward, we classified each document with IBRel, as described in the "Identifying Biomedical Relations" section.

6.2.2 Evaluation

Our experimental approach combined natural language processing techniques, as well as machine learning algorithms. The pipeline developed for this approach and the corpora used to evaluate each module are presented in Fig 6.1. The first module (B) processes the input text (A), extracting sentence and word boundaries, as well as token features such as lemma and part-of-speech. These features were necessary to develop and evaluate the other two modules. The NER module (C) consisted of named entity classifiers trained for miRNA and gene/protein entities, while the RE module (D) consisted of our method for miRNA-gene relation extraction, IBRel. Furthermore, we compared our method with supervised learning and co-occurrence approaches. Fig 6.1 also shows how each corpus was used, either to develop or evaluate the NER and RE modules. The corpora mentioned in Figs 6.1E, F, G, and H are the same ones mentioned in Table 6.1, except IBRel-CF, which was not used to develop or evaluate the system, but only as an independent case study.

As shown in Fig 6.1D, the miRNA-gene relation extraction module was evaluated on three corpora. Each corpus was developed using different methodologies and guidelines, therefore we consider this to be a robust evaluation. Using miRTex corpus, we studied the capacity to identify the relations of each document, while using Bagewadi's corpus, we studied the capacity to identify each relation mention from the text. The TransmiR corpus evaluation

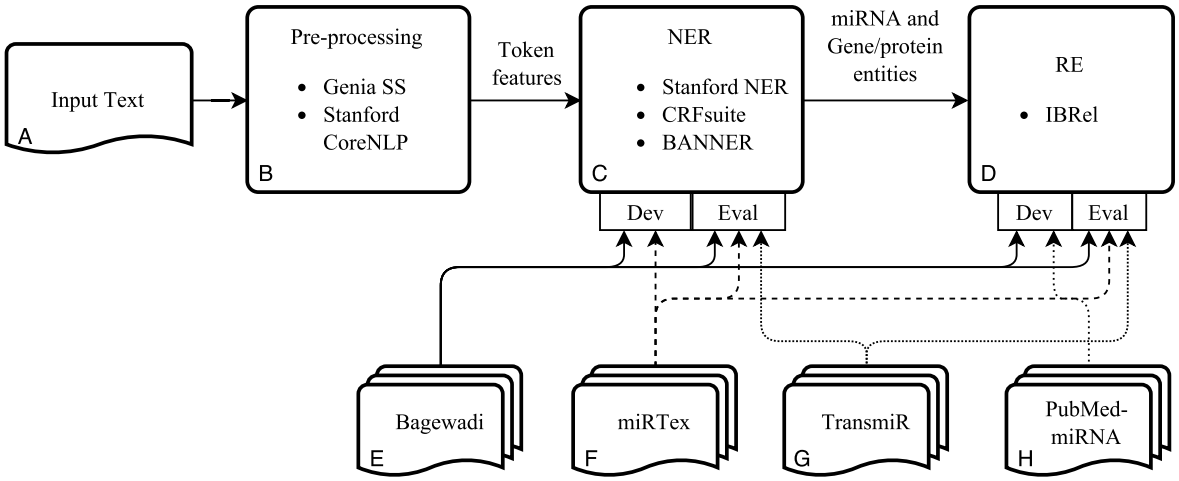


Figure 6.1: Pipeline used to perform the experiments. The input text (A) first goes through natural language processing tools to generate token features (B), then a named entity recognition module (C) to identify named entities and finally relation extraction (D) to extract relations between entities. Bagewadi (E), miRTex (F), TransmiR (G) and IBRel-miRNA (H) refer to the four corpora previously described.

provides a point of comparison to manually curated databases. However, this evaluation had some limitations. First, it is not possible to evaluate the extraction of relations independently from named entity recognition; if some entities are not correctly recognized, it will not be possible to extract relations that include those entities. Second, it may be the case that the system identified relations that were not in the database. However, it does not necessarily mean that they were incorrect since we used a different set of entries of the TransmiR database to train and evaluate the system: entries where the organism was different from "human" were used to train the model (IBRel-miRNA corpus) while the other entries were used for the TransmiR corpus gold standard. Third, we retrieved only abstracts related to the database entries. However, some relations from the database were not mentioned in the abstract, but only in the full text, figures, tables or supplementary material. These limitations should be taken into consideration when interpreting the results.

On Bagewadi's corpus, which contained relation mentions, we considered a true positive

if the offsets of the two entities of the pair matched the gold standard. For the other corpora, we normalized the text of each element of the relations to database identifiers from miRBase and UniProt [37]. This way, the possibility of false positives occurring due to nomenclature variation was reduced. We searched UniProt for the entry with the highest confidence that matched each protein entity, while for miRNAs we used a set of rules to match each miRNA entity to miRBase. We describe this process in greater detail in the "Biomedical Named Entity Recognition" section. Furthermore, we did not consider the direction of the relation when evaluating the results so that the order of the elements of each pair did not affect the results.

Three of the five corpora used were not annotated with named entities, hence it was necessary to perform and evaluate named entity recognition of miRNAs and genes. We used the test sets of miRTex and Bagewadi to evaluate this task since both were annotated with miRNA and gene mentions.

The evaluation measures used to evaluate the NER and RE modules were precision, recall, and F-score. These measures are commonly used to compare the performance of relation extraction methods on community challenges [2, 38]. Precision corresponds the fraction of relations retrieved that were relevant, while recall corresponds to the fraction of all relevant relations that were retrieved by the method. F-score corresponds to the harmonic mean between precision and recall. This measure it particularly important since it is usually trivial to obtain either high precision at the expense of a low recall, or vice-versa. However, these measures depend on the distribution of the corpus [39], so it can be difficult to compare results across different test sets.

6.2.3 Identifying Biomedical Relations

Our objective was to identify miRNA-gene regulatory relations in scientific abstracts without requiring additional manual data curation. We present a method, IBRel, to extract biomedical relations based on distant supervision. Our method requires only a set of doc-

uments, which can be easily retrieved from PubMed, and a knowledge base, which already exists for many biological problems. We focused on miRNA regulatory relations and selected the TransmiR database as the knowledge base. Each miRNA-gene pair mentioned in the same sentence was considered a potential miRNA-gene relation mention. These relations could be either a miRNA regulating the activity of a gene, or a gene or protein regulating the transcription of a miRNA. Existing approaches to extract this type of relation are based on fixed rules, which are difficult to adapt to other relations, or manually annotated corpora, which are costly to produce. Therefore, we considered miRNA regulation a relevant case-study to demonstrate the usefulness of IBRel to biological problems given the lack of annotated corpora.

The proposed method required a corpus to generate training instances. This corpus had to be larger than any other miRNA-gene corpora in terms of number of documents, and it should contain entities and relations relevant to miRNA regulation, i.e. the text should contain instances of miRNA-gene relations. We retrieved a corpus of 4,000 abstracts from PubMed, using the MeSH term "miRNA" (IBRel-miRNA corpus). Firstly, we applied a NER algorithm to recognize the miRNA and gene entities in this corpus. The NER algorithm was based on a machine learning classifier trained on the Bagewadi and BioCreative 2 GM task datasets, and both classifiers were evaluated on gold standards. Using the IBRel-miRNA corpus and the recognized entities, we trained a classifier for miRNA-gene relation extraction.

To train IBRel, we used the sparse multi-instance learning algorithm (sMIL) [40]. The sMIL algorithm is based on the assumption that the bags are sparse, meaning that only a few instances are positive in each bag. Although this algorithm was first applied to image classification, other authors have used it for relation extraction [41]. An abstract may mention each miRNA and gene multiple times but generally, due to word restrictions, the relation will be stated only once. This is the reason why this variation of multi-instance learning was chosen for this task.

It was necessary to define how the sMIL algorithm would be integrated into our method to extract biomedical relations. Multi-instance learning differs from traditional supervised learning in the sense that instead of using a training set composed of labeled instances it uses a training set composed of labeled bags of instances. The main challenge in adapting multi-instance learning to the biomedical domain was defining how to represent the data in the form of bags. In our case, each bag contains multiple relations. These bags can be positive, if at least one of the instances corresponds to a true relation, or negative if no instances in that bag are true. In a biomedical abstract, a given miRNA and gene may co-occur several times, while only some of those instances correspond to the description of a miRNA-gene relation. Take into consideration the sentence: "These abnormalities reflect the regulation of a cadre of modulators of SRF activity and actin dynamics by miR-143 and miR-145." (PMID 19720868); a relation is described between the gene SRF and two miRNAs. However, in the following sentence of that document: "Thus, miR-143 and miR-145 act as integral components of the regulatory network whereby SRF controls cytoskeletal remodeling and phenotypic switching of SMCs during vascular disease.", the same miRNAs and gene are mentioned but no relation is described.

To generate the bags for the sMIL algorithm, we considered an instance as a miRNA and gene co-occurrence in a sentence (Algorithm 1). Fig 6.2 contains an example of a sentence that produces one bag with two instances and another bag with one instance. The features used to represent each instance consisted of the words used before, between and after the two elements of the pair as well as their respective lemma, part-of-speech and named-entity tag (Person, Location, Organization, Numerical, Temporal, or Other) (Example 1). The size of the word window used was variable, and we experimented with window sizes 1, 3 and 5. We then converted these features into a bag-of-words representation using sci-kit learn [42]. These features were selected with the objective of being similar to the ones used by the supervised machine learning algorithm we chose to compare with IBRel.

Example 1 *Sparse Multi-instance learning instance example.*

Algorithm 1 Bag generation algorithm

```
1:  function GENERATE_BAGS(corpus, transmir_human)
2:      bags = []
3:      for sentence in corpus do
4:          for mirna in sentence do
5:              for gene in sentence do
6:                  bag = (mirna, gene)
7:                  instance_features = generate_features(bag, sentence)
8:                  if bag not in bags then
9:                      if bag in transmir_human then
10:                         bag.label = 1
11:                     else
12:                         bag.label = 0
13:                     bags.add(bag)
14:                 bags.add_instance_to_bag(bag, instance_features, bag_label)
15:     return bags
```

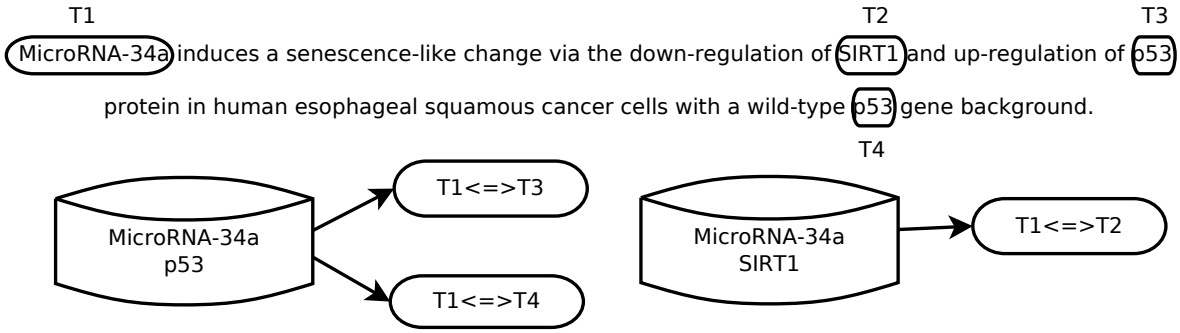


Figure 6.2: Multi-instance learning bags. For each sentence, we generated bags according to the distinct miRNA-gene pairs mentioned in the text. If a pair exists in the reference database, the bag is labeled as positive. Multi-instance learning assumes that at least one of the instances of a positive bag should describe a true relation.

- ***Sentence:*** *These abnormalities reflect the regulation of a cadre of modulators of SRF activity and actin dynamics by miR-143 and miR-145. (PMID 19720868)*
- ***Pair:*** *miR-143 - SRF*
- ***Label:*** *1*

- ***Feature vector:*** *(0-before-cadre-NN-O 1-before-of-IN-O 2-before-modulators-NNS-O 3-before-of-IN-O 0-middle-activity-NN-O 1-middle-and-CC-O 2-middle-actin-NN-O 3-middle-dynamics-NNS-O 4-middle-by-IN-O 0-end-and-CC-O 1-end-miR+145-NN-O 2-end-.-.-O)*

We did not manually annotate the relations of the training corpus, so it was necessary to explore a knowledge base to assign labels. This knowledge base had to contain relations of the same type as the ones to be extracted. For this purpose, we used the entries from TransmiR that were not related to the human species. This way, we avoided overlapping with the TransmiR corpus used for evaluation, which was generated using only the human entries. Each TransmiR entry contains a miRNA identifier as well as a gene name. We used these two columns to match with the miRNAs and genes found in the text. As shown in Algorithm 1, if the miRNA-gene pair existed in the human TransmiR database, the bag label was 1, otherwise, it was 0.

We trained a classifier for miRNA-gene relation extraction on bags generated from the IBRel-miRNA corpus and the TransmiR database, following Algorithm 1. The sMIL algorithm learned a classification model from the training data as described in "Corpora" and implemented by the miSVM package (`https://github.com/garydoranjr/misvm`). We used the default values of miSVM since we did not want to overfit the classifier to a particular dataset.

We evaluated IBRel on three datasets (Bagewadi, miRTex, and TransmiR). Those three datasets were chosen because two of them were manually annotated with miRNA-gene annotations while the other one was obtained using TransmiR database entries that were not used to train the classifier. We generated instances from each document and bags containing those instances, as previously described. If a bag was classified as positive, every instance in that bag was also classified as positive.

The confidence level of each prediction was estimated using the distance to the hyperplane, provided by the support vector machines algorithm. We used a logistic link function

to obtain the probability output, as suggested by Wahba [43]. This probability was given by Equation 6.1, where $f(x)$ is the uncalibrated value returned by the SVM.

$$P(class|input) = P(y = 1|x) = p(x) = \frac{1}{1 + exp(-f(x))} \tag{6.1}$$

If a relation was found in more than one document, we used the maximum confidence level obtained.

6.2.4 Supervised Machine Learning and Co-occurrence approaches

To assess the performance of IBRel on miRNA-gene relation extraction, we performed the same analysis using two other relation extraction approaches. First, we applied a co-occurrence approach which consisted in classifying every miRNA-gene pair in the same sentence as positive. This approach is considerably faster but tends to overestimate the number of relations, producing more false positives. However, some authors have obtained strong results using co-occurrence for relation extraction [44, 23]. For example, Bagewadi et. al. mention that their co-occurrence approach obtains similar results of machine learning approaches. The assumption is that due to restrictions on the word number in abstracts, a sentence that mentions two entities is likely to describe a relation between those two entities.

As another comparative approach, we used a variation of support vector machines, with a shallow linguistic kernel, as implemented by Giuliano et al. [5], to train a classifier on an annotated corpus. The advantage of kernel methods such as this one is the fact that no features have to be designed and tested. This kernel compares the sequence of tokens, lemmas, part-of-speech and named entities of each instance with the others. Tokens that refer to each argument are identified and substituted by a generic string so that the original text does not affect the algorithm. The label of each instance was 0 if it described relation, or 1 if it did not describe a relation. Example 2 provides a feature vector of a pair instance of this algorithm. Each element corresponds to a token and is constituted by its order in the sentence, original

text, lemma, part-of-speech, named-entity tag (Person, Location, Organization, Numerical, Temporal, or Other) and candidate identifier (A - Agent, T - Target).

Example 2 *Shallow Linguistic Kernel instance example.*

- ***Sentence:*** *These abnormalities reflect the regulation of a cadre of modulators of SRF activity and actin dynamics by miR-143 and miR-145. (PMID 19720868)*
- ***Pair:*** *miR-143 - SRF*
- ***Label:*** *1*
- ***Feature vector:*** *(0/These/these/DT/O/O, 1/abnormalities/abnormality/NNS/O/O, 2/reflect/reflect/VBP/O/O, 3/the/the/DT/O/O, 4/regulation/regulation/NN/O/O, 5/of/of/IN/O/O, 6/a/a/DT/O/O, 7/cadre/cadre/NN/O/O, 8/of/of/IN/O/O, 9/modulators/modulator/NNS/O/O, 10/of/of/IN/O/O, 11/#candidateb#/#candidateb#/NN/ENTITY/T, 12/activity/activity/N-N/O/O, 13/and/and/CC/O/O, 14/actin/actin/NN/O/O, 15/dynamics/dynamics/NNS/O/O, 16/by/by/IN/O/O, 17/#candidatea#/#candidatea#/NN/ENTITY/A, 18/and/and/CC/O/O, 19/miR-145/mir-145/NN/ENTITY/O, 20/././O/O)*

This kernel has been applied to biomedical text, for the extraction of relations between proteins [5] and chemical compounds [45]. The shallow linguistic kernel is a composite sequence kernel that uses both a local and global context window. We performed experiments using windows with size 1, 3 and 5. We used Bagewadi's corpus to train a miRNA-gene relation classifier using this kernel since this was the only corpus available that was manually annotated with that type of relation mention.

6.2.5 Biomedical Named Entity Recognition

The recognition of biomedical entities is a critical step to our method because the algorithms used require these entities to be annotated in the text. While gene/protein named

entity recognition is a task for which many systems have been developed, the same is not true regarding miRNAs. It was necessary to develop a method to recognize miRNA entities and evaluate both gene and miRNA named entity recognition methods. Then, each entity recognized was mapped to a database identifier. This step improves the quality of the information extracted by reducing lexical variation and by integrating external domain knowledge.

We applied an existing system for gene/protein named entity recognition, BANNER [46]. This system was evaluated on the three test sets used since we could not find published results on those datasets. BANNER is based on the conditional random fields algorithm [47]. This is a state-of-the-art algorithm used by NER systems that learns the patterns of tokens from an annotated gold standard. The model generated is then able to classify new text according to those patterns. BANNER contains a specific set of features based on orthographic, morphological and shallow syntax features. We used the model they trained for protein and gene named entity recognition on the BioCreative 2 GM task dataset. BANNER assigned a label to each token, expressing if that token was part of an entity or not.

We used the UniProt API to obtain the entry names corresponding to each gene entity. Example 3 provides an example of the query used, as well as the output obtained. Since this API does not provide a confidence score, we selected only the first entry obtained when sorted by their own internal score. Entities that were not mapped to the reference database were excluded. Since we are working with published papers, it is unlikely that the genes and proteins mentioned would be missing from the databases. UniProt was chosen instead of a more gene specific database to match both protein and gene entities to database identifiers because we wanted to identify as many entities as possible. Table 6.2 provides various examples of genes and proteins mapped to UniProt.

Example 3 *UniProt API example query and its output* `http://www.uniprot.org/uniprot/?query=insulin&sort=score&columns=id,entry%20name,reviewed,protein%20names,organism,&format=tab&limit=1`

Output: P06213 INSR_HUMAN reviewed Insulin receptor (IR) (EC 2.7.10.1) (CD anti-

gen CD220) [Cleaved into: Insulin receptor subunit alpha; Insulin receptor subunit beta] Homo sapiens (Human)

We performed basic pre-processing on the input text to extract features to train miRNA named entity classifiers on the text. Our system first splits the text into sentences, using the GENIA sentence splitter [48]. Each sentence is then processed by Stanford CoreNLP pipeline [49], to tokenize and extract lemmas, part-of-speech tags and named entity tags (proper noun, numerical or temporal entities) from the text.

We trained conditional random field classifiers on Bagewadi's corpus for miRNA named entity recognition. For each corpus, we trained two classifiers: one using Stanford NER with the default features and another with CRFsuite, using the features described in [50]. Our objective was to maximize the number and variety of entities found since this is a limiting step for relation extraction. It has been shown that combining classifiers training with different implementations and features can improve the performance of a text mining system [51].

miRNA entities were mapped to a list of human miRNA names extracted from miRBase, which includes the names of mature and pre-mature miRNAs, as well as deprecated names. We used some rules to reduce the variation of miRNA entities, in order to obtain better miRBase matches. These rules were based on the most common spelling variations of miRNAs. Sometimes authors mention multiple miRNA at the same time, for example: "mir-192/215", "mir-34a/b/c", "mir-143 and -145". We split a miRNA entity if it contained "/", " and " or ", ". However, this rule was not applied to Bagewadi's corpus because the

Table 6.2: Example of gene entities identified that were then matched with UniProt entries. Entity text refers to the original text found in the abstract, while Entry name and Entry ID refer to UniProt entries.

Entity text	Entry name	Entry ID
Smad	SMAD3_HUMAN	P84022
N-Myc	NDRG1_HUMAN	Q92597
Interferon regulatory factor 3	IRF3_HUMAN	Q14653
Egr-2	EGR2_HUMAN	P11161

guidelines specified that multiple miRNAs mentioned sequentially should be annotated as only one. Although miRNA nomenclature is well defined, some slight deviations appear in the literature. For example, sometimes "microRNA-" and "miRNA-" is used instead of "miR-". In some papers, there is no dash connecting the "miR-" prefix to the respective number, for example, "miR125a". Furthermore, human miRBase entries contain a "hsa-" prefix, which is not always used in the literature. We used simple post-processing rules to fix these variations. Then, each entity was matched to the list of miRNAs from miRBase, using fuzzy string matching. The confidence score of each match corresponded to the Levenshtein distance between the original text and the match. The Levenshtein distance is a string metric which is related to the minimum number of edits necessary to transform one string into another. Based on our experiments, we ignored matches with scores lower than 0.95 since many matches with those scores were incorrect. Table 6.3 provides some examples of the normalization process for miRNAs.

6.3 Results

We evaluated miRNA-gene relation extraction on three datasets: Bagewadi, miRTex and TransmiR (Figs 6.1D, E, F, G, and H). These were the datasets which were annotated with miRNA-gene relations, although TransmiR was annotated automatically. Table 6.4 presents the miRNA-gene relation extraction results on those datasets. The co-occurrence approach

Table 6.3: Example of miRNA entities identified that were then matched with miRBase entries. Entity text refers to the original text found in the abstract, while Entry name and Entry ID refer to miRBase entries.

Entity text	Entry name	Entry ID
miRNA-155	hsa-miR-155	MI0000681
miR-200	hsa-miR-200a	MI0000737
miR125a	hsa-mir-125a	MI0000469
microRNA-9	hsa-mir-9	MI0000466

consisted in classifying as true every miRNA-gene pair co-occurring in the same sentence We evaluated supervised learning with a shallow linguistic kernel (SL kernel), using a classifier trained on Bagewadi's corpus (supervised learning) and our method, IBRel, using a classifier trained on the IBRel-miRNA corpus. We used a fixed window of 3 on the SL kernel and IBRel, while we provide results for windows of size 1 and 5 in Supplementary Material.

Comparing the three methods in terms of F-score, the shallow linguistic kernel approach obtains the best score on two corpora (Bagewadi and miRTex), while the IBRel outperformed the others on the TransmiR corpus. Comparing in terms of precision, IBRel obtained the best score on two corpora (miRTex and Bagewadi), while the shallow linguistic kernel obtained the highest score on Bagewadi's corpus. With all three methods, the highest F-score obtained was on Bagewadi's corpus. However, the F-score obtained for miRTex and TransmiR using IBRel was close (0.413 and 0.383), while for the other two approaches, the F-score on TransmiR is lower than on miRTex (co-occurrence: 0.623 and 0.25; kernel: 0.654 and 0.130).

We also evaluated miRNA and gene entity recognition using conditional random fields on the same datasets (Figs 6.1C, E, F, G, and H) since this is a required step for the relation extraction approaches we used (Table 6.5). For Bagewadi and miRTex, we used the respective training and test sets to recognize miRNA entities, merging the results obtained with two conditional random fields classifiers. For the TransmiR corpus, we used the classifiers

Table 6.4: miRNA-gene relations extraction evaluation results on each corpus, comparing co-occurrence, supervised and IBRel (window size = 3). P, R and F refer to precision, recall and F-score.

Method	Bagewadi's	miRTex	TransmiR
Co-occurrence	0.689	0.623	0.250
SL kernel	**0.757**	**0.654**	0.130
IBRel	0.532	0.383	**0.413**

trained on the miRTex corpus, which obtained higher values on its own evaluation. Gene entity recognition on every corpus was performed using BANNER. On TransmiR, the results obtained were lower than on the other two corpora, particularly for gene entities. This issue is related to how that corpus was developed and will be discussed in the following section.

After evaluating our method, we used it extract miRNA-gene relations from a set of abstracts about cystic fibrosis and miRNAs (Table 6.6). The purpose of this study was to demonstrate the applicability of our method to a specific domain. These abstracts were removed from the training set (IBRel-miRNA corpus) to avoid any bias when developing IBRel 6.1. We were able to extract 27 relations, between 18 different miRNAs and 12 different genes. A total of 11 relations between the CFTR gene and a miRNA were found, which was to be expected since CFTR is the gene responsible for cystic fibrosis and the abstracts chosen dealt in most part with miRNA involvement in this disease. The maximum confidence level corresponds to the highest confidence of all instances of that particular relation. The confidence level of each instance was calculated by estimating the distance to the hyperplane, given by Equation 6.1. The relations with the highest confidence were also found in more sentences and abstracts.

6.4 Discussion

Our method obtained better results when applied to the TransmiR corpus. When compared to the supervised learning approach, the F-score on this corpus was improved by 0.283

Table 6.5: Entity recognition evaluation results on each corpus, for miRNA and gene entities. P, R and F refer to precision, recall and F-score.

	miRNA			Gene		
Gold standard	P	R	F	P	R	F
Bagewadi's	0.902	0.936	0.919	0.814	0.580	0.677
miRTex	0.934	0.948	0.941	0.803	0.788	0.795
TransmiR	0.726	0.651	0.687	0.255	0.618	0.361

Table 6.6: miRNA-gene relations extracted from the IBRel-CF corpus using IBRel, ordered by maximum confidence level.

miRNA	Gene	Sentences	Documents	Max. Confidence	Correct
hsa-mir-494	CFTR	10	5	0.996	Y
hsa-mir-93	CXCL8	6	1	0.978	Y
hsa-mir-101-1	CFTR	8	3	0.96	Y
hsa-mir-224	SLC4A4	5	1	0.937	Y
hsa-mir-145	CFTR	5	3	0.871	Y
hsa-mir-193b	BRCA1	2	1	0.86	N
hsa-mir-193b	CFTR	2	1	0.857	Y
hsa-mir-155	AKT1	4	1	0.828	Y
hsa-miR-199a-5p	AKT1	5	1	0.807	Y
hsa-mir-183	IDH2	3	1	0.763	Y
hsa-mir-155	CXCL8	5	2	0.736	Y
hsa-mir-125b-1	CFTR	4	1	0.709	Y
hsa-mir-125a	LIN28A	5	1	0.705	N
hsa-mir-224	CFTR	4	1	0.655	Y
hsa-mir-99b	LIN28A	5	1	0.651	N
hsa-mir-99b	KRT18	3	1	0.65	N
hsa-mir-126	TOM1L1	2	1	0.647	Y
hsa-miR-199a-5p	CAV1	5	1	0.642	Y
hsa-miR-509-3p	CFTR	3	2	0.613	Y
hsa-mir-125a	KRT18	3	1	0.58	N
hsa-mir-221	ATF6	3	1	0.546	Y
hsa-mir-145	SMAD3	3	1	0.543	Y
hsa-mir-138-1	CFTR	3	1	0.539	Y
hsa-mir-99b	CFTR	2	1	0.519	Y
hsa-mir-223	CFTR	3	1	0.513	Y
hsa-mir-125a	CFTR	2	1	0.512	Y
hsa-let-7e	LIN28A	3	1	0.508	N

with our method. For example, the supervised classifier was not able to identify the miRNA-gene relations in "Hence, miR-192 and miR-215 can act as effectors as well as regulators of p53" (PMID 19074875), while IBRel identified both relations. Consequentially, we were not able to find any relations described similarly to that example in Bagewadi's corpus. This type of error contributed to the difference in recall. Using a larger corpus, more sentence structures are taken into account, leading to a more flexible classifier.

On the miRTex corpus, our method obtained higher precision but lower recall, resulting in a lower F-score (0.383). It was not possible to train a classifier on this corpus using supervised learning since it was not annotated with relation mentions. For this reason, we used the classifier trained on Bagewadi's corpus. The increase in precision of 0.047 using distant supervision on miRTex corpus reinforces the idea that our approach is more adaptable.

The supervised learning approach obtained higher results on the Bagewadi and miRTex corpora. Since the training set was annotated with the same criteria as the test set, any classifier trained on that training set is more in the line with the test set annotations. The main source of error with the supervised learning approach were sentences where the miRNAs and genes had similar functions. For example, in the sentence "These data implicate hsa-miR-30b, hsa-miR-30d and KHDRBS3 as putative oncogenic target(s) of a novel recurrent medulloblastoma amplicon at 8q24.22-q24.23." (PMID 19584924), there is no miRNA-gene relation, although the words used are similar to the ones that would be used if the relation was between a miRNA and gene.

The co-occurrence approach obtained the highest recall because it classified every miRNA-gene pair in a sentence as a true relation. The precision obtained for Bagewadi and miRTex was close to the other two approaches. This may be due to the fact that since they were manually annotated, the documents were more relevant for the type of relations annotated. The abstracts selected for those two corpora are more likely to contain sentences describing relations than a random selection. Therefore, miRNA-gene pairs in the same sentence would often be related. Compared to IBRel, the co-occurrence approach obtained better F-score on

Bagewadi and miRTex. For the TransmiR corpus, our method outperformed co-occurrence on precision and F-score by 0.212 and 0.163, respectively. The TransmiR corpus has fewer relations per entity than Bagewadi and miRTex (Table 6.1), which may explain why our method performed better than co-occurrence in this case. Our method improved the results of the corpus where the co-occurrence approach was less effective.

Comparing our results to other published results on miRNA-gene relation extraction, the proposed method obtained lower F-score values. For example, Bagewadi et. al. [23] reported an F-score of 0.760 on their corpus. The authors used a linear kernel to obtain that result while using the SL kernel we obtain a similar F-score of 0.757. Using IBRel we obtained a lower F-score 0.532. However, these authors developed and evaluated their approach only on their dataset, which is understandable since they were the first to develop a manually annotated corpus containing information about miRNA-gene relations. Li et. al. [24] developed a rule-based approach to extract document-level relations, obtaining an F-score of 0.94 on their own manually annotated dataset (miRTex corpus) and 0.87 on Bagewadi's corpus. In this case, our best F-score on their dataset was 0.654 using SL kernel and 0.383 using IBRel, which is lower than the values reported by the authors. However, the approach used by these authors cannot be easily adaptable to other domains. This is the reason why in relation extraction community challenges, teams generally use machine learning approaches instead of designing rules specific for that challenge. Since IBRel could be applied to any biomedical relation represented in a knowledge base, it has more reusability than rule-based methods, which are specific for a biological problem.

Since we did not annotate the IBRel-CF corpus, we manually evaluated the results obtained. We identified some relations extracted from the corpus that were not correct. From the 27 relations extracted, we identified 6 errors. There is no mention of the gene BRCA1 in the document where the relation between that gene and hsa-mir-193b was extracted. This was due to a mapping error, where the string "uPA", referring to urokinase plasminogen activator (PLAU), was incorrectly mapped to BRCA1. This error could be fixed using acronym

extension so that the extended form of the gene is mapped instead of the acronym. The three relations with the gene LIN28A are incorrect. Although this gene regulates the expression of several miRNAs, those relations were not described in the text. This error occurred because some miRNAs were recognized as genes, and in this case, they were incorrectly mapped to the LIN28A gene. One possible solution to this problem is to use semantic similarity to improve the mapping process. Considering that entities mentioned in the same sentence should be semantically related, PLAU would be more semantically similar to the other genes mentioned than BRCA1. Therefore, semantic similarity could be used as a threshold to choose better mappings.

6.4.1 Evaluation of miRNA and Gene Entity Recognition

We were able to recognize miRNA and gene entities from the three corpora. Regarding miRNAs, this task was not difficult since miRNA nomenclature is standardized and thus not as ambiguous as other biomedical entities. In the case of Bagewadi's corpus, the F-score obtained was similar to the reported inter-annotator agreement (difference of 0.002 for miRNA and 0.075 for gene). On the miRTex corpus, we obtained higher F-score values for both miRNA and gene entities. The results obtained with the TransmiR corpus were lower since this evaluation was limited by some factors. The main one was the fact that not all relations were mentioned in the abstract of the articles. For example, every document with a relation between hsa-let-7a-1 and a gene also contained a relation between other miRNAs from the let-7 family and that gene. However, this was never mentioned in the abstract. This error accounted for 42 false negatives. Another type of error was due to some miRNAs and genes mentioned in the abstract that were not part of the TransmiR database. For example, PMID 20093556 mentions 6 miRNAs, but only one miRNA-gene relation exists in the database. This type of error contributed to lower precision values for miRNA and gene entity recognition when compared to the other two corpora.

6.5 Conclusion

In this paper, we showed that our method performed better on a dataset based on a manually curated database, while, as expected, supervised learning performed better on manually annotated datasets, developed for text mining applications. The main contributions of this paper are IBRel, a method for extraction of biomedical relations from texts using only existing resources, and a dataset of miRNA-gene relations automatically extracted and manually validated. The method we developed was evaluated for miRNA-gene relation extraction, where it outperformed supervised learning on the case where no specific training set was available.

A second contribution is the dataset obtained using our method for cystic fibrosis. We applied IBRel to a set of 51 abstracts about cystic fibrosis, published in the last 5 years. From the 27 miRNA-gene relations extracted, 21 of those were found to be correct in the context of cystic fibrosis. While this approach was not flawless, it should be of interest to researchers working on this subject since there are still few reliable resources for identifying miRNA-gene relations in disease-specific contexts. We intend to apply this approach to other diseases and develop a platform to visualize the information extracted.

The results obtained in this work suggest that our method can still be improved. For example, we can optimize the parameters of miSVM to this task using cross-validation on the datasets used. We intend to use ontologies to better annotate the corpus generated for distant supervision. Semantic similarity has been used before to extract protein-protein interactions [52] and drug-target interactions [53]. By computing the semantic similarity between the entities mentioned in a document, we can identify which are more likely to be associated. The similarity between two genes can be calculated using the semantic similarity between the two sets of Gene Ontology terms annotated to them. We have previously explored semantic similarity techniques for drug name recognition [54] and drug-drug relation extraction [50]. We used semantic similarity between two chemical entities on ChEBI as a feature for an ensemble classifier, obtaining higher precision values. Another type of approach we wish

to explore is crowd-sourcing. Other authors have used crowd-sourcing to improve multi-instance learning results [55]. The idea is to use machine learning algorithms to correctly classify a wide range of cases and use crowd-sourcing to solve the most difficult cases.

Acknowledgments

This work was supported by the Portuguese Fundação para a Ciência e Tecnologia (https://www.fct.mctes.pt/) through the PhD Grant ref. PD/BD/106083/2015 to AL, PEst-OE/BIA/UI4046/2011 (BioFIG) to LAC, and UID/CEC/00408/2013 (LaSIGE) to AL and FMC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] Isabel Segura Bedmar, Paloma Martínez and María Herrero Zazo. 'Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)'. In: *Proceedings of the Seventh International Workshop on Semantic Evaluation*. Association for Computational Linguistics. 2013.

[2] Martin Krallinger et al. 'Overview of the protein-protein interaction annotation extraction task of BioCreative II'. In: *Genome Biol* 9.2 (2008), p. 1.

[3] Markus Bundschus et al. 'Extraction of semantic biomedical relations from text using conditional random fields'. In: *BMC Bioinformatics* 9.1 (2008), p. 1.

[4] Dmitry Zelenko, Chinatsu Aone and Anthony Richardella. 'Kernel methods for relation extraction'. In: *J Mach Learn Res* 3.Feb (2003), pp. 1083–1106.

[5] Claudio Giuliano, Alberto Lavelli and Lorenza Romano. 'Exploiting shallow linguistic information for relation extraction from biomedical literature'. In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Vol. 18. 2006. Citeseer. 2006, pp. 401–408.

[6] Alessandro Moschitti. 'A study on convolution kernels for shallow semantic parsing'. In: *Proc Conf Assoc Comput Linguist Meet*. Association for Computational Linguistics. 2004, p. 335.

[7] David P Bartel. 'MicroRNAs: genomics, biogenesis, mechanism, and function'. In: *Cell* 116.2 (2004), pp. 281–297.

[8] Qinghua Jiang et al. 'miR2Disease: a manually curated database for microRNA deregulation in human disease'. In: *Nucleic Acids Res* 37.suppl 1 (2009), pp. D98–D104.

[9] Lin He and Gregory J Hannon. 'MicroRNAs: small RNAs with a big role in gene regulation'. In: *Nat Rev Genet* 5.7 (2004), pp. 522–531.

[10] George S Mack. 'MicroRNA gets down to business'. In: *Nat Biotechnol* 25.6 (2007), pp. 631–638.

[11] Ana Kozomara and Sam Griffiths-Jones. 'miRBase: annotating high confidence microRNAs using deep sequencing data'. In: *Nucleic Acids Research* 42.D1 (2014), pp. D68–D73. DOI: 10.1093/nar/gkt1181. URL: http://dx.doi.org/10.1093/nar/gkt1181.

[12] Yang Li et al. 'HMDD v2.0: a database for experimentally supported human microRNA and disease associations'. In: *Nucleic Acids Res* (2013), pp. D1070–4.

[13] Chih-Hung Chou et al. 'miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database'. In: *Nucleic acids research* 44.D1 (2016), pp. D239–D247.

[14] Rajesh Chowdhary et al. 'A database of annotated promoters of genes associated with common respiratory and related diseases'. In: *American Journal of Respiratory Cell and Molecular Biology* 47.1 (2012), pp. 112–119. ISSN: 10441549. DOI: 10.1165/rcmb.2011-0419OC.

[15] Xiangxiang Zeng, Xuan Zhang and Quan Zou. 'Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks'. In: *Briefings in bioinformatics* 17.2 (2016), pp. 193–203.

[16] Quan Zou et al. 'Similarity computation strategies in the microRNA-disease network: a survey'. In: *Briefings in functional genomics* 15.1 (2016), pp. 55–64.

[17] Mohamed Hamed et al. 'TFmiR: A web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks'. In: *Nucleic Acids Research* 43.W1 (2015), W283–W288. ISSN: 13624962. DOI: 10.1093/nar/gkv418.

[18] Y Liu et al. 'Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources'. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* PP.99 (2016), p. 1. ISSN: 1545-5963. DOI: 10.1109/TCBB.2016.2550432.

[19] Li Guo et al. 'miR-isomiRExp: a web-server for the analysis of expression of miRNA at the miRNA/isomiR levels'. In: *Scientific Reports* 6.February (2016), p. 23700. ISSN: 2045-2322. DOI: 10.1038/srep23700. URL: http://www.nature.com/articles/srep23700.

[20] B Stuart Murray et al. 'An in silico analysis of microRNAs: mining the miRNAome'. In: *Mol Biosyst* 6.10 (2010), pp. 1853–1862.

[21] Haroon Naeem et al. 'miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature'. In: *BMC Bioinformatics* 11.1 (2010), p. 1.

[22] Jingshan Huang et al. 'A semantic approach for knowledge capture of MicroRNA-Target gene interactions'. In: *Proceedings - 2015 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015* (2015), pp. 975–982. DOI: 10.1109/BIBM.2015.7359816.

[23] Shweta Bagewadi et al. 'Detecting miRNA mentions and relations in biomedical literature'. In: *F1000Research* 3 (2014).

[24] Gang Li et al. 'miRTex: A Text Mining System for miRNA-Gene Relation Extraction'. In: *PLoS Comput Biol* 11.9 (2015), e1004391.

[25] Benjamin Rozenfeld and Ronen Feldman. 'Clustering for unsupervised relation identification'. In: *Proceedings of the sixteenth ACM conference on information and knowledge management*. ACM. 2007, pp. 411–418.

[26] Benjamin Rozenfeld and Ronen Feldman. 'High-performance unsupervised relation extraction from large corpora'. In: *Proceedings - IEEE International Conference on Data Mining, ICDM* (2006), pp. 1032–1037. ISSN: 15504786. DOI: 10.1109/ICDM.2006.82.

[27] Mike Mintz et al. 'Distant supervision for relation extraction without labeled data'. In: *Proc Conf Assoc Comput Linguist Meet*. Association for Computational Linguistics. 2009, pp. 1003–1011.

[28] Thomas G Dietterich, Richard H Lathrop and Tomás Lozano-Pérez. 'Solving the multiple instance problem with axis-parallel rectangles'. In: *Artificial intelligence* 89.1 (1997), pp. 31–71.

[29] Sebastian Riedel, Limin Yao and Andrew McCallum. 'Modeling relations and their mentions without labeled text'. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2010, pp. 148–163.

[30] Bonan Min et al. 'Distant Supervision for Relation Extraction with an Incomplete Knowledge Base'. In: *Proceedings of NAACL-HLT*. 2013, pp. 777–782.

[31] Mark Craven and Johan Kumlien. 'Constructing biological knowledge bases by extracting information from text sources.' In: *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB-99)* (1999), pp. 77–86. ISSN: 1553-0833. URL: http://www.aaai.org/Papers/ISMB/1999/ISMB99-010.pdf.

[32] Ke Ravikumar et al. 'Literature mining of protein-residue associations with graph rules learned through distant supervision.' In: *Journal of biomedical semantics* 3 Suppl 3.Suppl 3 (2012), S2. ISSN: 2041-1480. DOI: 10.1186/2041-1480-3-S3-S2. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3465209%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[33] Hoifung Poon, Kristina Toutanova and Chris Quirk. 'Distant supervision for cancer pathway extraction from text'. In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 20.January (2015), pp. 120–31. ISSN: 2335-6936. DOI: 10.1142/9789814644730_0013. URL: http://www.ncbi.nlm.nih.gov/pubmed/25592574.

[34] Witold Filipowicz, Suvendra N Bhattacharyya and Nahum Sonenberg. 'Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?' In: *Nat Rev Genet* 9.2 (2008), pp. 102–114.

[35] Shyam Ramachandran et al. 'Post-transcriptional regulation of cystic fibrosis transmembrane conductance regulator expression and function by microRNAs'. In: *"Am J Respir Cell Mol Biol* 49.4 (2013), pp. 544–551.

[36] Juan Wang et al. 'TransmiR: a transcription factor–microRNA regulation database'. In: *Nucleic Acids Res* 38.suppl 1 (2010), pp. D119–D122.

[37] The UniProt Consortium. 'UniProt: a hub for protein information'. In: *Nucleic Acids Research* 43.D1 (2015), pp. D204–D212. DOI: 10.1093/nar/gku989. URL: http://dx.doi.org/10.1093/nar/gku989.

[38] Jin-Dong Kim et al. 'Overview of BioNLP'09 shared task on event extraction'. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*. Association for Computational Linguistics. 2009, pp. 1–9.

[39] David MW Powers. 'The problem with kappa'. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2012, pp. 345–355.

[40] Razvan C Bunescu and Raymond J Mooney. 'Multiple instance learning for sparse positive bags'. In: *Proceedings of the 24th international conference on Machine learning*. ACM. 2007, pp. 105–112.

[41] Yu-Ju Chen and Jane Yung-jen Hsu. 'Chinese relation extraction by multiple instance learning'. In: *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*. 2016, pp. 105–112.

[42] F. Pedregosa et al. 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[43] Grace Wahba. 'Multivariate function and operator estimation, based on smoothing splines and reproducing kernels'. In: *Nonlinear Modeling and Forecasting, SFI Studies in the Scienes of Complexity*. Vol. 12. Addison-Wesley. 1992, pp. 95–112.

[44] Arun Ramani et al. 'Integrating Co-occurrence Statistics with Information Extraction for Robust Retrieval of Protein Interactions from Medline'. In: *Proceedings of the HLT-NAACL Workshop on Linking Natural Language Processing and Biology: Towards deeper biological literature analysis (BioNLP-2006)* June (2006), pp. 49–56. DOI: 10.3115/1567619.1567628.

[45] Isabel Segura-Bedmar, Paloma Martinez and Cesar de Pablo-Sánchez. 'Using a shallow linguistic kernel for drug–drug interaction extraction'. In: *Journal of biomedical informatics* 44.5 (2011), pp. 789–804.

[46] Robert Leaman, Graciela Gonzalez et al. 'BANNER: an executable survey of advances in biomedical named entity recognition.' In: *Pacific symposium on biocomputing*. Vol. 13. 2008, pp. 652–663.

[47] John Lafferty, Andrew McCallum and Fernando Pereira. 'Conditional random fields: Probabilistic models for segmenting and labeling sequence data'. In: *Proc Int Conf Mach Learn*. Vol. 1. 2001, pp. 282–289.

[48] Rune Sætre et al. 'AKANE system: protein-protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask'. In: *Proceedings of the Second BioCreative Challenge Workshop*. Madrid. 2007, pp. 209–212.

[49] Christopher D. Manning et al. 'The Stanford CoreNLP Natural Language Processing Toolkit'. In: *Association for Computational Linguistics (ACL) System Demonstrations*. 2014, pp. 55–60. URL: http://www.aclweb.org/anthology/P/P14/P14-5010.

[50] Andre Lamurias, João D Ferreira and Francisco M Couto. 'Identifying interactions between chemical entities in biomedical text'. In: *J Integr Bioinform* 11.3 (2014), p. 247.

[51] Robert Leaman, Chih-Hsuan Wei and Zhiyong Lu. 'NCBI at the BioCreative IV CHEMDNER Task: Recognizing chemical names in PubMed articles with tmChem'. In: *BioCreative Challenge Evaluation Workshop*. Vol. 2. 2013, p. 34.

[52] Shobhit Jain and Gary D Bader. 'An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology.' In: *BMC bioinformatics* 11.1 (2010), p. 562. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-562. URL: http://www.biomedcentral.com/1471-2105/11/562.

[53] Guillermo Palma, Maria-esther Vidal and Louiqa Raschid. 'Drug-Target Interaction Prediction Using Semantic Similarity and Edge Partitioning'. In: *ISWC* 8796 (2014), pp. 131–146.

[54] Andre Lamurias, João D. Ferreira and Francisco M. Couto. 'Improving chemical entity recognition through h-index based semantic similarity'. In: *Journal of Cheminformatics* 7.Suppl 1 (2015), S13. ISSN: 17582946. DOI: 10.1186/1758-2946-7-S1-S13. URL: http://www.jcheminf.com/content/7/S1/S13.

[55] Truc-Vien T Nguyen and Alessandro Moschitti. 'Joint distant and direct supervision for relation extraction.' In: *Proceedings of the 5th International Joint Conference on Natural Language Processing*. 2011, pp. 732–740.

7

Generating a Tolerogenic Cell Therapy Knowledge Graph from Literature

ANDRE LAMURIAS, JOÃO D. FERREIRA, LUKA A. CLARKE, FRANCISCO M. COUTO

Abstract

Tolerogenic cell therapies provide an alternative to conventional immunosuppressive treatments of autoimmune disease and address, among other goals, the rejection of organ or stem cell transplants. Since various methodologies can be followed to develop tolerogenic therapies, it is important to be aware and up to date on all available studies that may be relevant to their improvement. Recently, knowledge graphs have been proposed to link various sources of information, using text mining techniques. Knowledge graphs facilitate the automatic retrieval of information about the topics represented in the graph.

The objective of this work was to automatically generate a knowledge graph for tolerogenic cell therapy from biomedical literature. We developed a system, ICRel, based on ma-

chine learning to extract relations between cells and cytokines from abstracts. Our system retrieves related documents from PubMed, annotates each abstract with cell and cytokine named entities, generates the possible combinations of cell-cytokine pairs co-occurring in the same sentence, and identifies meaningful relations between cells and cytokines. The extracted relations were used to generate a knowledge graph, where each edge was supported by one or more documents. We obtained a graph containing 647 cell-cytokine relations, based on 3264 abstracts. The modules of ICRel were evaluated with cross-validation and manual evaluation of the relations extracted. The relation extraction module obtained an F-measure of 0.789 in a reference database, while the manual evaluation obtained an accuracy of 0.615. Even though the knowledge graph is based on information that was already published in other papers about immunology, the system we present is more efficient than the laborious task of manually reading all the literature to find indirect or implicit relations. The ICRel graph will help experts identify implicit relations that may not be evident in published studies.

7.1 Introduction

Tolerogenic cell therapies provide an alternative to conventional immunosuppressive treatments of autoimmune disease and address, among other goals, the rejection of organ or stem cell transplants [1]. These therapies aim at modulating the pathological immune response with minimal effect on the immune system. Antigen-presenting cells (APCs) can be induced to control the immune response by targeting specific T cell responses, avoiding general suppression of the immune system [2]. It is necessary to understand the underlying mechanisms of the immune system to develop tolerogenic cell therapies. Cytokines are small peptides involved in cell signaling, which can be used to induce tolerance in APCs [3]. Immune cells express cytokines and their respective receptors. High-throughput sequencing techniques have improved our knowledge about cell signaling, introducing a variety of

information about how cytokines are used by the immune system. This information is important to understand and develop new methods to isolate, culture and induce tolerance in APCs.

Biomedical information is often presented to the community through published literature, including information about human autoimmune diseases and therapies to treat them. There are knowledge bases aiming at organizing the findings provided by the literature through a single access point. Populating such knowledge bases is, therefore, important for biomedical research, in particular, because they allow computational methods to find patterns in the data, thus generating new hypotheses to be tested experimentally. If a cell produces the same cytokine receptors as another cell, and a new cytokine is found to interact with the first cell, it is plausible that new cytokine could also affect the second cell. This type of inference, also known as ABC model [4], is only possible if the results of many studies are analyzed together.

The scientific community has shown interest in curating databases about cells and cytokines. For example, the National Center for Biotechnology Information (NCBI) provides a compilation of several biomedical and genomic resources [5], including the Entrez Gene database[6]. This database contains entries for the genes associated with cytokines, and each entry contains useful information about that cytokine, such as interactions, pathways, and gene ontology annotations. There are also resources specific to cytokine information. The Cytokine Reference is an online database of information on cytokines and receptors, compiled from the literature by experts [7]. This database contains links to other databases such as MEDLINE and GenBank, and can be searched by cytokine, cell or disease. Another relevant database is the Cytokine & Cells Online Pathfinder Encyclopedia (COPE) [8], which focuses on the interactions between cell types through cytokines. The current version of COPE contains 45k entries, including a cell type dictionary of 3k entries. These efforts show the importance of information structures for cells and cytokines. Therefore, the development of computational methods to structure this information would benefit researchers working in

this domain.

These computational methods require two conditions: (i) the information is readable by computers, and (ii) it is comprehensive, encoding the up-to-date collective knowledge of the community. Both these tasks are currently subject to intensive research. Converting heterogeneous data formats to a common language and merging the data is one approach to the first task. For example, Bio2RDF converts heterogeneous data from several datasets into RDF, a standard data model based on the specification of links between data elements [9].

As for the second task, the information stored in many biomedical datasets is the result of manual processing of documents, which is becoming less practical, since the number of published documents increases at a high rate. A more feasible approach is to use automatic text mining methods to process documents and generate a knowledge graph for a given topic. In a knowledge graph, nodes correspond to real world entities while edges represent relationships between the entities. A widely popular knowledge graph is the one integrated with Google search. This graph is generated from web documents, and organizes information about various topics, such as people, places, and works of art, to improve the quality of the search results delivered to the users[1]. Recent works have demonstrated how biological knowledge graphs can be extracted from documents, based on protein-protein, [10], miRNA-gene [11] and drug-target interactions [12]. While these graphs provide important efforts to link the discoveries of various manuscripts, there is still a need for automatic methods that can create specialized graphs and update them as more works are published.

This manuscript presents the system, ICRel (Identifying Cellular Relations), that we developed, based on machine learning, to extract cell-cytokine relations from documents and generate a knowledge graph. ICRel was trained and evaluated with the immuneXpresso database to extract meaningful relations between cells and cytokines in documents. We did not aim at finding novel information, instead we demonstrate the utility of the system by studying the graph generated by ICRel, in particular, the nodes associated with APCs. Therefore,

[1] https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html

the contributions of this manuscript are: (i) the open source ICRel system that generates a cell-cytokine graph from biomedical abstracts, and (ii) the knowledge graph obtained using ICRel on a set of documents relevant to tolerogenic antigen-presenting cell therapy. ICRel was able to identify cytokines associated with tolerogenic antigen presenting cells that were missing from the immuneXpresso database. The code and results obtained with ICRel are available at https://github.com/lasigeBioTM/ICRel.

7.2 Material and Methods

The objective of ICRel is to automatically generate a knowledge graph relevant to tolerogenic cell therapy from a given corpus. The system was written in Python 3.5 and its code is openly available[2]. The methodology used can be adapted to other domains, by selecting an appropriate set of documents and reference database. Figure 7.1 presents the pipeline of ICRel, describing the input and output of each module, while Figure 7.2 provides an example of an abstract being processed by each module. The first module retrieves abstracts from PubMed into an internal database, according to a given query specified as input. The second module identifies named entities with an external tool, requiring one lexicon for each entity type to be identified. In this case, we had a lexicon for cell names and another for cytokines. The third module combines all cell-cytokine pairs identified within a sentence to generate instances for the machine learning classifier and to calculate the pair frequency score. Finally, the fourth module classifies each pair, assigns a confidence score and generates a graph based on the pairs that were classified as positive. The remainder of this section describes in detail the data and methods used to develop this system.

[2] https://github.com/lasigeBioTM/ICRel

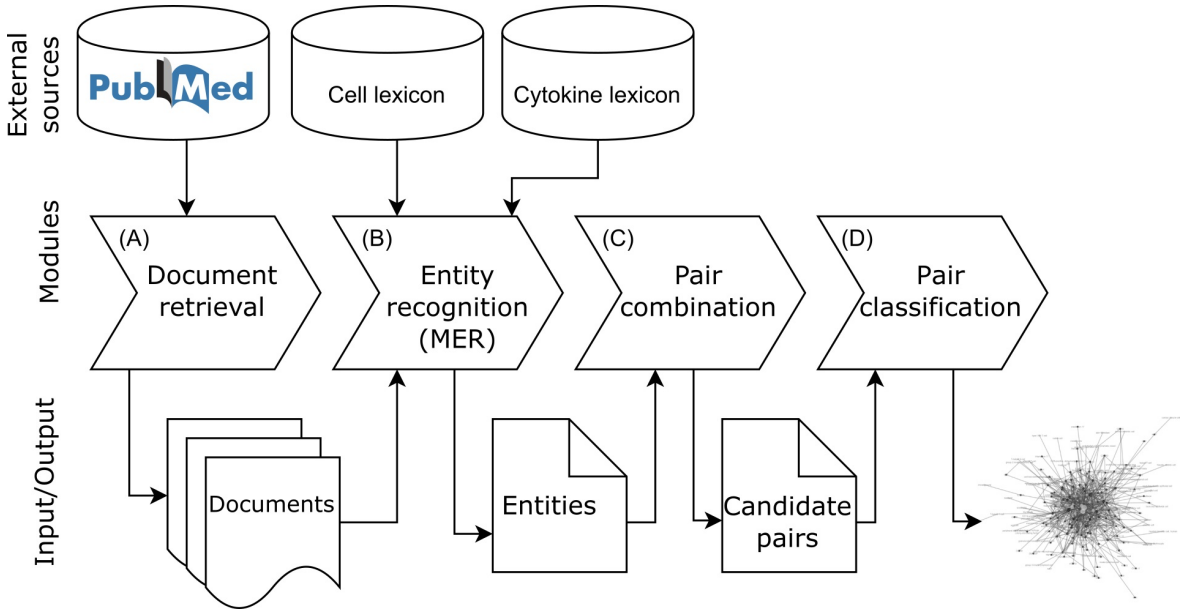


Figure 7.1: Pipeline of the ICRel system. This first module (A) retrieves documents from PubMed, the second module (B) annotates cell and cytokine entities in each document using the Cell Ontology and Cytokine registry, the third module (C) combines the cells and cytokines mentioned in the sentence and the fourth module (D) classifies each pair and generates the graph.

7.2.1 Datasets

A previous study provided a database of interactions between cytokines and cells, named immuneXpresso [14]. Although this database was generated using automatic information extraction methods, its contents were evaluated with two manually curated databases, regarding the interactions containing B cells. The authors obtained a 20% false negative rate and no false positives. Even though we have no other guarantee that all entries of this database are correct, we considered this database as a silver standard due to the evaluation scores reported by the authors. A gold standard would require each entry to be manually validated by different domain experts. Since we could not find a gold standard for cytokine-cell interactions in abstracts, we used this silver standard to train and evaluate our method using 5-fold cross-validation. In previous studies, this type of methodology has been shown to be

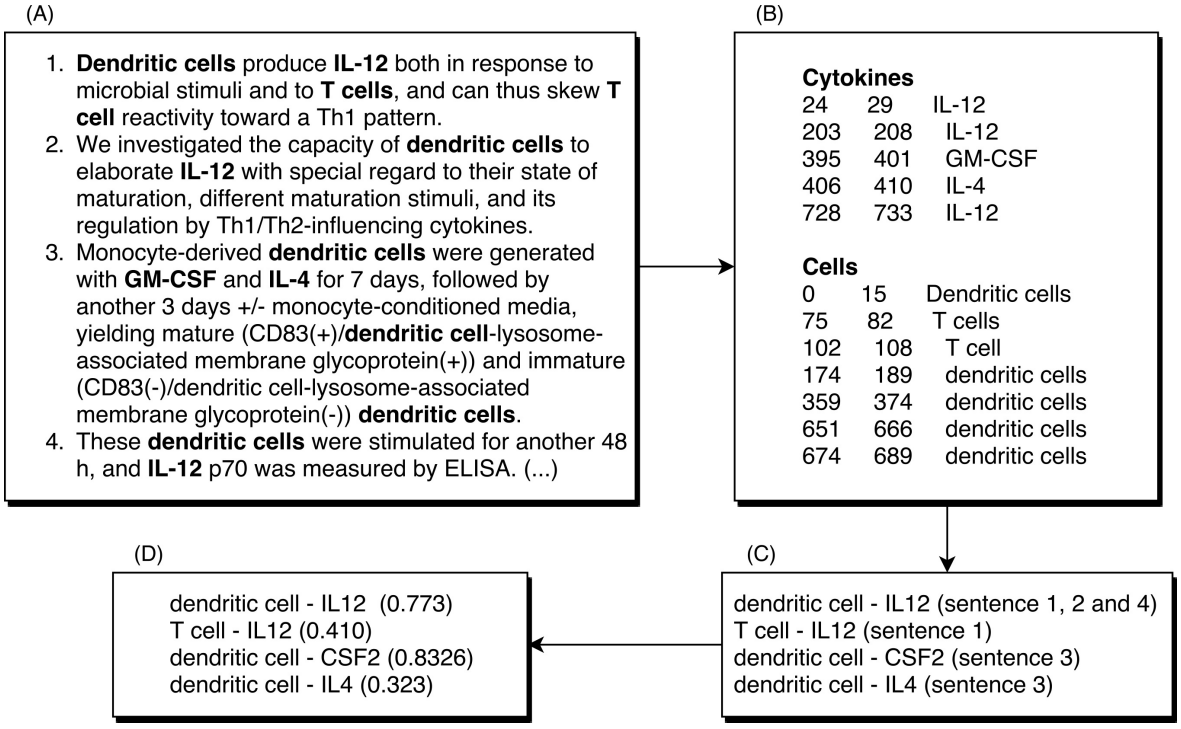


Figure 7.2: Example of an abstract being processed by the ICRel system. We show the first four sentences of the abstract of the article [13]. The first box (A) shows these sentences, numbered and with cells and cytokines bolded manually. The second box (B) shows the entities recognized automatically, where the numbers at the start of each line represent the first and last character offset of the entity. The third box (C) shows the possible cell-cytokine combinations using the sentences shown. The fourth box (D) shows the confidence scores obtained with our system for those pairs. It should be noted that those scores were obtained using several documents and not just the example shown.

useful for information extraction evaluations [15, 16].

Each entry of the immuneXpresso database represents an interaction between a cytokine and a cell found in the literature. The interactions are supported by one or more abstracts, and they have the following attributes: direction (cell to cytokine or vice-versa), sentiment (Positive, Negative or Unknown), number of papers, and e-score. The sentiment reflects if the interaction indicates up-regulation (positive) or down-regulation (negative). Each interaction can be found in the associated abstracts, in at least one sentence mentioning both the cytokine

and cell. We retrieved these abstracts from PubMed and associated each entry with the respective abstracts. A total of 25,347 abstracts were considered for this silver standard.

Our main objective was to develop an automatic system to generate a knowledge graph about cellular tolerogenic therapies, focusing on those that use APCs. Hence, we retrieved a corpus of documents related to this topic using the MeSH term "Antigen-Presenting Cells", which should include most published abstracts with information relevant to our graph. We restricted this query to abstracts published from January 2015 to August 2017, to avoid overlapping with immuneXpresso, which has no abstracts published after 2015. Using this query, we obtained 3264 abstracts, which were then annotated with cytokine and cell named entities. We expect that the information obtained by our system can be complementary to this database, which is not focused on any specific topic besides immunology. Furthermore, our system can automatically process new abstracts and add new relations to the graph.

7.2.2 Named entity recognition

Each abstract of our datasets contained named entities corresponding to concepts relevant to tolerogenic cell therapies. We were interested specifically in references to cells and cytokines in these abstracts. To this end, we established a lexicon of cell and cytokine names. The cell lexicon is based on the Cell Ontology [17] (version: 2017-07-29). We compiled all the concept labels and corresponding synonyms, resulting in a total of 8503 terms. For cytokines, we used a cytokine registry[3], which includes several synonyms for each cytokine, corresponding to a total of 7242 terms (version: November 2015). In both cases, each synonym was mapped to a reference string: Cell Ontology concept label in the case of cells and Entrez name in the case of cytokines. This way, we could associate the same entities mentioned across various documents through different synonyms, as long as those synonyms were considered in our lexicon.

We employed MER [18] to identify named entities in the abstracts. MER matches a list

[3] http://immport.org/immport-open/public/reference/cytokineRegistry

of terms (lexicon) to their mentions in the text, returning the characters of the entities found. For example, in the sentence "The dendritic cells were safely tolerated.", MER would return the characters from 4 to 19, which correspond to the text "dendritic cells". Figure 7.2B shows an example of the output of MER for an abstract. This tool has the advantage of being easy to adapt to any entity type, it does not require annotated training data, and it is lightweight in terms of computational resources. We ran MER for each entity type (cell and cytokine) on each abstract. Due to its simplicity, MER has some limitations, for example, it is not able to use context to recognize entities, and it is susceptible to orthographic variations. To increase the number of entities recognized, we added plural variants of every cell name to the lexicon with the Python package inflect. This way, in the previous example, "dendritic cells" would be matched to the "dendritic cell" concept of the Cell Ontology, even if the text is not a perfect match. Furthermore, we removed common words such as "light" and "killer" from the cytokine lexicon, since these words could also appear in other contexts, for example, as part of "natural killer cell". We found these words by comparing the lexicon to a list of common English words. The main limitation of MER is that the lexicon may be incomplete and some references to cells and cytokines in the documents will be missed. However, by using a large corpus, our assumption is that only rare variants will not be identified since most journals recommend a specific nomenclature for cells and proteins.

7.2.3 Cell-cytokine relation extraction

A classifier is a model capable of assigning labels to new data according to a specific function learned from the training data. Supervised machine learning algorithms learn to classify instances (in this case, pairs) by adjusting a function to the labels of each instance of the training set. Generally, these algorithms require the training data to consist of a matrix where each line corresponds to an instance and each column to a feature. We consider an instance to be a specific combination of cell and cytokine, while the features consist of the words used in sentences where that pair co-occurs. A classifier should be evaluated to

understand how useful it can be to predict the labels of new data. This type of evaluation is done by comparing the real labels assigned by experts to the labels predicted by the classifier. Figure 7.3A shows the workflow of the training and evaluation process of a supervised machine learning approach using 5-fold cross-validation. Cross-validation consists of iteratively partitioning the dataset in folds, using all but one of the folds to train a classifier. This classifier is used to predict labels for the remaining fold, which are then compared to the original labels. In a 5-fold cross-validation, this process is repeated 5 times, and an average of the scores obtained in each iteration is used to estimate the quality of the classifier. Afterwards, a classifier can be trained using the whole dataset.

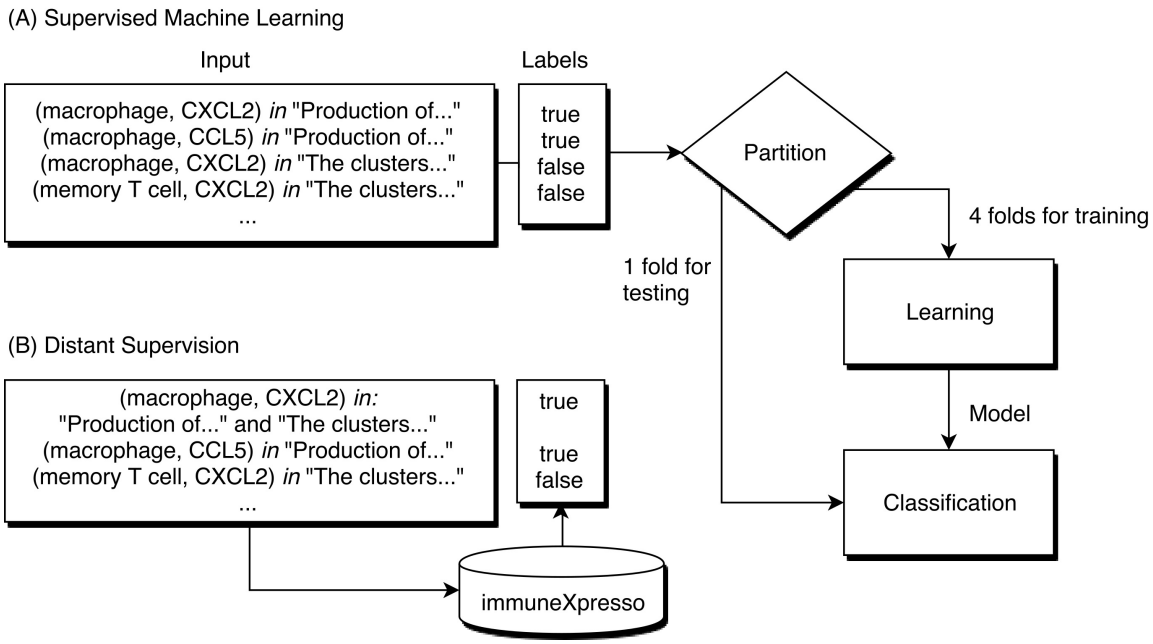


Figure 7.3: Demonstration of a machine learning workflow for cell-cytokine pair classification. (A) The label of each pair is known, and the learning algorithm trains a classifier based on these labels. Using 5-fold cross-validation, at each iteration 4 folds are used for training and 1 for testing. (B) Using distant supervision, the labels of each instance are not known, instead, a database assigns a label according to the existence of an entry corresponding to that pair.

We consider a knowledge graph to be a set of facts associated with a specific domain

using the RDF data model, i.e., specified by predicate-verb-object triplets. In our case, the knowledge graph is constituted by cell-cytokine interactions, where the focus is on the predicate and objects, which are cells and cytokines, with no specific order. An instance is any co-occurrence of a specific cell-cytokine pair within a sentence. We consider various types of relations, where a cell expresses a cytokine, or a cytokine affects the behavior of a cell. We are interested only in direct relations, where there are no intermediaries to the relation described. This includes cases of up- and down-regulation, signaling, activation, and stimulation, for example. However, we are not interested in cases where the relation is negated (e.g. the cell does not express the cytokine) or hypothetical (e.g. the authors consider that a similar cell may express the same cytokine). For each pair, at least one sentence must explicitly state the existence of the relation for it to be considered a positive instance. That sentence may contain other information, such as the mechanism of the relation, experimental details or other cells and cytokines.

Distant supervision assumes that if a relation between two entities is stated in a database, it can be assumed that whenever those two entities co-occur in a document a relation between them is described (Figure 7.3B). We used distant supervision to generate a dataset for training since it is not easy to obtain labeled training data for most domains For example, it would be assumed that every sentence in the abstract of the paper [13] that mentions both dendritic cells and IL-12 is supporting that relation, including this sentence: “These dendritic cells were stimulated for another 48 h, and IL-12 p70 was measured by ELISA”. Although this assumption does not take into account the semantics of the text, it has been shown that distant supervision can be useful to extract relations from documents [19]. In this work, we adopted immuneXpresso as the reference database. As previously mentioned, this database was generated automatically, however, the authors report a high accuracy when compared to experimental data.

The machine learning algorithm used by ICRel, multi-instance learning (MIL), organizes instances in bags, which consist simply of sets of instances with a common property. All

instances are negative if the bag label is negative, or at least one instance is positive if the bag label is positive. Therefore, there is no need to manually label the relations in the documents. This approach can be applied to relation extraction, assuming that the instances are potential relations and the bags contain instances of the same pair of entities. Figure 7.2C shows an example of the way the instances are organized in bags, where each line corresponds to a different bag. Each bag has a label, which can be positive if the database contains an entry establishing a relation between the two elements of the bag, or negative otherwise. Using a machine learning algorithm, a classifier can be trained to classify new instances. This classifier will assign a confidence score to each bag. It is a reasonable assumption that an interaction is stated in a single sentence, so we consider only pairs of entities mentioned within a sentence.

Besides the labels of each bag, the MIL algorithm uses a feature representation of each instance to train a classifier. In our case, the feature representation of each instance is based on a window of words around each entity of the pair. We used a context window of size three, meaning that at most three words before and after each entity were considered. Each word was represented by its lemma so that variations of the same root word did not affect the learning process. Words that were part of named entities were represented by their respective entity type, to avoid any bias towards specific entities, and words that appeared in less than 1% of the documents were not considered, to reduced noise caused by text artifacts. Then, we generated tf-idf weights for each word, to obtain a vector representation of each instance. Tf-idf corresponds to the product between term frequency (tf) and inverse document frequency (idf), and it is used to estimate the relative importance of each word in a corpus. This is required since machine learning algorithms require numeric vectors. The weights generated during the training phase were also applied to new data. In summary, each document was converted to sets of instances (bags), with each instance corresponding to a feature vector obtained with tf-idf weighting.

We observed that only some sentences in each abstract described relations between cells

and cytokines, while the other sentences presented other types of information, such as definitions or experimental parameters. This would be an issue to traditional approaches relation extraction because there is a larger proportion of negative pairs (no direct and explicit relation is described in the text) than positive. In our preliminary experiments, we found that often less than 10% of the pairs in a document are positive. Therefore, it was necessary to use an algorithm that takes into account the sparsity of the data. We tested variations of MIL and found that sparse MIL (sMIL) [20] provided the best results. This algorithm is based on support vector machines, with an adapted objective function to account for the reduced number of positive labels. This new cost function assumes that smaller positive bags are more informative, weighting the feature vector of each positive bag according to its number of instances.

Our system contains a classifier trained using all entries and documents from the immuneXpresso database, corresponding to about 25k abstracts, using the methods described above. ICRel extracts relations from documents by transforming the text into feature vectors and then applying this classifier. The trained classifier predicts the label of a bag but does not predict the individual label of its instances. This means that it is not possible to know the exact sentence where the interaction is described. However, this information is sufficient for our purposes, since we know that each extracted relation has at least one sentence supporting it.

We used two different measures to classify an instance: the confidence score assigned by the machine learning classifier, and the number of sentences associated with a pair, which we call the pair frequency. The classifier confidence score was based on the distance to the hyperplane given by the sMIL algorithm, as described in [21]. The pair frequency was calculated as the number of abstracts where that pair co-occurs in a sentence divided by the total number of abstracts in the corpus. We expect that pairs mentioned in more documents are more likely to have been correctly identified. Both scores were used to study how precision and recall varies when using a threshold. As the threshold increases, recall should decrease

while precision increases.

7.2.4 Knowledge graph for tolerogenic cell therapy

The proposed ICRel system can extract candidate entries to generate a cytokine-cell graph. Each candidate entry is supported by the sentences where it was found, a classifier confidence score and its frequency. Since each cell and cytokine entity was normalized to a reference database, we can associate relations described over many documents, even if the authors use various nomenclatures. Furthermore, since we used the Cell Ontology as the reference for cell names, its axioms can be explored to expand the graph.

To obtain a knowledge graph for tolerogenic cell therapy, we first obtained a set of 3264 documents about APCs. This set of documents does not overlap with the documents used to train the classifier, which includes only documents published before 2015. The same documents should not be used for training and testing machine learning classifiers because the classifier will have a biased performance on the training documents, leading to an overestimation of the quality of the results. Instead, we can simply match the immuneXpresso relations with our graph to obtain more knowledge.

The extracted relations were imported to Cytoscape [22] to visualize the graph. The ICRel graph is an undirected bipartite graph where each edge corresponds to a cell-cytokine relation. We compared our graph to the one obtained with immuneXpresso, by considering it also as an undirected graph. We computed standard properties of the two graphs, such as diameter and center nodes, with the Python package NetworkX [23]. Furthermore, since our system is focused on obtaining information about tolerogenic cell therapies, we explored the information contained by each graph relevant to this type of therapy.

We considered that a manual evaluation of the automatically generated knowledge graph was necessary to estimate the quality of the information. We sampled a set of 60 edges to be manually validated by three human curators. Each curator validated 30 edges, with a set of 15 edges common to all three, to calculate the inter-annotator agreement. Each curator

accepted an edge if there was at least one sentence supporting it in the corpus, and rejected otherwise. We asked to classify the cause of each rejection to understand the sources of error of our graph. The inter-annotator agreement was measured using Fleiss' kappa, an adaptation of Cohen's kappa for multiple annotators [24]. The classifications of the curators were used to estimate the accuracy of the graph.

7.3 Results

The silver standard described in section 7.2.1 is composed of 25,347 abstracts and a total of 4445 cell-cytokine relations, without considering direction or any other attribute. The silver standard did not contain any information about entities mentioned in the abstracts that did not participate in cell-cytokine relations. We identified 185,243 cells and 189,457 cytokines mentions in these abstracts, which we then used to extract relations using the distant supervision approach. Considering that only 26,357 cell and 25,946 cytokines mentions exist in the immuneXpresso database, we identified about seven times more entities. Notice that these numbers refer to total mentions, i.e., any cell or cytokine may be mentioned more than once across the abstracts. We obtained a precision of 0.366 and recall of 0.853 when comparing with this silver standard. We estimate that the low precision is due to entities that do not participate in interactions, and, as such, are not considered in the silver standard used. For our objective, it is more important to recognize most of the cell and cytokines mentioned in the abstracts because the relation classifier will train and identify new relations based on those entities. Therefore, a recall of 0.853 indicates that most of the cell and cytokine names were identified.

We ran a 5-fold cross-validation on the silver standard documents to evaluate the performance of our system. We randomly divided the documents into 5 partitions and iteratively trained a classifier on the documents and respective relations of 4 partitions and tested on the documents of the other one. Then we compared the relations obtained on each iteration with

the silver standard, to calculate precision and recall. Using the classifier confidence score of each prediction, we can use it as a threshold to observe how it affects precision and recall. We compared this approach with only using the pair frequency, which was given by the number of documents where the cell and cytokine appeared within a sentence divided by the total number of documents. For both cases, we tested several threshold values and calculated precision, recall and F-measure assuming that only pairs with scores above the threshold were predicted as positive. Table 7.1 compares the confidence score calculated by the classifier with the pair frequency, at the threshold where the highest F-measure was obtained. Figure 7.4 shows the precision-recall curve obtained by ranking the pairs by classifier confidence or pair frequency. In this figure, we can see that for the same recall values, the distant supervision approach has higher precision than the frequency approach, hence it can provide higher quality results. At the highest recall values, the precision of the frequency approach is slightly higher, and for maximum recall, the precision is the same in both cases since the only difference is how the pairs are ranked. However, the classifier confidence score has a larger area under the curve (0.881 vs. 0.850). The area under the PR curve is used as an estimate of the quality of a classifier in cases where the distribution of the labels is skewed [25].

We generated a graph from the immuneXpresso database to compare with the graph generated using ICRel. This graph is composed of cell-cytokine relations found automatically in 25k abstracts from 1988 to 2015, resulting in 432 nodes and 2495 edges. The authors of this database provided other properties for each relation, such as direction and degree. However, since our system did not provide this type of information, we considered all interactions

Table 7.1: Results obtained with cross-validation on the immuneXpresso silver standard using the classifier confidence score and pair frequency at the threshold where the highest F-measure was obtained.

	Precision	Recall	F1-score	Threshold
Pair frequency	0.753	0.718	0.735	0.126
ICRel	0.911	0.696	0.789	0.918

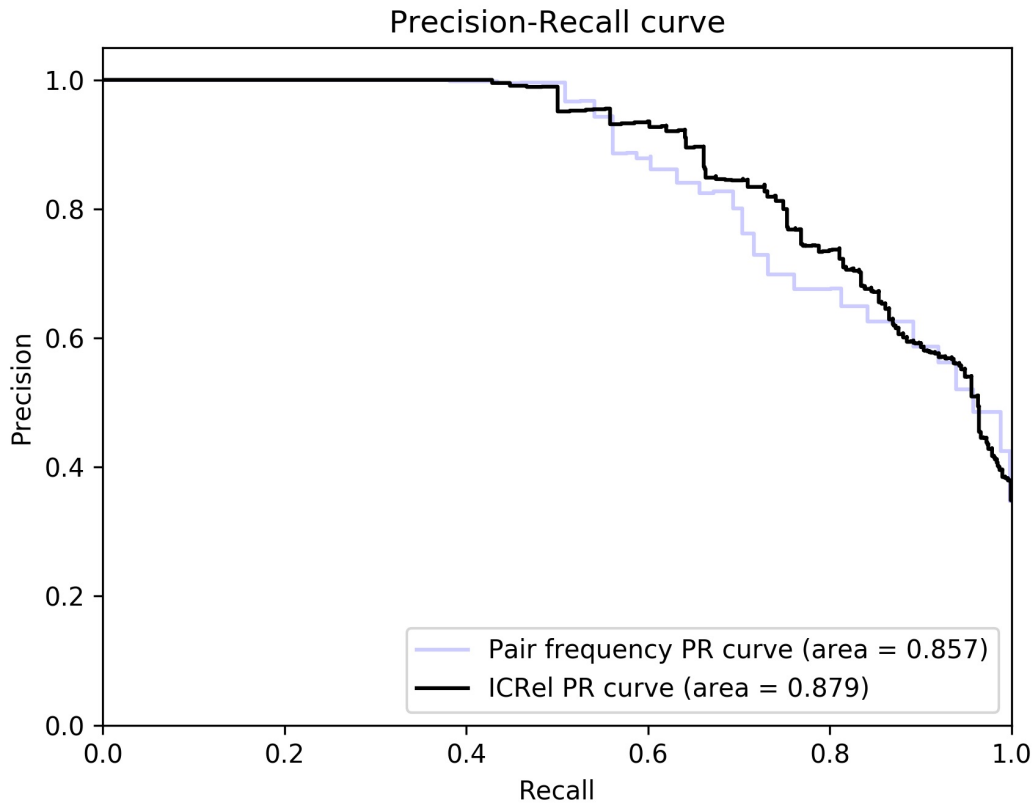


Figure 7.4: Precision-recall curves obtained using the classifier confidence score and pair frequency.

regardless of their properties.

The ICRel graph contains 212 nodes and 647 edges, extracted from 3264 abstracts. Each edge is supported by at least one sentence from these abstracts, with an average of 2.87 sentences per edge. Furthermore, each edge has a confidence value given by the classifier. We calculated the Pearson correlation between this confidence value and the number of sentences associated with the two nodes. We obtained a correlation of 0.666, which indicates that while the two variables are positively correlated, this correlation is not very strong. The diameter of this graph is 7, which is one edge larger than the immuneXpresso graph. Overall, the immuneXpresso graph contains more nodes and edges, which is expected since it was

derived from a larger number of documents than the ICRel graph. Figure 7.5 presents an overview representation of the ICRel graph, while Table 7.2 provides a comparison between the two graphs. The files used to generated the graph are provided as supplementary material. Data Sheet 1 is a table where each line is an edge of the graph and the PubMed IDs of the documents are included, while Data Sheet 2 contains the sentences which support each of the edges.

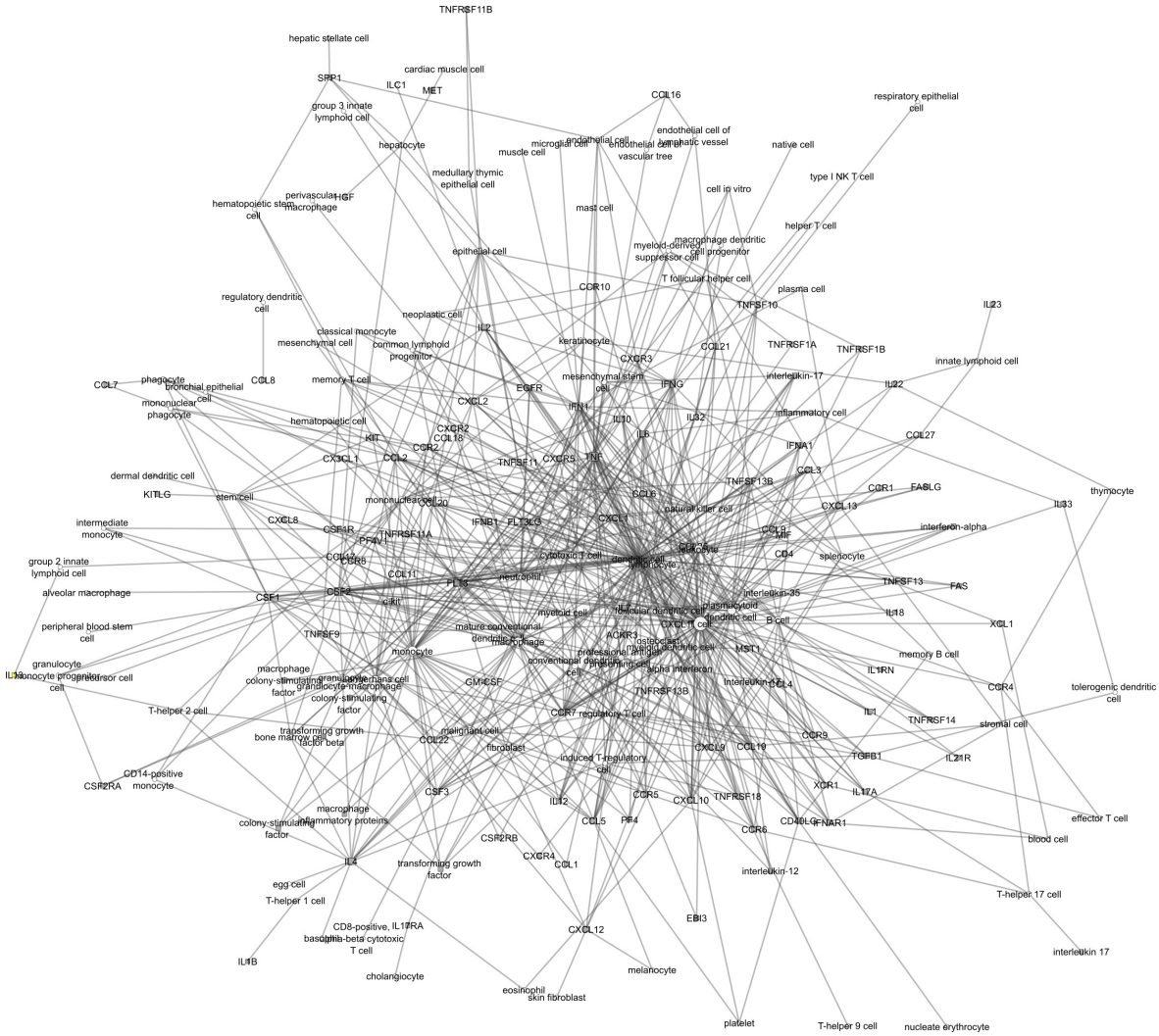


Figure 7.5: Overview of the ICRel knowledge graph. Cells are represented as white circles while cytokines are gray squares.

Table 7.2: Comparison of ICRel and immuneXpresso graphs in terms of number of nodes, edges, abstracts used, and diameter.

	ICRel	ImmuneXpresso
Nodes	212	433
Cells	93	295
Cytokines	119	138
Edges	647	2509
# abstracts	3264	25347
Diameter	7	6

Regarding the manual evaluation of the graph, the accuracy obtained was of 0.615. We obtained a kappa score of 0.600, which can be considered an adequate level of agreement [26]. In the following section, we summarize the most common sources of error found in this evaluation.

7.4 Discussion

Our work demonstrates how text mining solutions can be used to automatically generate a knowledge graph relevant to tolerogenic cell therapy. A reference database is required to train a classifier based on a specific type of relation. Due to the lack of databases about immunological therapies, we could only train and evaluate our system on immuneXpresso. As such, we were also limited in terms of type of relation to extract, since it had to be a relation described in that database. However, cytokines have been shown to be therapeutic agents in various diseases such as diabetes mellitus and multiple sclerosis. Cytokines also have important roles in the production of APCs [3]. It is relevant to understand the relation described in the literature between cells and cytokine since these could suggest novel approaches to tolerogenic cell therapy. Our graph contains these relations and can be integrated with other sources of information through the unique identifiers provided by the Cell Ontology or Entrez databases.

We compared the confidence score given by our classifier with a frequency-based ap-

proach, where the ranking score is given by the frequency of a cell-cytokine pair in the corpus. We found that the score given by the classifier is more accurate than the pair frequency. This is also supported by the low correlation between the classifier confidence and number of sentences supporting that pair (0.666). Our system learns how to classify relations using the context words as features. A cell-cytokine pair may be mentioned in multiple documents, but if the context words used are not similar to other positive pairs, it will not be classified as such. This is the main advantage of machine learning methods, along with the possibility of improving the classifier with more validated data.

Most of the processing time necessary to run our system consists of training the classifier. This part of the process takes more time and memory as more documents are considered for training since each document introduces new words and entities. In our case, the training itself took about one day. However, once the classifier is trained, a new set of documents can be processed relatively quickly.

7.4.1 Comparison between ICRel and immuneXpresso graphs

The main point of comparison of our graph is the one created by [14], which we refer to as the immuneXpresso graph. This graph is larger than ours, containing more nodes and edges. However, it is important to consider that immuneXpresso was created using a more generic set of documents, that were retrieved using the keywords “Immunology and Allergy” and “General Science”, from a span of about 50 years. We demonstrated the usefulness of our system by generating a knowledge graph focused on one particular subject and using only abstracts published in the past two years. We expect that the number of relations extracted by our system would increase with a larger set of documents. Our assumption is that a more limited and focused set of documents should result in a graph with more relevant information to the subject of study.

We first compared the information stored in each graph in general terms. As shown in the results section, despite the difference in size, both graphs have a similar diameter. The

diameter corresponds to the shortest distance between the two most distant nodes of a graph. As an example, Figure 7.6 shows a subgraph containing the union of the longest paths of each graph with at least three nodes in common. There are three edges in this subgraph that are shared between the two graphs (T cell<->IL4, IL4<->T-helper 2 cell and T-helper 2 cell<->IL13). These associations that exist in both graphs show that ICRel can extract well studied cell-cytokines relations, while in Section 7.4.2 we show examples of extracted relations from recent papers that could not be found in the immuneXpresso graph.

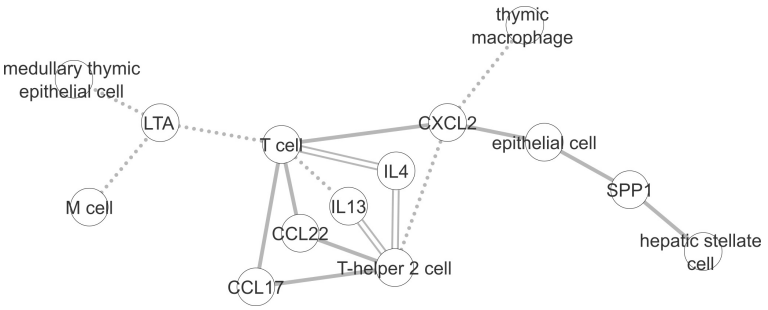


Figure 7.6: Subgraph created using the longest paths of the ICRel and immuneXpresso graphs with at least three nodes in common. Solid line corresponds to the edges of the ICRel graph, dashed line to the immuneXpresso graph and double line to both.

Comparing the relations described by each graph, we can observe various differences. The nodes in the center of the immuneXpresso graph (the center is the set of nodes whose distance to any other node is less or equal to the radius) are all cytokines (TGFB and TNG) while the ICRel graph has two cytokines (IL-6 and CSF2) and two cells (dendritic cell and T-cell) in the center. Dendritic cells are APCs, while T-cells can be targeted by APCs. Both cytokines CSF2 and IL-6 are also relevant to APCs since the former is used to differentiate APCs and the latter is produced by dendritic cells.

To better understand the degree of novelty of ICRel we divided its edges in four categories: (i) edges in common with the immuneXpresso graph; (ii) edges where the nodes existed in the immuneXpresso graph but were not connected; (iii) edges containing only one node that existed in the immuneXpresso graph; and (iv) edges where the two nodes did not exist

in the immuneXpresso graph. Table 7.3 shows the total of edges for each of these categories.

The two graphs have 132 nodes and 195 edges in common. The top five nodes that were in these edges were T cells (36), macrophages (20), TNF (19), CSF2 (17) and dendritic cells (15). Considering only nodes that were common to both graphs, ICRel found 178 new relations. For example, ICRel identified a relation between mononuclear cells and CSF2, supported by six documents.

The ICRel graph has 76 nodes (23 cells, 53 cytokines) that were not in the other graph. Of the new cytokines identified, most were actually genes coding cytokine receptors. However, we believe that these are as relevant to understand cell-cytokine relations as the cytokines themselves. A cell that produces a cytokine receptor is intrinsically associated with that cytokine. We found that 14 of the 76 new nodes were actually in the immuneXpresso database under different synonyms. For example, we identified the expressions "alpha interferon" and "interferon-alpha", but we were not able to associate with IFNA, which is how it is represented in immuneXpresso. These synonyms should be considered in future analysis to facilitate the integration of different knowledge graphs.

The ICRel graph contains 256 edges with one new node, and 18 where the two nodes were new. The top five nodes of this category were T cells (27), dendritic cells (25), FLT3 (16), CCR7 (16) and monocytes (16). While the immuneXpresso graph contained many edges with T cells and dendritic cells, ICRel identified even more cytokines related to those cells. The FLT3 receptor is associated with the differentiation of dendritic cells, which might explain why our graph contains more edges with this cytokine receptor. CCR7 is a cytokine

Table 7.3: Degree of novelty of ICRel vs. immuneXpresso.

Category of edge	#
present in both graphs	195
unique to ICRel w/ common nodes	178
unique to ICRel w/ a unique node to ICRel	256
unique to ICRel w/ both nodes unique to ICRel	18
Total	647

receptor annotated with the Gene Ontology term "positive regulation of dendritic cell antigen processing and presentation", which was recognized by our system due to an entry in the cytokine registry that we used.

7.4.2 Manual evaluation

We manually evaluated a partition of the ICRel graph to understand how a classifier trained on the immuneXpresso dataset would perform on a different corpus. This evaluation was performed by three researchers, who we refer to as curators, who read the sentences associated with 60 relations and determined if the cell-cytokine relation was supported by the text. The curators were given the same description of what was considered a relation, similar to the one presented in section 7.2.3 of this manuscript. We observed that the curators did not agree in some cases, leading to an inter-annotator agreement of 0.600, based on 15 relations. Since this value represented only a moderate agreement, we analyzed the cases where the curators disagreed. Our system considered both cytokine and cytokine receptors, and it was not clear to the curators which one was relevant. For example, one of the sentences contained the following text: "Flt3 ligand (Flt3L)"; our system recognized both FLT3LG and FLT3 and as cytokines, while FLT3 is actually a cytokine receptor. It is reasonable to assume that a cell associated with FLT3LG is also associated with its receptor, however, since it is not explicitly stated in the sentence, it caused ambiguity among the curators.

The accuracy obtained with the manual evaluation of the graph was of 0.615. The most common errors were indirect relation between the cytokine and cell, i.e., whenever there is a third element that affects both cytokine and cell. For example, consider the pair (CXCL2, memory T cell) in the sentence "(...) perivascular macrophages that are activated by IL-1a produced by keratinocytes and dDCs that are attracted by these macrophages through **CXCL2** signalling, both of which are essential for the efficient activation of **memory T cells** in situ.". Although both elements of the pair are mentioned in the sentence, there is not a direct relation described, instead, they are both directly associated to keratinocytes and

dDCs.

Another common source of error is the incorrect recognition of named entities, both cytokines and cells. For example, in every sentence mentioning "granulocyte macrophage colony-stimulating factor", macrophage was recognized as a cell entity. The cytokine registry we used to generate a list of synonyms contained some entries that were too ambiguous to be used by our system, such as acronyms that correspond to normal words. Although we were able to remove most of these synonyms, some cytokine synonyms stayed in the lexicon and generated named entity recognition errors. This is the case of immunoglobulin M (IgM), which was recognized as CD40LG since IGM is a synonym of that cytokine[4]. These errors are hard to prevent since it is not possible to have complete knowledge of which synonyms have multiple meanings. One possible solution to this problem consists in computing the semantic similarity of all entities of an abstract and using that value to exclude outliers. Assuming named entity recognition errors would have low similarity to the other entities, this method could improve the precision of our graph [27]. In the previous example, we expect that Immunoglobulin M and CD40LG would have low similarity to the other entities of that abstract.

To identify if the graph contains information relevant to APCs, we evaluated manually the edges containing the node "professional antigen-presenting cell". In the ICRel graph, this node is connected to 10 nodes: CCL19, CCL21, CCL5, CCR7, CSF2, CXCL12, IFN1, IL12, TGFB1 and TNF. Two of these cytokines (CSF2 and IL12) also appear associated with APCs in immuneXpresso. The ICRel graph contains the more generic IFN1, which includes two cytokines that appear associated with APCs in immuneXpresso (IFNA and IFNG). We confirmed the relations between APCs and its respective cytokines in the papers from where they were extracted (Table 7.4). By carefully analyzing the articles or the sentences provided in the supplementary material Data Sheet 2, it is possible to obtain more details about these relations. For example, [28] explain the roles of CCL19 and CCL21 in the migration of

[4] https://www.ncbi.nlm.nih.gov/gene/959

APCs to lymph nodes. Since our system identifies both cytokines and their receptors, it also identified a relation between CCR7, a chemokine receptor, and APCs. Even though CCR7 is associated with APCs, as explained in this article, it is out of the scope of the knowledge graph, which consists of cell-cytokines relations. [29] show that CXCL12 and CCL5 are relevant to the recruitment of APCs in early vitiligo. Although this is not directly related to tolerogenic therapies, understanding the mechanisms of APCs in disease can lead to new methods to generate and modulate the action of these cells. Further improvements could be added to ICRel in order to extract other attributes of each relation, such as directionality, temporality and magnitude. For example, by adapting the methods that we recently developed to classify the type, polarity, degree and modality of clinical events [30].

To understand if our method was able to find relations that were not yet well studied, we compared the cytokines associated with APCs and dendritic cells on ICRel and immuneXpresso (Table 7.4). ImmuneXpresso was generated using abstracts up to 2015, excluding that year. Only 2 of the 10 cytokines from ICRel were also found in immuneXpresso. Seven cytokines were found to be associated with APCs in papers from recent years. One cytokine receptor (CCR7) was also found to be associated with APCs and dendritic cells by our system. Our system as able to correctly extract this new information and organize it in a knowledge graph. We also studied the edges containing the node "dendritic cell", which is a type of professional APC. The ICRel graph contains 64 edges associated with dendritic cells, of which 49 were not found in immuneXpresso. Dendritic cells and APCs had 7 edges in common in the ICRel graph (IFN1, CCR7, IL12, CSF2, TNF, CCL5 and CCL19). Comparing to the immuneXpresso graph, we can see that most of the cytokines associated with dendritic cells were found to be associated with APCs by ICRel. Since there is no overlap in the source documents, this means that while these cytokines were first reported to be associated with dendritic cells, other APCs types have also been studied, such as epidermal Langerhans cells [28] and macrophages [34].

We found that immuneXpresso lacked information about specific tolerogenic cell types,

Table 7.4: Cytokines and receptors identified by ICRel as being associated with APCs. The second column indicates the reference of the abstract where that relation was found. The following columns indicate if that cytokine was associated with APCs or dendritic cells in ICRel and immuneXpresso.

		ICRel		immuneXpresso	
Cell type	Reference	APC	DC	APC	DC
CCL19	[28]	•	•		•
CCL21	[28]	•			•
CCR7	[28]	•	•		
CCL5	[29]	•	•		
CXCL12	[29]	•			•
CSF2	[31]	•	•	•	•
IFN1	[32]	•	•		•
IL12	[33]	•	•	•	•
TGFB1	[34]	•			•
TNF	[35]	•	•		•

given that the version of the Cell Ontology used did not contain them. Thus, we added a list of 13 tolerogenic APC types to the lexicon so that relations containing these cells could also be detected. This led to the identification of 8 relations containing tolerogenic APCs (Table 7.5). The majority of these relations included myeloid-derived suppressor cells (MDSC). The system identified relations between MDSC and TNF, TNFRSF1A, and TNFRSF1B. While TNFRSF1A and TNFRSF1B are actually cytokine receptors, the article that mentions them (source article) describes the effects of gene deletion of both the cytokine and the receptors in carcinogenesis [36]. The relation between MDSC and IL10 was extracted from a review article about the role of these cells in inflammatory diseases [37]. Another relation extracted was between tolerogenic dendritic cells and TGFB1. In this case, the source article establishes the importance of TGFB1 in immunotherapies using tolerogenic dendritic cells [38].

Table 7.5: Relations of tolerogenic APC types found by the ICRel system.

Cell	Cytokine	Reference
tolerogenic dendritic cell	TGFB1	[38]
tolerogenic dendritic cell	IL33	[39]
regulatory dendritic cell	CCL8	[40]
myeloid-derived suppressor cell	TNF	[36]
myeloid-derived suppressor cell	TNFRSF1B	[36]
myeloid-derived suppressor cell	TNFRSF1A	[36]
myeloid-derived suppressor cell	CXCL2	[41]
myeloid-derived suppressor cell	IL10	[37]

7.4.3 Conclusion and future directions

Due to its initial stage, there is a lack of openly available databases about tolerogenic cell therapy. Although commercial databases such as COPE and Cytokine Reference exist, these depend on manual curation. It is time-consuming to manually develop and then update databases with newly found information from published papers. Our ICRel system presents a solution to this issue, by using machine learning to automatically generate a knowledge graph of cell-cytokine relations. Using the knowledge graph, experts can then find more facts to store in their own databases, or help them formulate new hypotheses that need further study. Our system obtained higher precision values when compared to a frequency based approach.

We demonstrated the usefulness of the system by focusing on antigen presenting cells relevant to tolerogenic cell therapy. There have been various advancements in our understanding of immune mechanisms and pathways that are dysregulated in autoimmune diseases, and active in transplant rejection, contributing to advancements in tolerogenic therapies. A better organization of the current knowledge about this process would benefit the development of new treatments and clinical trials. The knowledge graph contained relations between APCs that were found only in recent papers, thus showing how our system can lead to a more complete information structure on this topic. Furthermore, we identified multiple associations between specific tolerogenic APCs and cytokines. We believe that our proposed system has a large potential to help practicing cell biologists or cell therapy experts in identifying

relevant relationships that can only be found by exploring various scientific articles in an integrated way. It was not our aim to find novel or specialized information but rather show the feasibility of the system, and to use examples for guiding practitioners and experts on how to take advantage of it.

The work presented in this manuscript has two major applications. The first is information retrieval systems that can use the information from our graph to integrate various sources of information. This is the case of Bio2RDF [9], which stores several biomedical databases, such as KEGG, PubMed and HGNC, in RDF format. The Bio2RDF project is an effort to link the entries of these databases using normalized URIs. Since our system matches each cytokine to the Entrez database and each cell to the Cell Ontology, it should be simple to integrate our graph with other databases for information retrieval. Another major application is recommendation systems. It is useful for a researcher working with a specific group of cell lines to know which other cells could also fit in that group. There are various methods to provide this type of recommendation, one of them consisting in exploring the structure of the graph to compute similarity measures. A recommender system could then suggest cells that interact with the same cytokines as the cells in the group. By integrating with external sources, it would be possible to suggest cytokines associated with specific diseases, chemicals or genes.

Author Contributions

Conceptualization: AL FMC. Funding acquisition: LAC FMC. Investigation: AL JF LAC FMC. Methodology: AL FMC. Project administration: LAC FMC. Software: AL. Supervision: LAC FMC. Validation: AL JF LAC FMC. Visualization: AL LAC FMC. Writing: original draft: AL LAC FMC. Writing: review and editing: AL JF LAC FMC.

Funding

This work was supported by the Portuguese Fundação para a Ciência e Tecnologia (https://www.fct.mctes.pt/) through the PhD Grant ref. PD/BD/106083/2015 to AL, UID/MULTI/04046/2013 (BioISI) to LAC, and UID/CEC/00408/2013 (LaSIGE) to AL, JF and FMC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We thank the reviewers for their valuable comments and suggestions, and in particular the feedback from reviewer 2 that helped us make the manuscript more appealing to immunologists.

References

[1] Adrian E. Morelli and Angus W. Thomson. 'Tolerogenic dendritic cells and the quest for transplant tolerance'. In: *Nature Reviews Immunology* 7.8 (2007), pp. 610–621. ISSN: 1474-1733. DOI: 10.1038/nri2132. URL: http://www.nature.com/doifinder/10.1038/nri2132.

[2] C. M U Hilkens and J. D. Isaacs. 'Tolerogenic dendritic cell therapy for rheumatoid arthritis: Where are we now?' In: *Clinical and Experimental Immunology* 172.2 (2013), pp. 148–157. ISSN: 00099104. DOI: 10.1111/cei.12038.

[3] Sergio Rutella, Silvio Danese and Giuseppe Leone. 'Tolerogenic dendritic cells: Cytokine modulation comes of age'. In: *Blood* 108.5 (2006), pp. 1435–1440. ISSN: 00064971. DOI: 10.1182/blood-2006-03-006403.

[4] Don R Swanson. 'Medical literature as a potential source of new knowledge.' In: *Bulletin of the Medical Library Association* 78.1 (1990), p. 29.

[5] Abigail Acland et al. 'Database resources of the national center for biotechnology information'. In: *Nucleic acids research* 42.Database issue (2014), p. D7.

[6] Donna Maglott et al. 'Entrez Gene: gene-centered information at NCBI'. In: *Nucleic acids research* 33.suppl_1 (2005), pp. D54–D58.

[7] J. J. Oppenheim et al. *The Online Cytokine Reference Database*. 2000.

[8] Horst Ibelgaufts. *COPE - Cytokines & Cells Online Pathfinder Encyclopedia"*. 2016. URL: http://www.cells-talk.com/.

[9] François Belleau et al. 'Bio2RDF: Towards a mashup to build bioinformatics knowledge systems'. In: *Journal of Biomedical Informatics* 41.5 (2008), pp. 706–716. ISSN: 15320464. DOI: 10.1016/j.jbi.2008.03.004.

[10] Damian Szklarczyk et al. 'The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible'. In: *Nucleic acids research* 45.D1 (2017), pp. D362–D368.

[11] Andre Lamurias, Luka A Clarke and Francisco M Couto. 'Extracting microRNA-gene relations from biomedical literature using distant supervision'. In: *PloS ONE* 12.3 (2017), e0171929.

[12] Muhammed A Yildirim et al. 'Drug—target network'. In: *Nature Biotechnology* 25.10 (2007), pp. 1119–1126. ISSN: 1087-0156. DOI: 10.1038/nbt1338. URL: http://www.nature.com/doifinder/10.1038/nbt1338.

[13] S Ebner et al. 'Production of IL-12 by human monocyte-derived dendritic cells is optimal when the stimulus is given at the onset of maturation, and is further enhanced by IL-4.' In: *Journal of immunology* 166.1 (2001), pp. 633–641. ISSN: 0022-1767. DOI: 10.4049/jimmunol.166.1.633.

[14] Shai S Shen-Orr et al. 'Towards a cytokine-cell interaction knowledgebase of the adaptive immune system.' In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2009), pp. 439–50. ISSN: 2335-6936. DOI: 10.1016/j.pestbp.2011.02.012.Investigations. URL: http://www.ncbi.nlm.nih.gov/pubmed/19209721%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2709757.

[15] Dietrich Rebholz-Schuhmann et al. 'CALBC silver standard corpus'. In: *Journal of bioinformatics and computational biology* 8.01 (2010), pp. 163–179.

[16] Ning Kang, Erik M van Mulligen and Jan A Kors. 'Training text chunkers on a silver standard corpus: can silver replace gold?' In: *BMC Bioinformatics* 13.1 (2012), p. 17. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-17.

[17] Terrence F Meehan et al. 'Logical Development of the Cell Ontology'. In: *BMC Bioinformatics* 12.1 (2011), p. 6. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-6. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-6.

[18] Francisco M Couto, Luis F Campos and Andre Lamurias. 'MER: a minimal named-entity recognition tagger and annotation server'. In: (2017).

[19] Mike Mintz et al. 'Distant supervision for relation extraction without labeled data'. In: *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP* August (2009), pp. 1003–1011. ISSN: 1932432469. DOI: 10.3115/1690219.1690287.

[20] Razvan C Bunescu and Raymond J Mooney. 'Multiple instance learning for sparse positive bags'. In: *Proceedings of the 24th International Conference on Machine Learning (2007)* 79.June (2007), pp. 105–112. DOI: 10.1145/1273496.1273510. URL: http://portal.acm.org/citation.cfm?doid=1273496.1273510.

[21] Grace Wahba. 'Multivariate function and operator estimation, based on smoothing splines and reproducing kernels'. In: *Nonlinear Modeling and Forecasting, SFI Studies in the Scienes of Complexity*. Vol. 12. Addison-Wesley. 1992, pp. 95–112.

[22] Melissa S Cline et al. 'Integration of biological networks and gene expression data using Cytoscape'. In: *Nature protocols* 2.10 (2007), p. 2366.

[23] Aric Hagberg, Pieter Swart and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Laboratory (LANL), 2008.

[24] Joseph L Fleiss. 'Measuring nominal scale agreement among many raters.' In: *Psychological bulletin* 76.5 (1971), p. 378.

[25] Jesse Davis and Mark Goadrich. 'The relationship between Precision-Recall and ROC curves'. In: *Proceedings of the 23rd international conference on Machine learning - ICML '06* (2006), pp. 233–240. ISSN: 14710080. DOI: 10.1145/1143844.1143874. arXiv: 1609.07195. URL: http://portal.acm.org/citation.cfm?doid=1143844.1143874.

[26] Mary L McHugh. 'Lessons in biostatistics Interrater reliability : the kappa statistic'. In: *Biochemica Medica* 22.3 (2012), pp. 276–282. ISSN: 1330-0962. DOI: 10.11613/BM.2012.031. arXiv: 9809069v1 [arXiv:gr-qc].

[27] Andre Lamurias, João D. Ferreira and Francisco M. Couto. 'Improving chemical entity recognition through h-index based semantic similarity'. In: *Journal of Cheminformatics* 7.Suppl 1 (2015), S13. ISSN: 17582946. DOI: 10.1186/1758-2946-7-S1-S13. URL: http://www.jcheminf.com/content/7/S1/S13.

[28] Steven A Bryce et al. 'ACKR4 on Stromal Cells Scavenges CCL19 To Enable CCR7-Dependent Trafficking of APCs from Inflamed Skin to Lymph Nodes'. In: *The Journal of Immunology* 196.8 (2016), pp. 3341–3353.

[29] Ahmed F Rezk et al. 'Misbalanced CXCL12 and CCL5 chemotactic signals in vitiligo onset and progression'. In: *Journal of Investigative Dermatology* 137.5 (2017), pp. 1126–1134.

[30] A. Lamurias et al. 'ULISBOA at SemEval-2017 Task 12: Extraction and classification of temporal expressions and events'. In: *10th International Workshop on Semantic Evaluation (SemEval)*. 2017. URL: http://nlp.arizona.edu/SemEval-2017/pdf/SemEval179.pdf.

[31] Renuka Ramanathan et al. 'Effect of mucosal cytokine administration on selective expansion of vaginal dendritic cells to support nanoparticle transport'. In: *American Journal of Reproductive Immunology* 74.4 (2015), pp. 333–344.

[32] Meagan O'Brien et al. 'CD4 receptor is a key determinant of divergent HIV-1 sensing by plasmacytoid dendritic cells'. In: *PLoS pathogens* 12.4 (2016), e1005553.

[33] Sweena M Chaudhari et al. 'Deficiency of HIF1α in antigen-presenting cells aggravates atherosclerosis and type 1 T-helper cell responses in mice'. In: *Arteriosclerosis, thrombosis, and vascular biology* (2015), ATVBAHA–115.

[34] Fayaz Ahmad Mir, Laura Contreras-Ruiz and Sharmila Masli. 'Thrombospondin-1-dependent immune regulation by transforming growth factor-β2-exposed antigen-presenting cells'. In: *Immunology* 146.4 (2015), pp. 547–556.

[35] Matthew A Burchill, Beth A Tamburini and Ross M Kedl. 'T cells compete by cleaving cell surface CD27 and blocking access to CD70-bearing APCs'. In: *European journal of immunology* 45.11 (2015), pp. 3140–3149.

[36] Andrea Sobo-Vujanovic et al. 'Inhibition of soluble tumor necrosis factor prevents chemically induced carcinogenesis in mice'. In: *Cancer immunology research* 4.5 (2016), pp. 441–451.

[37] Yewon Kwak, Hye-Eun Kim and Sung Gyoo Park. 'Insights into myeloid-derived suppressor cells in inflammatory diseases'. In: *Archivum immunologiae et therapiae experimentalis* 63.4 (2015), pp. 269–285.

[38] Amy E Anderson et al. 'Tolerogenic dendritic cells generated with dexamethasone and vitamin D3 regulate rheumatoid arthritis CD4+ T cells partly via transforming growth factor-β1'. In: *Clinical & Experimental Immunology* 187.1 (2017), pp. 113–123.

[39] Sumate Ampawong and Pornanong Aramwit. 'Tolerogenic responses of CD206+, CD83+, FOXP3+, and CTLA-4 to sericin/polyvinyl alcohol/glycerin scaffolds relevant to IL-33 and HSP60 activity'. In: *Histol Histopathol* 31 (2016), pp. 1011–27.

[40] Hirohisa Mekata et al. 'Expression of regulatory dendritic cell-related cytokines in cattle experimentally infected with Trypanosoma evansi'. In: *Journal of Veterinary Medical Science* 77.8 (2015), pp. 1017–1019.

[41] Yuan Zhuang et al. 'A pro-inflammatory role for Th22 cells in Helicobacter pylori-associated gastritis'. In: *Gut* (2014), gutjnl–2014.

8

BO-LSTM: Classifying relations via long short-term memory networks along biomedical ontologies

ANDRE LAMURIAS, DIANA SOUSA, LUKA A CLARKE AND FRANCISCO M COUTO

Abstract

Recent studies have proposed deep learning techniques, namely recurrent neural networks, to improve biomedical text mining tasks. However, these techniques rarely take advantage of existing domain-specific resources, such as ontologies. In Life and Health Sciences there is a vast and valuable set of such resources publicly available, which are continuously being updated. Biomedical ontologies are nowadays a mainstream approach to formalize existing knowledge about entities, such as genes, chemicals, phenotypes, and disorders. These resources contain supplementary information that may not be yet encoded in

training data, particularly in domains with limited labeled data.

We propose a new model to detect and classify relations in text, BO-LSTM, that takes advantage of domain-specific ontologies, by representing each entity as the sequence of its ancestors in the ontology. We implemented BO-LSTM as a recurrent neural network with long short-term memory units and using open biomedical ontologies, specifically Chemical Entities of Biological Interest (ChEBI), Human Phenotype, and Gene Ontology. We assessed the performance of BO-LSTM with drug-drug interactions mentioned in a publicly available corpus from an international challenge, composed of 792 drug descriptions and 233 scientific abstracts. By using the domain-specific ontology in addition to word embeddings and WordNet, BO-LSTM improved the F1-score of both the detection and classification of drug-drug interactions, particularly in a document set with a limited number of annotations. We adapted an existing DDI extraction model with our ontology-based method, obtaining a higher F1 score than the original model. Furthermore, we developed and made available a corpus of 228 abstracts annotated with relations between genes and phenotypes, and demonstrated how BO-LSTM can be applied to other types of relations.

Our findings demonstrate that besides the high performance of current deep learning techniques, domain-specific ontologies can still be useful to mitigate the lack of labeled data.

8.1 Background

Current relation extraction methods employ machine learning algorithms, often using kernel functions in conjunction with Support Vector Machines [1, 2] or based on features extracted from the text [3]. In recent years, deep learning techniques have obtained promising results in various Natural Language Processing (NLP) tasks [4], including relation extraction [5]. These techniques have the advantage of being easily adaptable to multiple domains, using models pre-trained on unlabeled documents [6]. The success of deep learning for text

mining is in part due to the high quantity of raw data available and the development of word vector models such as word2vec [7] and GloVe [8]. These models can use unlabeled data to predict the most probable word according to the context words (or vice-versa), leading to meaningful vector representations of the words in a corpus, known as word embeddings.

A high volume of biomedical information relevant to the detection of Adverse Drug Reactions (ADRs), such as Drug-Drug Interactions (DDI), is mainly available in articles and patents [9] A recent review of studies about the causes of hospitalization in adult patients has found that ADRs were the most common cause, accounting for 7% of hospitalizations [10]. Another systematic review focused on the European population, identified that 3.5% of hospital admissions were due to ADRs, while 10.1% of the patients experienced ADRs during hospitalization [11].

The knowledge encoded in the ChEBI (Chemical Entities of Biological Interest) ontology is highly valuable for detection and classification of DDIs, since it provides not only the important characteristics of each individual compound but also, more importantly, the underlying semantics of the relations between compounds. For instance, dopamine (CHEBI:18243), a chemical compound with several important roles in the brain and body, can be characterized as being a catecholamine (CHEBI:33567), an aralkylamino compound (CHEBI:64365) and an organic aromatic compound (CHEBI:33659) (Fig. 8.1). When predicting if a certain drug interacts with dopamine, its ancestors will provide additional information that is not usually directly expressed in the text. While the reader can consult additional materials to better understand a biomedical document, current relation extraction models are trained solely on features extracted from the training corpus. Thus, ontologies confer an advantage to relation extraction models due to the semantics encoded in them regarding a particular domain. Since ontologies are described in a common machine-readable format, methods based on ontologies can be applied to different domains and incorporated with other sources of knowledge, bridging the semantic gap between relation extraction models, data sources, and results [12].

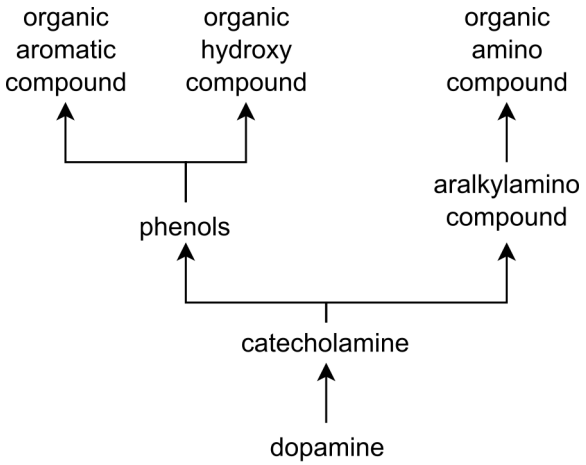


Figure 8.1: An excerpt of the ChEBI ontology showing the first ancestors of dopamine, using "is-a" relationships.

8.1.1 Deep learning for biomedical NLP

Current state-of-the-art text mining methods employ deep learning techniques, such as Recurrent Neural Networks (RNN), to train classification models based on word embeddings and other features. These methods use architectures composed of multiple layers, where each layer attempts to learn a different kind of representation of the input data. This way, different types of tasks can be trained using the same input data. Furthermore, there is no need to manually craft features for a specific task.

Long Short-Term Memory (LSTM) networks have been proposed as an alternative to regular RNN [13]. LSTMs are a type of RNN that can handle long dependencies, and thus are suitable for NLP tasks, which involve long sequences of words. When training the weights of an RNN, the contribution of the gradients may vanish while propagating for long sequences of words. LSTM units account for this vanishing gradient problem through a gated architecture, which makes it easier for the model to capture long-term dependencies. Recently, LSTMs have been applied to relation extraction tasks in various domains. Miwa and Bansal [14] presented a model that extracted entities and relations based on bidirectional tree-structured and sequential LSTM-RNNs. The authors evaluated this model on three data-

sets, including the SemEval 2010 Task 8 dataset, which defines 10 general semantic relations types between nominals [15].

Bidirectional LSTMs have been proposed for relation extraction, obtaining better results than one-directional LSTMs on the SemEval 2010 dataset [16]. In this case, at each time step, there are two LSTM layers, one that reads the sentence from left to right, and another that reads from right to left. The output of both layers is combined to produce a final score.

The model proposed by Xu et al. [17] combines Shortest Dependency Paths (SDP) between two entities in a sentence with linguistic information. SDPs are informative features for relations extraction since these contain the words of the sentence that refer directly to both entities. This model has a multichannel architecture, where each channel makes use of information from a different source along the SDP. The main channel, which contributes the most to the performance of the model, uses word embeddings trained on the English Wikipedia with word2vec. Additionally, the authors study the effect of adding channels consisting of the part-of-speech tags of each word, the grammatical relations between the words of the SDP, and the WordNet hypernyms of each word. Using all four channels, the F1-score of the SemEval 2010 Task 8 was 0.0135 higher than when using only the word embeddings channel. Although WordNet can be considered an ontology, its semantic properties were not integrated in this work, since only the word class is extracted, and the relations between classes are not considered.

Deep learning approaches to DDI classification have been proposed in recent years, using the SemEval 2013: Task 9 DDI extraction corpus to train and evaluate their performance. Zhao et al. [18] proposed a syntax convolutional neural network for DDI extraction, using word embeddings. Due to its success on other domains, LSTMs have also been used for DDI extraction [19, 20, 21, 22]. Xu et al. [21] proposed a method that combines domain-specific biomedical resources to train embedding vectors for biomedical concepts. However, their approach uses only contextual information from patient records and journal abstracts and does not take into account the relations between concepts that an ontology provides.

While these works are similar to ours, we present the first model that makes use of a domain-ontology to classify DDIs.

8.1.2 Ontologies for biomedical text mining

While machine learning classifiers trained on word embeddings can learn to detect relations between entities, these classifiers may miss the underlying semantics of the entities according to their respective domain. However, the semantics of a given domain are, in some cases, available in the form of an ontology. Ontologies aim at providing a structured representation of the semantics of the concepts in a domain and their relations [23]. In this paper, we consider a domain-specific ontology as a directed acyclic graph where each node is a concept (or entity) of the domain and the edges represent known relations between these concepts [24]. This is a common representation of existing biomedical ontologies, which are nowadays a mainstream approach to formalize knowledge about entities, such as genes, chemicals, phenotypes, and disorders.

Biomedical ontologies are usually publicly available and cover a large variety of topics related to Life and Health Sciences. In this paper, we use ChEBI, an ontology for chemical compounds with biological interest, where each node corresponds to a chemical compound [25]. The latest release of ChEBI contains nearly 54k compounds and 163k relationships. Note that, the success of exploring a given biomedical ontology for performing a specific task can be easily extended to other topics due to the common structure of biomedical ontologies. For example, the same measures of metadata quality have been successfully applied to resources annotated with different biomedical ontologies [26].

Other authors have previously combined ontological information with neural networks, to improve the learning capabilities of a model. Li et al. [27] mapped each word to a WordNet sense disambiguation to account for the different meanings that a word may have and the relations between word senses. Ma et al. [28] proposed the LSTM-OLSI model, which indexes documents based on the word-level contextual information from the DBpedia

ontology and document-level topic modeling. Some authors have explored graph embedding techniques, converting relations to a low dimensional space which represents the structure and properties of the graph [29]. For example, Kong et al. [30] combined heterogeneous sources of information, such as ontologies, to perform multi-label classification, while Dasigi et al. [31] presented an embedding model based on ontology concepts to represent word tokens.

However, few authors have explored biomedical ontologies for relation extraction. Textpresso is a project that aims at helping database curation by automatically extracting biomedical relations from research articles [32]. Their approach incorporates an internal ontology to identify which terms may participate in relations according to their semantics. Other approaches measure the similarity between the entities and use the value as a feature for a machine learning classifier [33]. One of the teams that participated in the BioCreative VI ChemProt task used ChEBI and Protein Ontology to extract additional features for a neural network model that extracted relation between chemicals and proteins [34]. To the best of our knowledge, our work is the first attempt at incorporating ancestry information from biomedical ontologies with deep learning to extract relations from text.

In this manuscript, we propose a new model, BO-LSTM that can explore domain information from ontologies to improve the task of biomedical relation extraction using deep learning techniques. We compare the effect of using ChEBI, a domain-specific ontology, and WordNet, a generic English language ontology, as external sources of information to train a classification model based on LSTM networks. This model was evaluated on a publicly available corpus of 792 drug descriptions and 233 scientific abstracts annotated with DDIs relevant to the study of adverse drug effects. Using the domain-specific ontology in addition to word embeddings and WordNet, BO-LSTM improved the F1-score of the classification of DDIs by 0.0207. Our model was particularly efficient with document types that were less represented in the training data. Moreover, we improved the F1-score of an existing DDI extraction model by 0.022 by adding our proposed ontology information, and demonstrated its

applicability to other domains by generating a corpus of gene-phenotype relations and training our model on that corpus. The code and results obtained with the model can be found on our GitHub repository (https://github.com/lasigeBioTM/BOLSTM), while a Docker image is also available (https://hub.docker.com/r/andrelamurias/bolstm), simplifying the process of training new classifiers and applying them to new data. We also made available the corpus produced for gene-phenotype relations, where each entity is mapped to an ontology concept. These results support our hypothesis that domain-specific information is useful to complement data-intensive approaches such as deep learning.

8.2 Methods

In this section, we describe the proposed BO-LSTM model in detail, as shown in Fig. 8.2, with a focus on the aspects that refer to the use of biomedical ontologies.

8.2.1 Data preparation

The objective of our work is to identify and classify relations between biomedical entities found in natural language text. We assume that the relevant entities are already recognized. Therefore, we process the input data in order to generate instances to be classified by the model. Considering the set of entities E mentioned in a sentence, we generate $\binom{E}{2}$ instances of that sentence. We refer to each instance as a candidate pair, identified by the two entities that constitute that pair, regardless of the order. A relation extraction model will assign a class to each candidate pair. In some cases, it is enough to simply classify the candidate pairs as negative or positive, while in other cases different types of positive relations are considered.

An instance should contain the information necessary to classify a candidate pair. Therefore, after tokenizing each sentence, we obtain the Shortest Dependency Path (SDP) between the entities of the pair. For example, in the sentence "Laboratory Tests Response to Plenaxis$_{e1}$

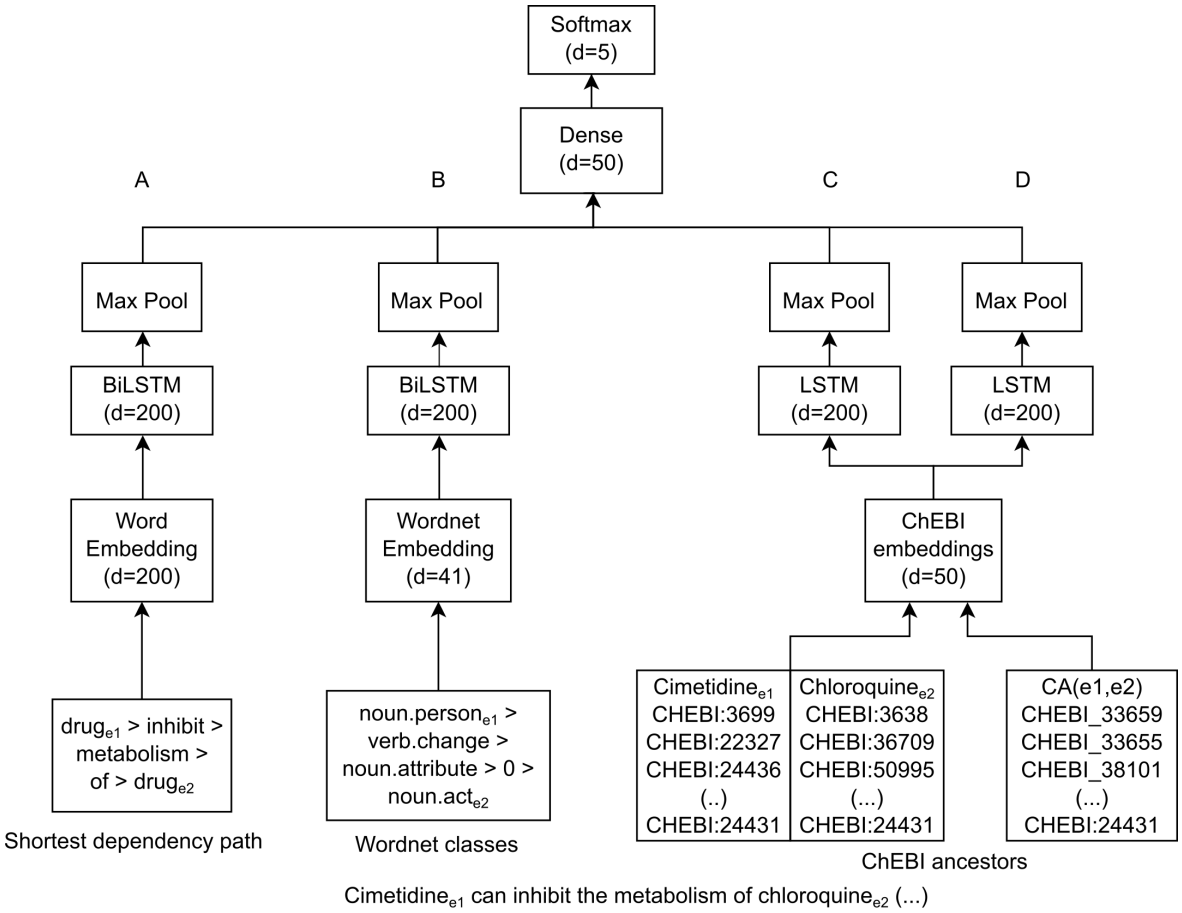


Figure 8.2: BO-LSTM Model architecture, using a sentence from the Drug-Drug Interactions corpus as an example. Each box represents a layer, with an output dimension, and merging lines represent concatenation. We refer to (A) as the Word embeddings channel, (B) the WordNet channel and (C) the ancestors concatenation channel and (D) the common ancestors channel.

should be monitored by measuring serum total testosterone$_{e1}$ concentrations just prior to administration on Day 29 and every 8 weeks thereafter", the shortest path between the entities would be Plenaxis - Response - monitored - by - measuring - concentrations - testosterone. For both tokenization and dependency parsing, we use the spaCy software library (https://spacy.io/). The text of each entity that appears in the SDP, including the candidate entities, is replaced by the generic string to reduce the effect of specific entity names on the

model. For each element of the SDP, we obtain the WordNet hypernym class using the tool developed by Ciaramita and Altun [35].

To focus our attention on the effect of the ontology information, we use pre-trained word embedding vectors. Pyysalo et al. [36] released a set of vectors trained on PubMed abstracts (nearly 23 million) and PubMed Central full documents (nearly 700k), with the word2vec algorithm [7]. Since these vectors were trained on a large biomedical corpus, it is likely that its vocabulary will contain more words relevant to the biomedical domain than the vocabulary of a generic corpus.

We match each entity to an ontology concept so that we can then obtain its ancestors. Ontology concepts contain an ID, a preferred label, and, in most cases, synonyms. While pre-processing the data, we match each entity to the ontology using fuzzy matching. The adopted implementation uses the Levenshtein distance to assign a score to each match.

Our pipeline first attempts to match the entity string to a concept label. If the match has a score equal to or higher than 0.7 (determined empirically), we accept that match and assign the concept ID to that entity. Otherwise, we match to a list of synonyms of ontology concepts. If that match has a score higher than the original score, we assign the ID of the matched synonym to the entity, otherwise, we revert to the original match. It is preferable to match to a concept label since these are more specific and should reflect the most common nomenclature of the concepts. This way, every entity was matched to a ChEBI concept, either to its preferred label or to a synonym. Due to the automatic linking method used, we cannot assume that every match is correct, but fuzzy matching has been used for similar purposes [59], so we can assume that the best match is chosen. We matched 9020 unique entities to the preferred label and 877 to synonyms, and 1283 unique entities had an exact match to either a preferred label or synonym.

The DDI corpus used to evaluate our method has a high imbalance of positive and negative relations, which hinders the training of a classification model. Even though only entities mentioned in the same sentence are considered as candidate DDIs, there is still a ratio of

1:5.9 positive to negative instances. Other authors have suggested reducing the number of negative relations through simple rules [37, 38]. We excluded from training and automatically classify as negative the pairs that fit the following rules:

- entities have the same text (regardless of case): in nearly every case a drug does not interact with itself;
- the only text between the candidate pair is punctuation: consecutive entities, in the form of lists and enumerations, are not interacting, as well as instances where the abbreviation of an entity is introduced;
- both entities have anti-positive governors: we follow the methodology proposed by [37], where the headwords of entities that do not interact are used to filter less informative instances.

With this filtering strategy, we used only 15697 of the 27792 pairs of the training corpus, obtaining a ratio of 1:3.5 positive to negative instances.

We developed a corpus of 228 abstracts annotated with human phenotype-gene relations, which we refer to as the HP corpus, to demonstrate how our model could be applied to other relation extraction tasks. This corpus was based on an existing corpus that were manually annotated with 2773 concepts of the Human Phenotype Ontology [39], corresponding to 2170 unique concepts. The developers of the Human Phenotype Ontology made available a file that links phenotypes and genes that are associated with the same diseases. Each gene of this file was automatically annotated on the HP corpus through exact string matching, resulting in 360 gene entity mentions. Then, we assumed that every gene-phenotype pair that co-occurred in the same sentence was a positive instance if this relation existed in the file. While the phenotype entities were manually mapped to the Human Phenotype Ontology, we had to employ an automatic method to obtain the most representative Gene Ontology [40, 41] concept of each gene, giving preference to concepts inferred from experiments. We applied the same pre-processing steps as for the DDI corpus, except for entity matching and negative

instance filtering. This corpus is available at https://github.com/lasigeBioTM/BOLSTM/tree/master/HP%20corpus.

8.2.2 BO-LSTM model

The main contribution of this work is the integration of ontology information with a neural network classification model. A domain-specific ontology is a formal definition of the concepts related to a specific subject. We can define an ontology as a tuple $< C, R >$, where C is the set of concepts and R the set of relations between the concepts, where each relation is a pair of concepts (c_1, c_2) with $c_1, c_2 \in E$. In our case, we consider only subsumption relations (is-a), which are transitive, i.e. if $(c_1, c_2) \in R$ and $(c_2, c_3) \in R$, then we can assume that (c_1, c_3) is a valid relation. Then, the ancestors of concept c are given by

$$Anc(c) = a : (c, a) \in T \tag{8.1}$$

where T is the transitive closure of R on the set E, i.e., the smallest relation set on E that contains R and is transitive. Using this definition, we can define the common ancestors of concepts c_1 and c_2 as

$$CA(c_1, c_2) = Anc(c_1) \cap Anc(c_2) \tag{8.2}$$

and the concatenation of the ancestors of concepts c_1 and c_2 as

$$Conc(c_1, c_2) = Anc(c_1) \oplus Anc(c_2) \tag{8.3}$$

We consider two types of representations of a candidate pair based on the ancestry of its elements: the first consisting of the concatenation of the sequence of ancestors of each entity; and second, consisting of the common ancestors between both entities. Each set of ancestors is sorted by its position in the ontology so that more general concepts are in the first positions and the final position is the concept itself. Common ancestors are also used in

some semantic similarity measures [42, 43, 44], since they normally represent the common information between two concepts. Due to the fact that in some cases there can be almost no overlap between the ancestors of two concepts, the concatenation provides an alternative representation.

We first represent each ontology concept as a one-hot vector v_c, a vector of zeros except for the position corresponding to the ID of the concept. The ontology embedding layer transforms these sparse vectors into dense vectors, known as embeddings, through an embedding matrix $M \in \mathbb{R}^{D \times C}$, where D is the dimensionality of the embedding layer and C is the number of concepts of the ontology. Then, the output of the embedding layer is given by

$$f(c) = M \cdot v_c$$

In our experiments, we set the dimensionality of the ontology embedding layer as 50, and initialized its values randomly. Then, these values were tuned during training through back-propagation.

The sequence of vectors representing the ancestors of the terms is then fed into the LSTM layer. Fig. 8.3 exemplifies how we adapted this architecture to our model, using a sequence of ontology concepts as input. After the LSTM layer, we use a max pool layer which is then fed into a dense layer with a sigmoid activation function. We experimented with bypassing this dense layer, obtaining inferior results. Finally, a softmax layer outputs the probability of each class.

Each configuration of our model was trained through mini-batch gradient descent with the Adam algorithm [45] and with cross-entropy as the loss function, with a learning rate of 0.001 We used the dropout strategy [46] to reduce overfitting on the trained embeddings and weights. We used a dropout rate of 0.5 on every layer except the penultimate and output layers. We tuned the hyperparameters common to all configurations using only the word embeddings channel on the validation set. Each model was trained until the validation loss stopped decreasing. The experiments were performed on an Intel Xeon CPU (X3470 @ 2.93

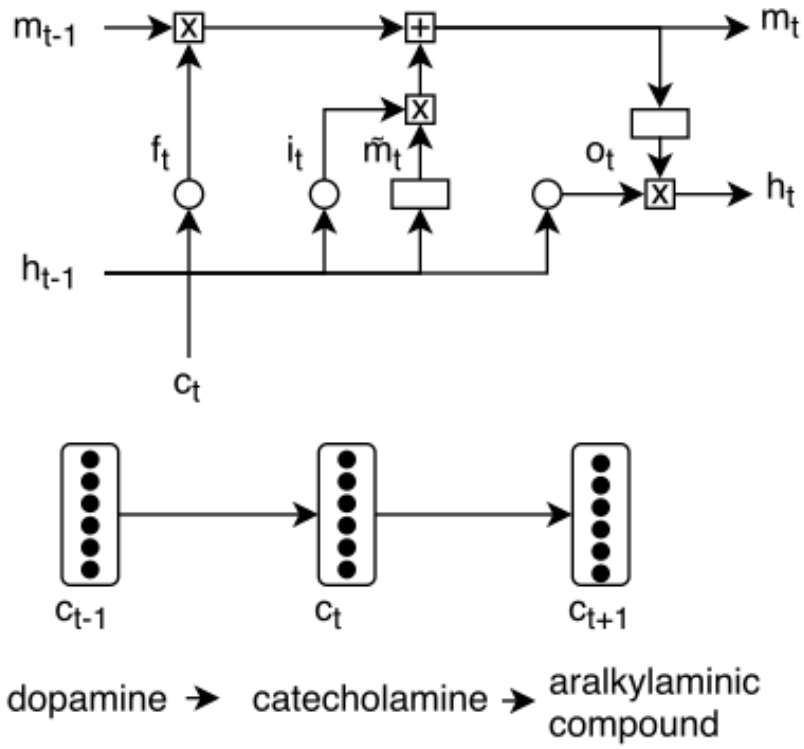


Figure 8.3: BO-LSTM unit, using a sequence of ChEBI ontology concepts as an example. Circle refers to sigmoid function and rectangle to $\tanh$, while "x" and "+" refer to element-wise multiplication and addition. h: hidden unit; $\tilde{m}$: candidate memory cell; m: memory cell; i input gate; f forget gate; o: output gate;

GHz) with 16 GB of RAM and on a GeForce GTX 1080 Ti GPU with 11GB of RAM.

The ChEBI and WordNet embedding layers were trained along with the other layers of the network. The DDI corpus contains 1757 of the 109k concepts of the ChEBI ontology. Since this is a relatively small vocabulary, we believe that this approach is robust enough to tune the weights. For the size of the WordNet embedding layer, we used 50 as suggested by Xu et al. [17], while for the ChEBI embedding layer, we tested 50, 100 and 150, obtaining the best performance with 50.

8.2.3 Baseline models

As a baseline, we implemented a model based on the SDP-LSTM model of Xu et al. [17] The SDP-LSTM model makes use of four types of information: word embeddings, part-of-speech tags, grammatical relations and WordNet hypernyms, which we refer to as channels. Each channel uses a specific type of input information to train an LSTM-based RNN layer, which is then connected to a max pooling layer, the output of the channel. The output of each channel is concatenated, and connected to a densely-connected hidden layer, with a sigmoid activation function, while a softmax layer outputs the probabilities of each class.

Xu et al. show that it is possible to obtain high performance on a relation extraction task using only the word representations channel. For this reason, we use a version of our model with only this channel as the baseline. We employ the previously mentioned pre-trained word embeddings as input to the LSTM layer.

Additionally, we make use of WordNet as an external source of information. The authors of the SDP-LSTM model showed that WordNet contributed to an improvement of the F1-score on a relation extraction task. We use the tool developed by Ciaramita and Altun [35] to obtain the WordNet classes of each word according to 41 semantic categories, such as "noun.group" and "verb.change". The embeddings of this channel were set to be 50-dimensional and tuned during the training of the model.

We adopted a second baseline model to make a stronger comparison with other DDI extraction models, based on the model presented by Zhang et al. [47]. Their model uses the sentence and SDP of each instance to train a hierarchical LSTM network. This model is constituted by two levels of LSTMs which learn feature representations of the sentence and SDP based on word, part-of-speech and distance to entity. An embedding attention mechanism is used to weight the importance of each word to the two entities that constitute each pair. We kept the architecture and hyperparameters of their model, and added another type of input, based on the common ancestors and concatenation of each entity's ancestors. We applied the same attention mechanism, so that the most relevant ancestors have a larger

weight on the LSTM. We ran the original Zhang et al. model to replicate the results, and then ran again with ontology information.

8.3 Results

We evaluated the performance of our BO-LSTM model on the SemEval 2013: Task 9 DDI extraction corpus [48]. This gold standard corpus consists of 792 texts from DrugBank [49], describing chemical compounds, and 233 abstracts from the Medline database [50]. DrugBank is a cheminformatics database containing detailed drug and drug target information, while Medline is a database of bibliographic information of scientific articles in Life and Health Sciences. Each document was annotated with pharmacological substances and sentence-level DDIs. We refer to each combination of entities mentioned in the same sentence as a candidate pair, which could either be positive if the text describes a DDI, or negative otherwise. In other words, a negative candidate is a candidate pair that is not described as interacting in the text. Each positive DDI was assigned one of four possible classes: mechanism, effect, advice, and int, when none of the others were applicable.

In the context of the competition, the corpus was separated into training and testing sets, containing both DrugBank and Medline documents. We maintained the test set partition and evaluated on it, as it is the standard procedure on this gold standard. After shuffling we used 80% of the training set to train the model and 20% as a validation set. This way, the validation set contained both DrugBank and Medline documents, and overfitting to a specific document type is avoided. It has been shown that the DDIs of the Medline documents are more difficult to detect and classify, with the best systems having almost a 30 point F1-score difference to the DrugBank documents [51].

We implemented the BO-LSTM model in Keras, a Python-based deep learning library, using the TensorFlow backend. The overall architecture of the BO-LSTM model is presented in Fig. 8.2. More details about each layer can be found in the Methods section. We focused

on the effect of using different sources of information to train the model. As such, we tuned the hyperparameters to obtain reasonable results, using as reference the values provided by other authors that have applied LSTMs to this gold standard [18, 19]. We first trained the model using only the word embeddings of the SDP of each candidate pair (Fig. 8.2A). Then we tested the effect of adding the WordNet classes as a separate embedding and LSTM layer (Fig. 8.2B) Finally, we tested two variations of the ChEBI channel: first using the concatenation of the sequence of ancestors of each entity (Fig. 8.2C), and second using the sequence of common ancestors of both entities (Fig. 8.2D).

Table 8.1 shows the DDI detection results obtained with each configuration using the evaluation tool provided by the SemEval 2013: Task 9 organizers on the gold standard, while Table 8.2 shows the DDI classification results, using the same evaluation tool and gold standard. The difference between these two tasks is that while detection ignores the type of interactions, the classification task requires identifying the positive pairs and also their correct interaction type. We compare the performance on the whole gold standard, and on each document type (DrugBank and Medline). The first row of each table shows the results obtained using an LSTM network trained solely on the word embeddings of the SDP of each candidate pair. Then, we studied the impact of adding each information channel on the performance of the model, and the effect of using all information channels, as shown in Fig. 8.2.

For the detection task, using the concatenation of ancestors results in an improvement of the F1-score in the Medline dataset, contributing to an overall improvement of the F1-score in the full test set. The most notable improvement was in the recall of the Medline dataset, where the concatenation of ancestors increased this score by 0.246. The usage of ontology ancestors did not improve the F1-score of detection of DDIs in the DrugBank dataset. In every test set, it is possible to observe that the concatenation of ancestors results in a higher recall while considering only the common ancestors is more beneficial to precision. Combining both approaches with the WordNet channel results in a higher F1-score.

Table 8.1: Evaluation scores obtained for the DDI detection task on the DDI corpus and on each type of document, comparing different configurations of the model.

	DDI test			DrugBank			Medline		
Configuration	P	R	F	P	R	F	P	R	F
Word embeddings	0.7551	0.6865	0.7192	0.7620	0.7158	0.7382	0.6389	0.377	0.4742
+ WordNet	0.716	0.6936	0.7046	0.7267	0.7143	0.7204	0.5800	0.4754	0.5225
+ Common Ancestors	**0.7661**	0.6738	0.7170	**0.7723**	0.7003	0.7345	**0.6667**	0.3607	0.4681
+ Concat. Ancestors	0.7078	0.7489	0.7278	0.7166	0.7578	0.7366	0.6032	**0.623**	**0.6129**
+ WordNet + Ancestors	0.6572	**0.8184**	**0.7290**	0.6601	**0.8385**	**0.7387**	0.5574	0.5574	0.5574

Evaluation metrics used: Precision (P), Recall (R) and F1-score (F). Each row represents the addition of an information source to the initial configuration.

Table 8.2: Evaluation scores obtained for the DDI classification task on the DDI corpus and on each type of document, comparing different configurations of the model.

	DDI test			DrugBank			Medline		
Configuration	P	R	F	P	R	F	P	R	F
Word embeddings	0.5819	0.5291	0.5542	0.5868	0.5512	0.5685	0.5000	0.2951	0.3711
+ WordNet	0.5754	0.5574	0.5663	0.5845	0.5745	**0.5795**	0.4600	0.3770	0.4144
+ Common Anc.	**0.5968**	0.5248	0.5585	**0.6045**	0.5481	0.5749	**0.5152**	0.2787	0.3617
+ Concat. Anc.	0.5282	0.5589	0.5431	0.5286	0.5590	0.5434	0.4921	**0.5082**	**0.5000**
+ WordNet + Anc.	0.5182	**0.6454**	**0.5749**	0.5171	**0.6568**	0.5787	0.4590	0.4590	0.4590

Evaluation metrics used: Precision (P), Recall (R) and F1-score (F). Each row represents the addition of an information source to the initial configuration.

Regarding the classification task (Table 8.2), the F1-score was improved on each dataset by the usage of the ontology channel. Considering only the common ancestors led to an improvement of the F1-score in the DrugBank dataset and on the full corpus, while the concatenation improved the Medline F1-score, similarly to the detection results.

To better understand the contribution of each channel, we studied the relations detected by each configuration by one or more channels, and which of those were also present in the gold standard. Fig. 8.4 and Fig. 8.5 show the intersection of the results of each channel in the full, DrugBank, and Medline test sets. We compare only the results of the detection task, as it is simpler to analyze and show the differences in the results of different configurations.

In Fig. 8.4, we can visualize false negatives as the number of relations unique to the gold standard and the false positives of each configuration as the number of relations that does not intersect with the gold standard. The difference between the values of this figure and the sum of their respective values in Fig. 8.5 is due to the system being executed once for each dataset. Overall 369 relations in the full test set were not detected by any configuration of our system, out of a total of 979 relations in the gold standard. We can observe that 60 relations were detected only when adding the ontology channels.

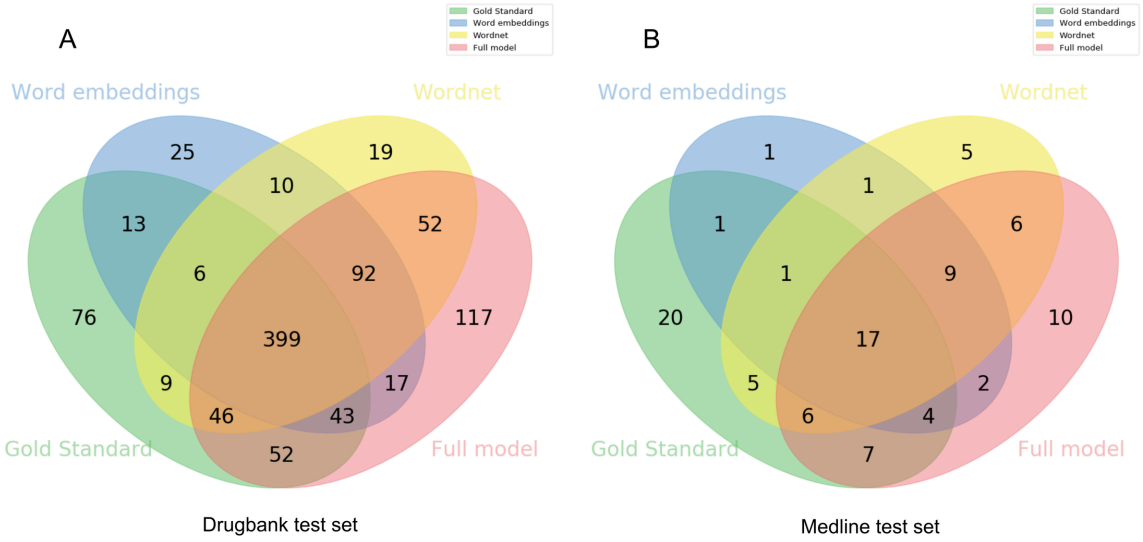


Figure 8.4: Venn diagram demonstrating the contribution of each configuration of the model to the results of the full test set. The intersection of each channel with the gold standard represents the number of true positives of that channel, while the remaining correspond to false negatives and false positives.

In the Medline test set, the ontology channel identified 7 relations that were not identified by any other configuration (Fig. 8.5B). One of these relations was the effect of quinpirole treatment on amphetamine sensitization. Quinpirole has 27 ancestors in the ChEBI ontology, while amphetamine has 17, and they share 10 of these ancestors, with the most informative being “organonitrogen compound”. While this information is not described in the original

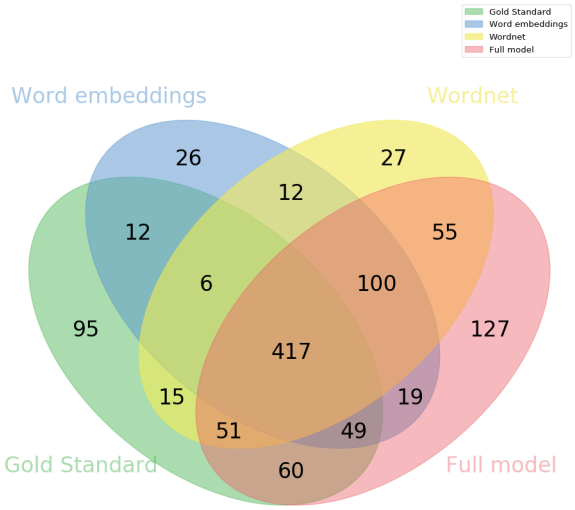


Figure 8.5: Venn diagram demonstrating the contribution of each configuration of the model to the DrugBank (A) and Medline (B) test set results. The intersection of each channel with the gold standard represents the number of true positives of that channel, while the remaining correspond to false negatives and false positives.

text, but only encoded in the ontology, it is relevant to understand if the two entities can participate in a relation. However, this comes at the cost of precision, since 10 incorrect DDIs were classified by this configuration.

To empirically compare our results with the state-of-the-art of the DDI extraction, we compiled the most relevant works on this task in Table 8.3. The first line refers to the system that obtained the best results on the original SemEval task [37, 52]. Since then, other authors have presented approaches for this task, most recently using deep learning algorithms. In Table 8.3 we compare the machine learning architecture used by each system, and the results reported by the authors. Since some authors focused only on the DDI classification task, we could not obtain the DDI detection results for those systems, hence the missing values. We were only able to replicate the results of Zhang et al.[47]. Since this system followed an architecture similar to ours, we adapted the model with our ontology-based channel, as described in the Methods section. This modification to the model resulted in an improvement

of 0.022 to the F1-score. Our version of this model is also available on our page along with the BO-LSTM model.

Table 8.3: Comparison of DDI extraction systems. The architectures mentioned are Support Vector Machines (SVM), Convolutional Neural Networks (CNN) and LSTMs.

System	Architecture	Best F1-score
FBK-irst [37]	SVM	0.651
SCNN [18]	CNN	0.686
Joint AB-LSTM [19]	LSTM	0.6939
Att-BLSTM [22]	LSTM	0.773
DLSTM [20]	LSTM	0.6839
BR-LSTM [21]	LSTM	0.7115
Zhang et al. 2018 [47]	LSTM	0.729
Zhang et al. 2018 + BO-LSTM	LSTM	0.751

We used the HP corpus to demonstrate the generalizability of our method. This case-study served only as a proof-of-concept, it was not our intent to measure the performance of the model, given the limited number of annotations and the dependence on the quality of using exact string matching to identify the genes. For example, we may have missed correct relations in the corpus, because they were not in the reference file or the gene name was not correctly identified.

Therefore, we used 60% (137 documents) of the corpus to train the model and 40% (91 documents) to manually evaluate the relations predicted with that model. For example, in the following sentence:

```
Multiple angiofibromas, collagenomas, lipomas, confetti-like
    hypopigmented macules and multiple gingival papules are
    cutaneous manifestations of MEN1 and should be looked for
    in both family members of patients with MEN1 and
    individuals with hyperparathyroidism of other MEN1-
    associated tumors.
```

, the model identified the relation between the phenotype “angiofibromas” and the gene “MEN1”. One recurrently identified relation by our model that was not present on the phenotype-gene associations file is between the phenotype ’neurofibromatosis’ and the gene ’NF2’:

Clinical and genetic data of 10 patients with neurofibromatosis 2 (NF−2) are presented.

Despite this relation not being described in the previous sentence , it is predicted given its presence in the phenotype-gene associations files. With a larger number of annotations in the training corpus, we expect this error to disappear.

8.4 Discussion

Comparing the results across the two types of documents, we can observe that our model was most beneficial to the Medline test set. This set contains only 1301 sentences from 142 documents for training, while the DrugBank set contains 5675 sentences from 572 documents. Naturally, the patterns of the DrugBank documents will be easier to learn than the ones of the Medline documents because more examples are shown to the model. Furthermore, the Medline set has 0.18 relations per sentence, while the DrugBank set has 0.67 relations per sentence. This means that DDIs are described much more sparsely than in the DrugBank set. This demonstrates that our model is able to obtain useful knowledge that is not described in the text.

One disadvantage of incorporating domain information in a machine learning approach is that it reduces its applicability to other domains. However, biomedical ontologies have become ubiquitous in biomedical research. One of the most successful cases of a biomedical ontology is the Gene Ontology, maintained by the Gene Ontology Consortium [40]. The Gene Ontology defines over 40,000 concepts used to describe the properties of genes. This project is constantly updated, with new concepts and relations being added every day. How-

ever, there are ontologies for more specific subjects, such as microRNAs [53], radiology terms [54] and rare diseases [55]. BioPortal is a repository of biomedical ontology, currently hosting 685 ontologies. Furthermore, while manually labeled corpora are created specifically to train and evaluate text mining applications, ontologies have diverse applications, i.e., they are not developed for this specific purpose.

We evaluate the proposed model on the DDI corpus because it is associated with a SemEval task, and for this reason, it has been the subject of many studies since its release. However, while applying our model to a single domain, we designed its architecture so it can fit any other domain-specific ontology. To demonstrate this, we developed a corpus of gene-phenotype relations annotated with Human Phenotype and Gene ontology concepts, and applied our model to it. Therefore, the methodology proposed can be easily followed to apply to any other biomedical ontology that describes the concepts of a particular domain. For example, the Disease Ontology [56], that describes relations between human diseases, could be used with the BO-LSTM model on a disease relation extraction task, as long as there is an annotated training corpus.

While we studied the potential of domain-specific ontologies based only on the ancestors of each entity, there are other ways to integrate semantic information from ontologies into neural networks. For example, one could consider only the ancestors with the highest information content, since those would be the most helpful to characterize an entity. The information content can be estimated either by the probability of a given term in the ontology or in an external dataset. Alternatively, a semantic similarity measure that accounts for non-transitive relations could be used to find similar concepts to the entities of the relation [57], or one that considers only the most relevant ancestors [58]. The quality of the ontology embeddings could also be improved by pre-training on a larger dataset, which would include a wider variety of concepts.

8.5 Conclusions

This work demonstrates how domain-specific ontologies can improve deep learning models for classification of biomedical relations. We developed a model, BO-LSTM which combines biomedical ontologies with LSTM units to detect and classify relations in text. In this manuscript, we demonstrate that ontologies can improve the performance of deep learning techniques for biomedical relation extraction, in particular for situations with a limited number of annotations available, which was the case of the Medline dataset. Furthermore, we explored how it can be adapted to other relation extraction domains, for example, gene-phenotype relations. Considering that biomedical ontologies are openly available and regularly updated as the knowledge on the domain progresses, they should be considered important information sources for relation extraction.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The data and code used for this study are available at https://github.com/lasigeBioTM/BOLSTM.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by FCT through funding of the DeST: Deep Semantic Tagger project, ref. PTDC/CCI-BIO/28685/2017, LaSIGE Research Unit, ref. UID/CEC/00408/2013 and BioISI, ref. ID/MULTI/04046/2013. AL is recipient of a fellowship from BioSys PhD programme (ref PD/BD/106083/2015) from FCT (Portugal).

Author's contributions

All authors read and approved the final manuscript.

Acknowledgements

We acknowledge the help of Nuno Dionisio in setting up the machine to run the experiments.

References

[1] Dmitry Zelenko et al. 'Kernel Methods for Relation Extraction'. In: *Journal of Machine Learning Research* 3 (2003), pp. 1083–1106. ISSN: 1532-4435. DOI: 10.3115/1118693.1118703.

[2] Frank Reichartz, Hannes Korte and Gerhard Paass. 'Semantic relation extraction with kernels over typed dependency trees'. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10* (2010), p. 773. DOI: 10.1145/1835804.1835902. URL: http://dl.acm.org/citation.cfm?doid=1835804.1835902.

[3] Nanda Kambhatla. 'Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations'. In: *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions* (2004), p. 22. DOI: 10.3115/1219044.1219066. URL: http://portal.acm.org/citation.cfm?doid=1219044.1219066.

[4] Ronan Collobert et al. 'Natural language processing (almost) from scratch'. In: *Journal of Machine Learning Research* 12.Aug (2011), pp. 2493–2537.

[5] A. Lamurias and F. Couto. 'Text Mining for Bioinformatics using Biomedical Literature'. In: *Reference Module in Life Sciences (Encyclopedia of Bioinformatics and Computational Biology)*. Vol. 1. Oxford: Elsevier, 2019. DOI: https://doi.org/10.1016/B978-0-12-809633-8.20409-3.

[6] Dumitru Erhan et al. 'Why Does Unsupervised Pre-training Help Deep Learning?' In: *J. Mach. Learn. Res.* 11 (Mar. 2010), pp. 625–660. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1756006.1756025.

[7] Tomas Mikolov et al. 'Distributed Representations of Words and Phrases and Their Compositionality'. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111–3119. URL: http://dl.acm.org/citation.cfm?id=2999792.2999959.

[8] Jeffrey Pennington, Richard Socher and Christopher D. Manning. 'GloVe: Global Vectors for Word Representation'. In: *Empirical Methods in Natural Language Pro-*

cessing (EMNLP). 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[9] Chung Chi Huang and Zhiyong Lu. 'Community challenges in biomedical text mining over 10 years: Success, failure and the future'. In: *Briefings in Bioinformatics* 17.1 (2016), pp. 132–144. ISSN: 14774054. DOI: 10.1093/bib/bbv024.

[10] Abdullah Al Hamid et al. 'A systematic review of hospitalization resulting from medicine-related problems in adult patients'. In: *British Journal of Clinical Pharmacology* 78.2 (2014), pp. 202–217. ISSN: 13652125. DOI: 10.1111/bcp.12293.

[11] Jacoline C Bouvy, Marie L De Bruin and Marc A Koopmanschap. 'Epidemiology of adverse drug reactions in Europe: a review of recent observational studies'. In: *Drug safety* 38.5 (2015), pp. 437–453.

[12] D. Dou, H. Wang and H. Liu. 'Semantic data mining: A survey of ontology-based approaches'. In: *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*. Feb. 2015, pp. 244–251. DOI: 10.1109/ICOSC.2015.7050814.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 'Long short-term memory'. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[14] Makoto Miwa and Mohit Bansal. 'End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures'. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1105–1116. DOI: 10.18653/v1/P16-1105. URL: http://aclweb.org/anthology/P16-1105.

[15] Iris Hendrickx et al. 'SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals'. In: *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. Association for Computational Linguistics. 2009, pp. 94–99.

[16] Shu Zhang et al. 'Bidirectional long short-term memory networks for relation classification'. In: *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*. 2015, pp. 73–78.

[17] Yan Xu et al. 'Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths'. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. September. 2015, pp. 1785–1794. ISBN: 9781941643327. DOI: 10.18653/v1/D15-1206.

[18] Zhehuan Zhao et al. 'Drug drug interaction extraction from biomedical literature using syntax convolutional neural network'. In: *Bioinformatics* 32.November (2016), btw486. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw486. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw486.

[19] Sunil Kumar Sahu and Ashish Anand. 'Drug-Drug Interaction Extraction from Biomedical Text Using Long Short Term Memory Network'. In: *CEUR Workshop Proceedings* 1828 (Jan. 2017), pp. 53–59. ISSN: 16130073. DOI: 10.1145/2910896.2910898. arXiv: 1701.08303. URL: http://arxiv.org/abs/1603.07016%20http://dx.doi.org/10.1145/2910896.2910898%20http://arxiv.org/abs/1701.08302%20http://arxiv.org/abs/1701.08303.

[20] Wei Wang et al. 'Dependency-based long short term memory network for drug-drug interaction extraction'. In: *BMC Bioinformatics* 18.Suppl 16 (2017). ISSN: 14712105. DOI: 10.1186/s12859-017-1962-8.

[21] B. Xu et al. 'Leveraging Biomedical Resources in Bi-LSTM for Drug-Drug Interaction Extraction'. In: *IEEE Access* 6 (2018), pp. 33432–33439. DOI: 10.1109/ACCESS.2018.2845840.

[22] Wei Zheng et al. 'An attention-based effective neural model for drug-drug interactions extraction'. In: (2017), pp. 1–11. DOI: 10.1186/s12859-017-1855-x.

[23] F. Couto and A. Lamurias. 'Semantic similarity definition'. In: *Reference Module in Life Sciences (Encyclopedia of Bioinformatics and Computational Biology)*. Vol. 1. Oxford: Elsevier, 2019. DOI: https://doi.org/10.1016/B978-0-12-809633-8.20401-9.

[24] B. Smith et al. 'The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration'. In: *Nature biotechnology* 25.11 (2007), pp. 1251–1255.

[25] Janna Hastings et al. 'The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013'. In: *Nucleic Acids Research* 41.D1 (2013), pp. 456–463. ISSN: 03051048. DOI: 10.1093/nar/gks1146.

[26] João D Ferreira et al. 'Assessing Public Metabolomics Metadata, Towards Improving Quality'. In: *Journal of integrative bioinformatics* 14.4 (2017).

[27] Tong Shu Li et al. 'A crowdsourcing workflow for extracting chemical-induced disease relations from free text'. In: *Database* 2016 (2016), baw051. DOI: 10.1093/database/baw051. eprint: /oup/backfile/content_public/journal/database/2016/10.1093_database_baw051/2/baw051.pdf. URL: http://dx.doi.org/10.1093/database/baw051.

[28] Ningning Ma, Hai-tao Zheng B and Xi Xiao. 'An Ontology-Based Latent Semantic Indexing Approach Using Long Short-Term Memory Networks'. In: *Web and Big Data* 10366.2 (2017), pp. 185–199. DOI: 10.1007/978-3-319-63579-8. URL: http://link.springer.com/10.1007/978-3-319-63579-8.

[29] Palash Goyal and Emilio Ferrara. 'Graph embedding techniques, applications, and performance: A survey'. In: *arXiv preprint arXiv:1705.02801* (2017).

[30] Xiangnan Kong, Bokai Cao and Philip S. Yu. 'Multi-label Classification by Mining Label and Instance Correlations from Heterogeneous Information Networks'. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. Chicago, Illinois, USA: ACM, 2013, pp. 614–622. ISBN: 978-1-4503-2174-7. DOI: 10.1145/2487575.2487577. URL: http://doi.acm.org/10.1145/2487575.2487577.

[31] Pradeep Dasigi et al. 'Ontology-Aware Token Embeddings for Prepositional Phrase Attachment'. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 2089–2098. DOI: 10.18653/v1/P17-1191. URL: http://www.aclweb.org/anthology/P17-1191.

[32] Hans-Michael Michael Müller, Eimear E. Kenny and Paul W. Sternberg. 'Textpresso: an ontology-based information retrieval and extraction system for biological literature.' In: *PLoS Biology* 2.11 (2004), e309. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0020309.

[33] Andre Lamurias, João D Ferreira and Francisco M Couto. 'Identifying interactions between chemical entities in biomedical text'. In: *J Integr Bioinform* 11.3 (2014), p. 247.

[34] Ignacio Tripodi et al. 'Knowledge-base-enriched relation extraction'. In: *Proceedings of the Sixth BioCreative Challenge Evaluation Workshop. Bethesda, MD USA*. Vol. 1. 2017, pp. 163–166.

[35] Massimiliano Ciaramita and Yasemin Altun. 'Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger'. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2006, pp. 594–602.

[36] Sampo Pyysalo et al. 'Distributional Semantics Resources for Biomedical Text Processing'. In: *Proceedings of LBM 2013* (2013). URL: http://bio.nlplab.org/pdf/pyysalo13literature.pdf.

[37] Md Faisal Mahbub Chowdhury and Alberto Lavelli. 'FBK-irst: A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information'. In: *Atlanta, Georgia, USA* 351 (2013), p. 53.

[38] Sun Kim et al. 'Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach'. In: *Journal of biomedical informatics* 55 (2015), pp. 23–30.

[39] Sebastian Köhler et al. 'The Human Phenotype Ontology in 2017'. In: *Nucleic Acids Research* 45.D1 (2017), pp. D865–D876. DOI: 10.1093/nar/gkw1039. URL: http://dx.doi.org/10.1093/nar/gkw1039.

[40] M. Ashburner et al. 'Gene Ontology: tool for the unification of biology'. In: *Nature genetics* 25.1 (2000), pp. 25–29.

[41] No authors listed. 'Expansion of the Gene Ontology knowledgebase and resources'. In: *Nucleic Acids Res.* 45.D1 (Jan. 2017), pp. D331–D338.

[42] P Resnik. 'Using information content to evaluate semantic similarity in a taxonomy'. In: *International Joint Conference on Artificial Intelligence*. Vol. 14. Citeseer. 1995, pp. 448–453.

[43] Jay J. Jiang and David W. Conrath. 'Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy'. In: *CoRR* cmp-lg/9709008 (1997). URL: http://arxiv.org/abs/cmp-lg/9709008.

[44] Dekang Lin. 'An Information-Theoretic Definition of Similarity'. In: *Proceedings of the Fifteenth International Conference on Machine Learning*. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 296–304. ISBN: 1-

55860-556-8. URL: http://dl.acm.org/citation.cfm?id=645527.657297.

[45] Diederik P. Kingma and Jimmy Ba. 'Adam: A Method for Stochastic Optimization'. In: *CoRR* abs/1412.6980 (2014). arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980.

[46] Geoffrey E. Hinton et al. 'Improving neural networks by preventing co-adaptation of feature detectors'. In: *CoRR* abs/1207.0580 (2012). arXiv: 1207.0580. URL: http://arxiv.org/abs/1207.0580.

[47] Yijia Zhang et al. 'Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths'. In: *Bioinformatics* 34.5 (2018), pp. 828–835. DOI: 10.1093/bioinformatics/btx659. URL: http://dx.doi.org/10.1093/bioinformatics/btx659.

[48] María Herrero-Zazo et al. 'The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions'. In: *Journal of biomedical informatics* 46.5 (2013), pp. 914–920.

[49] D. S. Wishart et al. 'DrugBank 5.0: a major update to the DrugBank database for 2018'. In: *Nucleic Acids Res.* 46.D1 (Jan. 2018), pp. D1074–D1082.

[50] David L Wheeler et al. 'Database resources of the national center for biotechnology information'. In: *Nucleic acids research* 35.suppl_1 (2006), pp. D5–D12.

[51] Isabel Segura-Bedmar, Paloma Martínez and María Herrero-Zazo. 'Lessons learnt from the DDIExtraction-2013 shared task'. In: *Journal of biomedical informatics* 51 (2014), pp. 152–164.

[52] Isabel Segura Bedmar, Paloma Martínez and María Herrero Zazo. 'Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)'. In: *Proceedings of the Seventh International Workshop on Semantic Evaluation*. Association for Computational Linguistics. 2013.

[53] Vicky Dritsou et al. ‘miRNAO: An Ontology Unfolding the Domain of microRNAs.’ In: *IWBBIO*. 2014, pp. 989–1000.

[54] Curtis P Langlotz. *RadLex: a new method for indexing online educational materials*. 2006.

[55] Ana Rath et al. ‘Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users’. In: *Human mutation* 33.5 (2012), pp. 803–808.

[56] Warren A Kibbe et al. ‘Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data’. In: *Nucleic acids research* 43.D1 (2014), pp. D1071–D1078.

[57] Mingdong Ou et al. ‘Non-transitive hashing with latent similarity components’. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, pp. 895–904.

[58] Andre Lamurias, João D. Ferreira and Francisco M. Couto. ‘Improving chemical entity recognition through h-index based semantic similarity’. In: *Journal of Cheminformatics* 7.Suppl 1 (2015), S13. ISSN: 17582946. DOI: 10.1186/1758-2946-7-S1-S13. URL: http://www.jcheminf.com/content/7/S1/S13.

[59] Balu Bhasuran et al. ‘Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases’. In: *Journal of Biomedical Informatics* 64 (2016), pp. 1–9. ISSN: 15320464. DOI: 10.1016/j.jbi.2016.09.009. URL: http://dx.doi.org/10.1016/j.jbi.2016.09.009.

9

General discussion and conclusions

Due to the complexity of biological systems, it is necessary to combine the knowledge originating from various studies and research areas. The main method to communicate scientific knowledge is through scientific literature, leading to a steady increase of the number of documents stored on text repositories of scientific literature. Therefore, computational methods are essential to make use of this information, as the task of finding information relevant to a particular problem becomes increasingly difficult. Text mining aims at extracting information from text, making it easier to integrate various areas of research and leading to a better understanding of biological systems. This thesis presented several solutions that demonstrate the effectiveness of text mining to systems biology.

The text mining solutions presented in this thesis can effectively extract useful information and, as new improvements are made to the state-of-the-art, the quality of the extracted information will be higher. This can be seen in other domains where text mining has been applied, such as web pages and social media; in these domains more data is available, leading to more specific approaches and results with higher quality.

9.1 Summary of contributions

The main contribution of this thesis was the text mining solutions developed specifically for systems biology and disease network discovery, applied to various biomedical domains. More specifically, I focused on two main case studies: miRNA-gene regulations in Cystic Fibrosis and cell-cytokine relations for tolerogenic cell therapies. For each of these case studies, I applied the developed solution to new documents, generating a knowledge graph supported by the literature. We performed an overall evaluation of these knowledge graphs, obtaining positive results. This section provides an overview of the contributions related to each of the objectives initially defined in the Introduction chapter.

9.1.1 Objective 1

The first objective was the recognition of biomedical entities in text. I explored two solutions to this task: one that was simple to use, easily adapted to different domains but with limited quality of results (MER), and another that obtained better results but was less adaptable and more computationally demanding (IBEnt). Since each solution has its advantages and disadvantages, they cover different user needs: if the objective is to find information about a specific topic which has a limited range of entity names, MER would be more suitable, but for a topic that has more available resources, IBEnt would provide results with more quality. Named entity recognition is the basis of many text mining tasks, so it is fundamental to select a tool that can provide results good enough for downstream applications. For this reason, both solutions were evaluated on public datasets and community challenges. The work developed for this objective resulted in two software tools, two journal publications (both Q1 Scimago Journal Rank) and four participations in text mining challenges.

9.1.2 Objective 2

The second objective consisted in linking entities found in documents to concepts established by a reference knowledge base. This way, different nomenclatures for the same concepts are linked, and knowledge from different sources is more easily combined. This work was developed in parallel with Objective 3 since it does not depend directly on its results. One of the solutions developed for entity linking, MER, also incorporates a simple solution to this task: the entity names are linked directly to ontology concepts though string matching. MER provided a baseline for a more complex solution, PPR-SSM, that takes into account the other entities mentioned in the same document and maximizes the overall coherence. I developed a measure of coherence between two ontology concepts, which combined Personalized Pagerank, a ranking algorithm, with semantic similarity measures calculated on ontologies, improving the accuracy of the entity linking process. This solution obtained a maximum accuracy of 0.8039 on a biomedical case-study. A manuscript was written about this approach and submitted to an international conference (Core A).

9.1.3 Objective 3

The third objective consisted of extracting relations between biomedical entities found in documents. This objective was essential to establish a network of relations between concepts. However, due to the difficulty of this task, more complex methods were necessary. I developed a distant supervision solution that uses existing resources, such as documents, databases, and ontologies to extract specific types of relations, therefore mitigating the need for manually annotated documents by domain experts. Furthermore, I developed a solution based on a deep learning algorithm that integrates domain knowledge from ontologies, improving upon the baseline performance. Due to the variety of biomedical domains involved in systems biology, it is crucial that text mining solutions can be adapted with relative ease. The biomedical community has adopted biomedical ontologies as a way to formalize the

knowledge of a particular subject, for example, the Gene Ontology establishes the terms used to describe genes and gene products and is widely used in bioinformatics. For this reason, I focused on ontology-based solutions, since these can be applied to various biomedical domains that have their concepts formalized in an ontology. This component obtained a F-score of 0.751, which is comparable to the state-of-the-art for this task. The work developed for this objective resulted in two software tools, three journal publications (all Q1 Scimago Journal Rank), and one participation in a text mining challenge.

9.2 Future work

This thesis presented various solutions to biomedical text mining, which can work as modules of a larger system. Therefore, a natural next step would be to integrate the developed solutions as a single pipeline that could be used by other researchers. In this case, we have to consider two types of users: developers or knowledgeable users who may want to integrate and adapt the tools, and researchers who want to apply the tools to their data. For developers, a command-line interface to the pipeline would be ideal, while for researchers, a graphical user interface should be developed, along with a detailed tutorial. Furthermore, tools to visualize the extracted information can also be developed, making it easier to understand the concepts and relations.

Although the performance of these approaches can be optimized to obtain high-quality results, we preferred to demonstrate that they could be applied to various biological systems. The results obtained should be analyzed by experts to be effectively validated. However, this validation is a process that requires less effort than manual curation of a database, since the information to be validated is already filtered from a larger quantity of data.

I explored crowdsourcing approaches briefly during my doctoral work, however, in the future, a more extensive study could be done to compare domain experts and crowdsourced annotation as a method to obtain training data or to evaluate text mining results. Crowd-

sourcing could be particularly useful in Entity Linking since the crowd could select another candidate match in case the top one was not correct. This procedure is common in search engines, where the top result should be the most relevant, and user interaction is taken into consideration.

Another project to be explored in the future is an in-depth extraction of information from the literature about a set of diseases, integrating genes, metabolites, and phenotypes in a knowledge graph that could be used to do a comparative analysis and explore new hypotheses. By taking advantage of existing resources and the approaches presented in this thesis, it would be possible to perform large-scale information extraction on the biomedical literature. The knowledge graph generated can then be used to find new patterns and generate new hypotheses, which could then be validated with clinical data. The results of this validation could then lead to new findings, new hypotheses to be tested, and the improvement of the text mining approaches.