UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA ANIMAL

Extracting Phenotype-Gene Relations from Biomedical Literature Using Distant Supervision and Deep Learning

Diana Francisco de Sousa

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:

Professor Doutor Francisco José Moreira Couto

2019

Acknowledgements

Firstly, I would like to offer my special thanks to my teacher and supervisor, Professor Francisco Couto, for his guidance and support throughout this project. I also want to thank André Lamúrias, for his suggestions and corrections that were extremely important for the development, and conclusion of this dissertation. To my day-to-day LASIGE colleagues (Márcia, Telma, Joana, Sofia, Alexandra, Soraia, Miguel, Vinícius, Nuno, Adriano, Fernando, Rita, Margarida, Teresa, and Sofia), for coffee, lunch, and for all the great team spirit and working environment. To my friend Beatriz Lopes, for being present at every step of the way, understanding every struggle and every success. To all my other dear friends (Sofia, Clara, Íris, and Inês) with whom I can always count for great conversations and support, wherever they are in the world. To my parents that allowed me to have all these amazing opportunities, and believed in me always. To my sister for always having my back, and making me proud of her every day. Finally, I would like to thank Rafael Ramos for unconditional love and support.

Resumo

As relações entre fenótipos humanos e genes são fundamentais para entender completamente a origem de algumas abnormalidades fenotípicas e as suas doenças associadas. A literatura biomédica é a fonte mais abrangente dessas relações. Diversas ferramentas de extração de relações têm sido propostas para identificar relações entre conceitos em texto muito heterogéneo ou não estruturado, utilizando algoritmos de supervisão distante e aprendizagem profunda. Porém, a maioria dessas ferramentas requer um *corpus* anotado e não há nenhum *corpus* disponível anotado com relações entre fenótipos humanos e genes.

Este trabalho apresenta o *corpus Phenotype-Gene Relations* (PGR), um *corpus* padrão-prata de anotações de fenótipos humanos e genes e as suas relações (gerado de forma automática) e dois módulos de extração de relações usando um algoritmo de *distantly supervised multi-instance learning* e um algoritmo de aprendizagem profunda com ontologias biomédicas. O *corpus* PGR consiste em 1712 resumos de artigos, 5676 anotações de fenótipos humanos, 13835 anotações de genes e 4283 relações. Os resultados do *corpus* foram parcialmente avaliados por oito curadores, todos investigadores nas áreas de Biologia e Bioquímica, obtendo uma precisão de 87,01%, com um valor de concordância inter-curadores de 87,58%. As abordagens de supervisão distante (ou supervisão fraca) combinam um *corpus* não anotado com uma base de dados para identificar e extrair entidades do texto, reduzindo a quantidade de esforço necessário para realizar anotações manuais. A *distantly supervised multi-instance learning* aproveita a supervisão distante e um *sparse multi-instance learning algorithm* para treinar um classificador de extração de relações, usando uma base de dados padrão-ouro de relações entre fenótipos humanos e genes. As ferramentas de aprendizagem profunda de extração de relações, para tarefas de prospeção de textos biomédicos, raramente tiram proveito dos recursos específicos existentes para cada domínio, como as ontologias biomédicas. As ontologias biomédicas desempenham um papel fundamental, fornecendo informações semânticas e de ancestralidade sobre uma entidade. Este trabalho utilizou a *Human Phenotype Ontology* e a *Gene Ontology*, para representar cada par candidato como a sequência de relações entre os seus ancestrais para cada ontologia. O *corpus* de teste PGR foi aplicado aos módulos de extração de relações desenvolvidos, obtendo resultados promissores, nomeadamente 55,00% (módulo de aprendizagem profunda) e 73,48% (módulo de *distantly supervised multi-instance learning*) na medida-F. Este *corpus* de teste também foi aplicado ao BioBERT, um modelo de representação de linguagem biomédica pré-treinada para prospeção de texto biomédico, obtendo 67,16% em medida-F.

Palavras Chave: Literatura Biomédica, Extração de Relações, *Corpus* Padrão-Prata, Supervisão Distante, Aprendizagem Profunda.

Abstract

Human phenotype-gene relations are fundamental to fully understand the origin of some phenotypic abnormalities and their associated diseases. Biomedical literature is the most comprehensive source of these relations. Several relation extraction tools have been proposed to identify relations between concepts in highly heterogeneous or unstructured text, namely using distant supervision and deep learning algorithms. However, most of these tools require an annotated corpus, and there is no corpus available annotated with human phenotype-gene relations.

This work presents the Phenotype-Gene Relations (PGR) corpus, a silver standard corpus of human phenotype and gene annotations and their relations (generated in a fully automated manner), and two relation extraction modules using a distantly supervised multi-instance learning algorithm, and an ontology-based deep learning algorithm. The PGR corpus consists of 1712 abstracts, 5676 human phenotype annotations, 13835 gene annotations, and 4283 relations. The corpus results were partially evaluated by eight curators, all working in the fields of Biology and Biochemistry, obtaining a precision of 87.01%, with an inter-curator agreement score of 87.58%. Distant supervision (or weak supervision) approaches combine an unlabeled corpus with a knowledge base to identify and extract entities from text, reducing the amount of manual effort necessary. Distantly supervised multi-instance learning takes advantage of distant supervision and a sparse multi-instance learning algorithm to train a relation extraction classifier, using a gold standard knowledge base of human phenotype-gene relations. Deep learning relation extraction tools, for biomedical text mining tasks, rarely take advantage of existing domain-specific resources, such as biomedical ontologies. Biomedical ontologies play a fundamental role by providing semantic and ancestry information about an entity. This work used the Human Phenotype Ontology and the Gene Ontology, to represent each candidate pair as the sequence of relations between its ancestors for each ontology. The PGR test-set was applied to the developed relation extraction modules, obtaining promising results, namely 55.00% (deep learning module), and 73.48% (distantly supervised multi-instance learning module) in F-measure. This test-set was also applied to BioBERT, a pre-trained biomedical language representation model for biomedical text mining, obtaining 67.16% in F-measure.

Keywords: Biomedical Literature, Relation Extraction, Silver Standard Corpus, Distant Supervision, Deep Learning.

Resumo Alargado

A literatura biomédica é o principal meio que os investigadores utilizam para partilhar as suas descobertas, maioritariamente na forma de artigos, patentes e outros tipos de relatórios escritos. Um investigador interessado num tópico específico precisa de estar atualizado em relação aos trabalhos desenvolvidos sobre esse tópico. No entanto, o volume de informação textual disponível supera amplamente a capacidade de análise de um investigador, mesmo restringindo a um domínio específico. Não só isso, mas a informação textual disponível é geralmente apresentada num formato não estruturado ou altamente heterogéneo. Assim, a recuperação de informação relevante exige não só uma quantidade considerável de esforço manual, mas também é uma tarefa que consome demasiado tempo.

Os artigos científicos são a principal fonte de conhecimento para entidades biomédicas e as suas relações. Essas entidades incluem fenótipos humanos, genes, proteínas, substâncias químicas, doenças e outras entidades biomédicas inseridas em domínios específicos. Uma fonte abrangente de artigos sobre este tópico é a plataforma PubMed, que combina mais de 29 milhões de citações, fornecendo acesso aos seus metadados. O processamento desse volume de informação só é viável através de soluções de prospeção de texto.

Os métodos automáticos de Extração de Informação (EI) visam obter informações úteis de grandes conjuntos de dados. As soluções de prospeção de texto usam métodos de EI para processar documentos de texto. Os sistemas de prospeção de texto geralmente incluem tarefas de *Named-Entity Recognition* (NER), *Named-Entity Linking* (NEL) e Extração de Relações (ER). O NER consiste em reconhecer entidades mencionadas no texto, identificando o seu primeiro e último carácter. O NEL consiste em mapear as entidades reconhecidas a entradas numa determinada base de dados. A ER consiste em identificar relações entre as entidades mencionadas num determinado documento. Algumas das relações biomédicas comummente extraídas são as interações proteína-proteína, interações fármaco-fármaco e relações gene-doença.

A ER pode ser executada por diferentes métodos, a saber, por ordem de complexidade: coocorrência, baseados em padrões (criados manual e automaticamente), baseados em regras (criados manualmente e automaticamente) e aprendizagem automática (*feature-based*, *kernel-based*, *multi-instance* (MIL) e *recurrent neural networks* (RNN)). O método de *distantly supervised multi-instance learning* utiliza uma base de dados de relações padrão-ouro do domínio de interesse (supervisão distante) combinada com um *sparse multi-instance learning algorithm* (sMIL) para executar a ER. A supervisão distante pressupõe que qualquer frase que mencione um par de entidades correspondente a uma entrada na base de

dados provavelmente descreverá uma relação entre essas entidades. Essas relações candidatas podem ser usadas para treinar um classificador usando o algoritmo sMIL. Mais recentemente, técnicas de aprendizagem profunda, como a RNN, provaram obter excelentes resultados em várias tarefas de Processamento de Linguagem Natural (PNL), entre elas a ER. O sucesso da aprendizagem profunda para a PNL biomédica deve-se em parte ao desenvolvimento de modelos de vectores de palavras como o Word2Vec e, mais recentemente, o ELMo, o BERT, o GPT, o Transformer-XL e o GPT-2. Estes modelos aprendem representações vetoriais de palavras que capturam as relações sintáticas e semânticas de palavras e são conhecidos como *word embeddings*. As *Long Short-Term Memory* (LSTM) RNN constituem uma variante de redes neuronais artificiais apresentadas como uma alternativa às RNN. As redes LSTM lidam com frases mais complexas, sendo por isso mais adequadas à literatura biomédica. Em redes LSTM, é possível integrar fontes externas de conhecimento, como ontologias de domínio específico. As ontologias são formalmente organizadas em formatos legíveis por máquinas, facilitando a sua integração em modelos de extração de relações.

O desafio contemporâneo da análise genética é correlacionar os genes aos seus respetivos fenótipos. Os sistemas existentes que têm flexibilidade para serem aplicados na identificação e extração de relações entre fenótipos humanos e genes, oriundos da literatura biomédica, são escassos e limitados. Os principais desafios que eles enfrentam são a falta de dados anotados; dificuldades na identificação de entidades fenotípicas, que são compostas de múltiplas palavras, o que torna complexo a identificação das fronteiras de cada entidade; e uma escassez de especialistas para realizar a correção das relações identificadas. Todos os problemas acima mencionados geram a necessidade de uma criação automatizada de *corpora* e o desenvolvimento de sistemas de aprendizagem automática que possam lidar com a versatilidade das entidades genéticas e fenotípicas humanas e as suas relações, para melhor identificá-las e extraí-las do texto.

Este trabalho divide-se em três etapas, o *corpus Phenotype-Gene Relations* (PGR), um *corpus* padrão-prata de anotações de fenótipos humanos e genes e as suas relações (gerado de forma automática), e dois módulos de extração de relações usando um algoritmo de *distantly supervised multi-instance learning* e um algoritmo de aprendizagem profunda com ontologias biomédicas.

Para realizar a primeira etapa, precisamos de um *pipeline* que realize NER para reconhecer genes e entidades fenotípicas humanas, e ER para extrair e classificar uma relação entre cada fenótipo humano e gene identificado. O primeiro passo é coletar resumos de artigos usando a API do PubMed com palavras-chave definidas manualmente, ou seja, cada nome de cada gene que participa numa relação (presente numa base de dados), *homo sapiens* e *disease*. Em seguida, a etapa NER é realizada usando a ferramenta *Minimal Named-Entity Recognizer* (MER) para extrair menções de genes, e a ferramenta *Identifying Human Phenotypes* (IHP) para extrair menções de fenótipos humanos, a partir dos resumos dos artigos. Por fim, usando uma base de dados de relações padrão-ouro, fornecida pela *Human Phenotype Ontology* (HPO), as relações obtidas pela coocorrência das entidades na mesma frase são marcadas como *Conhecida* ou *Desconhecida*. As relacções marcadas com *Conhecida* são relações presentes na base de dados e as relações marcadas com *Desconhecida* são relações que não estão ainda identificadas ou que não existem. O *corpus* de teste foi criado selecionando aleatoriamente 260 relações para serem revistas por oito curadores (50 relações cada, com uma sobreposição de 20 relações), todos investigadores

nas áreas de Biologia e Bioquímica, obtendo uma precisão de 87,01%, com um valor de concordância inter-curadores de 87,58%.

Enquanto na primeira etapa se utiliza uma abordagem de supervisão distante para marcar cada relação extraída como *Conhecida* ou *Desconhecida*, na segunda etapa o *corpus* PGR sem anotações vai ser usado para aplicar a abordagem de *distantly supervised multi-instance learning*. Estas duas abordagens de supervisão distante diferem na forma como são aplicadas, como vai ser possível verificar na descrição das respetivas metodologias.

Na segunda etapa, o objetivo era usar o *corpus* gerado na primeira etapa combinado com uma base de dados (fornecida pelo HPO), que fornece exemplos para a relação que queríamos extrair, para aplicar *distantly supervised multi-instance learning*. A melhor característica desta abordagem de aprendizagem automática é o facto de ela não requerer as anotações de relações, apenas anotações das entidades, neste caso fenótipos humanos e genes, reduzindo a quantidade de esforço necessário para realizar anotações manuais.

Para a última etapa, o objetivo principal foi combinar algoritmos de RNN (aprendizagem profunda) com ontologias biomédicas para melhorar a identificação das relações entre fenótipos e genes humanos na literatura biomédica. As ontologias como o HPO e a *Gene Ontology* fornecem uma representação confiável dos seus respetivos domínios e podem ser usadas como camadas de representação de dados para extrair relações do texto. O sistema proposto representa cada par candidato como a sequência das relações entre as entidades ancestrais na sua respetiva ontologia e combina os *word embeddings* e a WordNet (uma ontologia genérica da língua inglesa) para produzir um modelo capaz de extrair as relações do texto.

O *corpus* de teste PGR foi aplicado aos módulos de extração de relações desenvolvidos, obtendo resultados promissores, nomeadamente 55,00% (módulo de aprendizagem profunda) e 73,48% (módulo de *distantly supervised multi-instance learning*) na medida-F. Este *corpus* de teste também foi aplicado ao BioBERT, um modelo de representação de linguagem biomédica pré-treinada para prospeção de texto biomédico, obtendo 67,16% em medida-F.

O uso de diferentes fontes de informação, como dados adicionais, para apoiar a procura automatizada de relações entre conceitos biomédicos contribui para o desenvolvimento de farmacogenómica, triagem de testes clínicos e identificação de reações adversas a medicamentos. A identificação de novas relações pode ajudar a validar os resultados de investigações recentes e até propor novas hipóteses experimentais.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter presents the motivation, objectives, general methodology, and contributions of this dissertation, as well as the overall document structure.

1.1 Motivation

Biomedical literature is the main medium that researchers use to share their findings, mainly in the form of articles, patents, and other types of written reports [Hearst, 1999]. A researcher working on a specific topic needs to be up-to-date with all developments regarding the work done on the same topic. However, the volume of textual information available widely surpasses the ability of analysis by a researcher even if restricting it to a domain-specific topic. Not only that, but the textual information available is usually in an unstructured or highly heterogeneous format. Thus, retrieving relevant information requires not only a considerable amount of manual effort but is also a time-consuming task.

Scientific articles are the primary source of knowledge for biomedical entities and their relations. These entities include human phenotypes, genes, proteins, chemicals, diseases, and other biomedical entities inserted in specific domains. A comprehensive source for articles on this topic is the PubMed[1] platform, combining over 29 million citations while providing access to their metadata. Processing this volume of information is only feasible by using text mining solutions.

Automatic methods for Information Extraction (IE) aim at obtaining useful information from large data sets [Lamurias and Couto, 2019b]. Text mining uses IE methods to process text documents. Text mining systems usually include Named-Entity Recognition (NER), Named-Entity Linking (NEL), and Relation Extraction (RE) tasks. NER consists of recognizing entities mentioned in the text by identifying the offset of its first and last character. NEL consists of mapping the recognized entities to entries in a given knowledge base. RE consists of identifying relations between the entities mentioned in a given document. A detailed definition of these tasks will be provided in Section 2.2.1. Some of the commonly

[1] https://www.ncbi.nlm.nih.gov/pubmed/

extracted biomedical relations are protein-protein interactions [Papanikolaou et al., 2015], drug-drug interactions [Lamurias et al., 2019] and disease-gene relationships [Kim et al., 2017].

RE can be performed by different methods, namely, by order of complexity, co-occurrence, pattern-based (manually or automatically created), rule-based (manually or automatically created), and machine learning (feature-based, kernel-based, multi-instance, and Recurrent Neural Networks (RNN)). Distantly supervised multi-instance learning uses a knowledge base of gold standard target relations (distant supervision) combined with a sparse multi-instance learning (sMIL) algorithm [Bunescu and Mooney, 2007] to perform RE. Distant supervision 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 [Lamurias et al., 2017]. These candidate relations can be used to train a classifier using the multi-instance algorithm. More recently, deep learning techniques, such as RNN, have achieved outstanding results at various Natural Language Processing (NLP) tasks, among them RE. The success of deep learning for biomedical NLP is in part due to the development of word vector language models like Word2Vec [Mikolov et al., 2013], and, more recently, ELMo [Peters et al., 2018], BERT [Devlin et al., 2018], GPT [Radford et al., 2018], Transformer-XL [Dai et al., 2019], and GPT-2 [Radford et al., 2019]. These models learn word vector representations also known as word embeddings that capture the syntactic and semantic word relationships. Long Short-Term Memory (LSTM) networks constitute a variant of artificial neural networks presented as an alternative to regular RNN [Hochreiter and Schmidhuber, 1997]. LSTM networks deal with more complex sentences, making them more fitting for biomedical literature.

The knowledge encoded in the various domain-specific ontologies, such as the Gene Ontology (GO) [Ashburner et al., 2000], the Chemical Entities of Biological Interest (ChEBI) ontology [Hastings et al., 2015], and the Human Phenotype Ontology (HPO) [Köhler et al., 2017] is deeply valuable for detection and classification of relations between different biomedical entities. Besides that these ontologies make available important characteristics about each entity, they also provide us with the underlying semantics of the relations between those entities, such as is-a relations. For example, *neoplasm of the endocrine system* (HP:0100568), a phenotypic abnormality that describes a tumor (abnormal growth of tissue) of the endocrine system **is-a** *abnormality of the endocrine system* (HP:0000818), and **is-a** *neoplasm by anatomical site* (HP:0011793), which in turn **is-a** *neoplasm* (HP:0002664) (Figure 1.1).

The information provided by the ancestors is not expressed directly in the text and can support or disprove an identified relation. Ontologies are formally organized in machine-readable formats, facilitating their integration in relation extraction models.

Using different sources of information, as additional data, to support automating searching for relations between biomedical concepts contributes to the development of pharmacogenomics, clinical trial screening, and adverse drug reaction identification [Luo et al., 2017]. Identifying new relations can help validate the results of recent research, and even propose new experimental hypotheses.

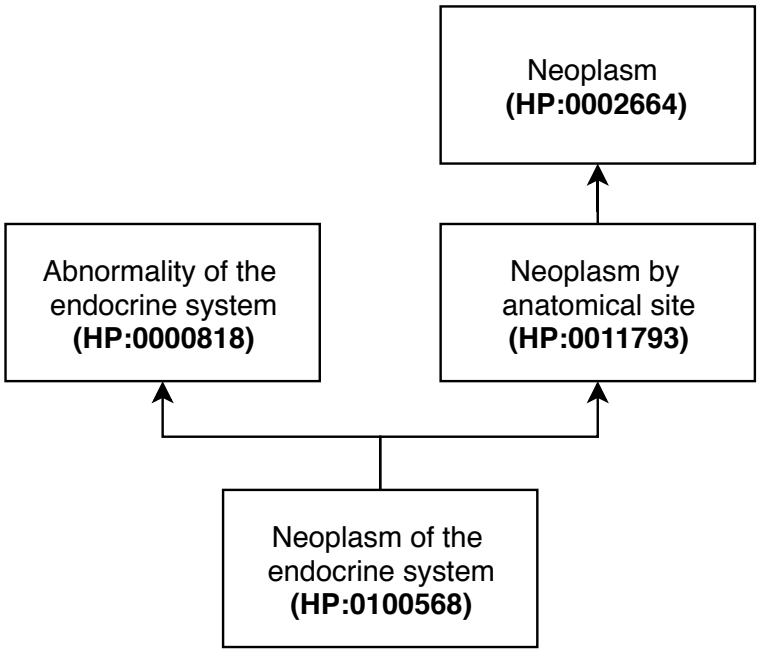

Figure 1.1: An excerpt of the HPO ontology showing the first ancestors of *neoplasm of the endocrine system*, using **is-a** relationships.

1.2 Objectives

The fundamental challenge of contemporary genetic analysis is correlating genes to their respective phenotypes. Existing systems that have the flexibility to be applied for the identification and extraction of human phenotype-gene relations, from biomedical literature, are scarce and limited. The main challenges that they face are the lack of annotated data sets; difficulties in the identification of phenotype entities, that are composed of multiple words, which makes name boundaries complex; and a scarcity of experts to perform curation of the identified relations. All of the aforementioned creates the need for automated corpora creation tools and the development of machine learning systems that can deal with the versatility of the gene and human phenotype entities and their relations, to better identify and extract them from text. Thus, the main goals of this work are:

1. Create a large and versatile silver standard corpus of human phenotype-gene relations.
2. Develop a distantly supervised multi-instance learning module that combines a knowledge base for automatic extraction of human phenotype-gene relations (added to the IBRel system [Lamurias et al., 2017]).
3. Develop a deep learning module for automatic extraction of human phenotype-gene relations, taking advantage of domain-specific ontologies, like the Human Phenotype Ontology (HPO) and the

Gene Ontology (added to the BO-LSTM system [Lamurias et al., 2019]).

First, the proposed pipeline should be able to generate a silver standard corpus based on articles dedicated to human phenotype-gene relations, using existing NER tools, and a gold standard relations knowledge base, provided by the HPO. Second, both machine learning systems (distantly supervised multi-instance learning, and deep learning) should be able to use the previous corpus to train a classifier and compare the classifications against a manually curated test-set.

The **hypothesis** of this dissertation is that information about human phenotype-gene relations can be efficiently extracted from biomedical literature using an automatically generated corpus, and machine learning techniques along with domain-specific ontologies.

1.3 Methodology

The overall methods to accomplish the proposed objectives can be divided into three stages, one for each objective. The first stage is the creation of a silver standard human phenotype-gene relations corpus (generated in a fully automated manner) (Chapter 3), the second and third stages are the development of a distantly supervised multi-instance learning module that combines a knowledge base, and the development of a deep learning module that takes advantage of domain-specific ontologies, both for automatic extraction of human phenotype-gene relations (Chapter 4).

To generate a silver standard for phenotype-gene relations, we need a pipeline that performs NER to recognize genes and human phenotype entities, and RE to extract and classify a relation between the identified human phenotype and gene entities. The first step is to gather abstracts using the PubMed API with manually defined keywords, namely, each gene name that participates in a relation (retrieved from a gold standard knowledge base of relations), *homo sapiens*, and *disease*. Then, the NER stage is performed using the Minimal Named-Entity Recognizer (MER) tool [Couto and Lamurias, 2018] to extract gene mentions, and the Identifying Human Phenotypes (IHP) tool [Lobo et al., 2017] to extract human phenotype mentions, from the abstracts. At last, using a gold standard relations knowledge base, provided by the HPO, the relations obtained by co-occurrence of the entities in the same sentence are marked *Known* or *Unknown*, and a subset (test-set) of the relations curated by domain experts. The *Known* relations are in the knowledge base and the *Unknown* relations are not yet identified or that do not exist. The test-set was created by randomly selecting 260 relations to be reviewed by eight curators (50 relations each, with an overlap of 20 relations), all researchers working in the areas of Biology and Biochemistry.

While in the first stage a distant supervision approach is used to mark the relations with *Known* or *Unknown*, in the second stage the unlabeled silver standard corpus is going to be used to apply the distantly supervised multi-instance learning approach. These two distant supervision approaches differ in the way they are applied, as we are going to see in the following chapters.

In the second stage, the goal is to use the corpus generated in the first stage unlabeled (annotated only

with entity mentions) combined with a knowledge base (provided by the HPO), that provides examples for the relations we wanted to extract, to apply distantly supervised multi-instance learning. The best feature of this machine learning approach is the fact that it does not require the relations annotations, only the human phenotype and gene entities mentions, reducing the amount of manual effort necessary.

For the last stage, the main goal is to combine RNN (deep learning) algorithms with biological ontologies to improve the identification of human phenotype-gene relations in biomedical literature. Ontologies such as the HPO and the Gene Ontology provide a reliable representation of their respective domains and can be used as data representation layers to extract relations from text. The proposed system is going to represent each candidate pair as the sequence of the relations between the entities ancestors in their respective ontology and combine word embeddings and WordNet (generic English language ontology) to produce a model able to extract the *Known* relations from text.

1.4 Contributions

The main contribution of this dissertation was a feasible solution to identify and extract human phenotype-gene relations from text, that may be applied to other types of biomedical relations. This dissertation created the first corpus specific to human phenotype-gene relations, in an attainable and reproducible way, and two different system modules to extract these type of relations from highly heterogeneous text. Both the silver standard corpus and the developed modules evaluation was done with a test-set curated by domain experts. This section provides an overview of the contributions related to each of the objectives initially defined in Section 1.2. One contribution that did not corresponded to the initially defined goals was a book chapter presenting the base concepts for neural networks using ontologies for RE:

- **Book Chapter Submitted** [Sousa et al., 2019b]: *Using Neural Networks for Relation Extraction from Biomedical Literature for the book Artificial Neural Networks: Methods and Applications* (Diana Sousa, André Lamúrias, and Francisco M. Couto) in the Springer "Methods in Molecular Biology" series.

1.4.1 Objective 1

Chapter 3 presents a pipeline to generate a silver standard human phenotype-gene relations corpus. The pipeline required the application of two NER tools and the availability of a list of gold standard relations. The evaluation of the corpus resorted to eight curators obtaining 87.01% in precision with an inter-agreement of 87.58%. The work developed for this objective resulted in one freely available silver standard corpus of human phenotype-gene relations[2] and one paper accepted for the proceedings of an international conference (Core A):

[2] https://github.com/lasigeBioTM/PGR

- **Paper Accepted** [Sousa et al., 2019a]: *A Silver Standard Corpus of Human Phenotype-Gene Relations* (Diana Sousa, André Lamúrias, and Francisco M. Couto) in the Proceedings of the 2019 North American Chapter of the Association for Computational Linguistics.

1.4.2 Objective 2

Section 4.1.1 presents a distantly supervised multi-instance learning module added to the IBRel system, to extract human phenotype-gene relations from text. The pipeline required a list of gold standard human phenotype-gene relations, the same as used in Chapter 3. The evaluation of the module resorted to the PGR test-set obtaining 73.48% in F-measure. The work developed for this objective produced a high-performance distantly supervised multi-instance learning module that can effectively extract human phenotype-gene relations from text.

1.4.3 Objective 3

Section 4.1.2 presents a deep learning module added to the BO-LSTM system, able to extract human phenotype-gene relations from text. The pipeline required the ontologies available for both type of entities (HPO and Gene Ontology). These were added to the module as data representation layers to feed the deep learning model. The evaluation of the module resorted to the PGR test-set obtaining 55.00% in F-measure. The work developed for this objective resulted in one journal publication (Q1 Scimago):

- **Paper Published** [Lamurias et al., 2019]: *BO-LSTM: Classifying Relations Via Long Short-term Memory Networks Along Biomedical Ontologies* (André Lamúrias, **Diana Sousa**, Luka A. Clarke, and Francisco M. Couto) in BMC Bioinformatics.

1.5 Document Structure

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

- **Chapter 2** (Related Work) introduces the basic concepts and resources that support RE techniques, namely, Natural Language Processing (NLP), text mining primary tasks, initial approaches for RE, distant supervision for RE, neural networks for RE, and evaluation measures.
- **Chapter 3** (A Silver Standard Corpus of Phenotype-Gene Relations) presents the work developed to create a silver standard corpus of human phenotype-gene relations, including methods, evaluation, results and discussion.
- **Chapter 4** (Extracting Phenotype-Gene Relations) presents the system modules developed (distantly supervised multi-instance and deep learning modules) to accommodate human phenotype-gene RE, with methods, evaluation, results and discussion, for each module, and a detailed comparison between the two.

- **Chapter 5** (Conclusion) discusses the main conclusions of this work, and indicates some directions for future work.

Chapter 2

Related Work

This chapter presents the basic concepts and resources that support Relation Extraction (RE) deep learning techniques, namely, Natural Language Processing (NLP), text mining primary tasks, initial approaches for RE, distant supervision for RE, neural networks for RE, and evaluation measures.

2.1 Natural Language Processing

Natural Language Processing (NLP) is an area in computer science that aims to derive meaning from unstructured or highly heterogeneous text written by humans. NLP covers several techniques that constitute pre-processing steps for the tasks described in Section 2.2. These NLP techniques have different goals and are often combined to obtain higher performance.

- **Tokenization**: has the purpose of breaking the text into tokens to be processed individually or as a sequence. These tokens are usually words but can also be phrases, numbers and other types of elements. The most straightforward form of tokenization is breaking the input text by whitespaces or punctuation. However, with scientific biomedical literature, that is usually descriptive and formal, we have to account for complex entities like human phenotype terms (composed of multiple words), genes (represented by symbols), and other types of structured entities. These entities tend to be morphological complex and need specialized tokenization pipelines. Some researchers use a compression algorithm [Sennrich et al., 2015], byte pair encoding (BPE), to account for biomedical vocabulary variability. BPE represents open vocabularies through a fixed-size vocabulary of variable-length character sequences, making it suitable for neural networks models, for instance.

- **Stemming and Lemmatization**: aims at reducing the variability of natural language by normalizing a token to its base form (stem) [Manning et al., 2008]. It can also take into account the context of the token, along with vocabulary and morphological analysis to determine the canonical form of the word (lemma). The stem can correspond only to a fragment of a word, but the lemma is always a real word. For instance, the stem of the word *having* is *hav* and the lemma is *have*.

- **Part-of-Speech Tagging**: consists of assigning each word of a sentence to the category where it belongs taking into account their context (e.g., verb or preposition). Each word can belong to more than one category. This feature is useful to gain information on the role of a word in a given sentence.
- **Parse Tree**: represents the syntactic structure of a sentence. There are two different types of parse trees: constituency-based parse trees and dependency-based parse trees. The main difference between the two is that the first distinguishes between the terminal and non-terminal nodes and the second does not (all nodes are terminal). In constituency-based parse trees, each node of the tree is either a *root* node, a *branch* node, or a *leaf* node. For each given sentence there is only one *root* node. The *branch* node connects to two or more *child* nodes, and the *leaf* node is terminal. These leaves correspond to the lexical tokens [Aho et al., 1986]. Dependency-based parse trees are usually simpler because they only identify the primary syntactic structure, leading to fewer nodes. Parse trees generate structures that are used as inputs for other algorithms and can be constructed based on supervised learning techniques.

2.2 Text Mining Primary Tasks

Text mining has become a widespread approach to identify and extract information from unstructured or highly heterogeneous text [Westergaard et al., 2018]. Text mining is used to extract facts and relationships in a structured form that can be used to annotate specialized databases and to transfer knowledge between domains [Fleuren and Alkema, 2015]. We may consider text mining as a sub-field of data mining. Thus, data mining algorithms can be applied if we transform text to a proper data representation, namely numeric vectors. Even if in recent years text mining tools have evolved considerably in number and quality, there are still many challenges in applying text mining to scientific biomedical literature. The main challenges are the complexity and heterogeneity of the written resources, which make the retrieval of relevant information, i.e., relations between entities, a non a trivial task. Text Mining tools can target different tasks together or separately. Some of the primary tasks are Named Entity Recognition (NER), Named-Entity Linking (NEL) and Relation Extraction (RE).

- **Named Entity Recognition (NER)**: seeks to recognize and classify entities mentioned in the text by identifying the offset of its first and last character. The workflow of this task starts by spliting the text in tokens and then labeling them into categories (part-of-speech (POS) tagging). Some tools that perform NER, used in this dissertation, are the Identifying Human Phenotypes tool (IHP) [Lobo et al., 2017] and the Minimal Named-Entity Recognizer tool (MER) [Couto and Lamurias, 2018] tools. IHP is a NER tool, specifically created to recognize HPO entities in unstructured text. It uses Stanford CoreNLP [Manning et al., 2014] for text processing and applies Conditional Random Fields trained with a rich feature set, combined with hand-crafted validation rules and a dictionary to improve the recognition of human phenotypes. MER is a NER tool which given

any lexicon or ontology (e.g., an OWL file) and an input text is able to return a list of recognized entities, their location, and links to their classes.

- **Named-Entity Linking (NEL)**: maps the recognized entities to entries in a given knowledge base. For instance, a gene can be written in multiple ways and mentioned by different names or acronyms in a text. NEL links all these different nomenclatures to one unique identifier. There are several organizations dedicated to providing identifiers, among them the National Center for Biotechnology Information (NCBI)[1] for genes, and the Human Phenotype Ontology (HPO) [Köhler et al., 2017] for phenotypic abnormalities encountered in human diseases. Also, the HUGO Gene Nomenclature Committee (HGNC) at the European Bioinformatics Institute[2] is responsible for approving unique symbols and names for human loci, including protein-coding genes, ncRNA genes, and pseudogenes, with the goal of promoting clear scientific communication. All approved symbols are stored in the HGNC database.

- **Relation Extraction (RE)**: identifies relations between entities (recognized manually or by NER) in a text. Tools mainly consider relations by the co-occurrence of the entities in the same sentence, but some progress is being made to extend this task to the full document (taking into account a global context) [Singhal et al., 2016].

The workflow of a typical RE system is presented in Figure 2.1.

2.3 Initial Approaches for Relation Extraction

Through the years, several approaches have been proposed to extract relations from biomedical literature [Lamurias et al., 2017]. Most of these approaches work on a sentence level to perform RE, due to the inherent complexity of biomedical literature.

- **Co-occurrence**: assumes that if two entities are mentioned in the same sentence (co-occur), it is likely that they are related. Usually, the application of this approach results in a higher recall (most of the entities co-occurring in a sentence participate in a relation), and lower precision. Some methods use frequency-based scoring schemes to eliminate relations identified by chance [Zweigenbaum et al., 2007]. Nowadays, most applications use co-occurrence as a baseline against more complex approaches [Bunescu et al., 2006].

- **Pattern-based**: uses manually defined and automatically generated patterns to extract relations. **Manually defined patterns** require domain expertise knowledge about the type of biomedical entities, their interactions, and the text subject at hand. Initial systems made use of regular expressions

[1] https://www.ncbi.nlm.nih.gov/
[2] http://www.genenames.org/

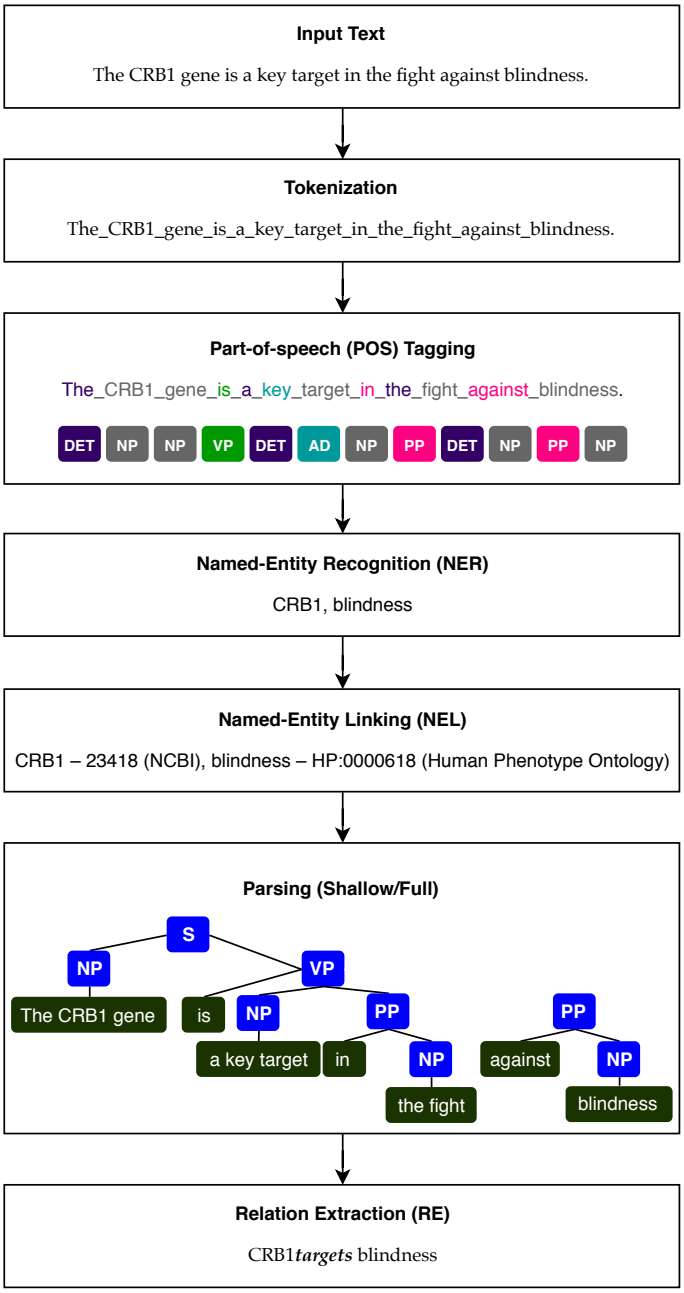

Figure 2.1: Workflow of a simplified RE system. **DET** is a determinant, **NP** is a noun, **VP** is a verb, **AD** is an adjective, and **PP** is a preposition. Text obtained from Alves and Wijnholds [2018].

to match word patterns that reflected a relation between two entities [Zhou et al., 2008], making use of a dictionary of words that express a relation, such as *trigger* and *stimulate*. Later systems introduce part-of-speech (POS) tagging, but this proven to be too naive, especially when applied

to complex sentences, such as the ones that we typically find in biomedical literature [Hao et al., 2005]. Opposite to the co-occurrence approaches, manually defined patterns frequently achieve high precision but tend to have poor recall. This approach does not generalize well, and therefore is difficult to apply to new unseen data. **Automatically generated patterns** encompass two main approaches, bootstrapping with *seeds* [Wang et al., 2011] and leveraging of the corpora [Liu et al., 2011]. The bootstrapping method uses a small set of relations known as *seeds* (e.g., gene-disease pairs). The first step is to identify the *seeds* in the data set and map the relation pattern they describe. The second step is to try to apply the mapped patterns to the data set to identify new pairs of relations that follow the same construction. Finally, expanding the original set of relations by adding these new pairs. When repeating all previous steps, if no more pairs are found, the process ends. Some systems apply distant supervision techniques to keep track of the validity of the added patterns. Distant supervision uses existing knowledge base entries as gold standards to confirm or discard a relation. This method is susceptible to noisy patterns, as the original set of relations grows. On the other hand, the leveraging of the corpora method makes immediately use of the entire data set to generate the patterns. This method requires a higher number of annotated relations and produces highly specific patterns, that are unable to match new unseen data. Automatically generated patterns can achieve a higher recall than manually defined patterns, but overall the noisy patterns continue damaging the precision. Nevertheless, there are a few efforts to reduce the number of noisy patterns [Nguyen et al., 2010].

- **Rule-based**: also uses manually defined and automatically generated rules from the training data to extract relations. Depending on the systems, the differences between pattern-based and ruled-based approaches can be minor. Ruled-based approaches not only use patterns but also additional restraints to cover issues that are difficult to express by patterns, such as checking for the negation of the relations [Koike et al., 2004]. Some ruled-based systems distance themselves from pattern-based approaches by replacing regular expressions with heuristic algorithms and sets of procedures [Rinaldi et al., 2007]. Similarly to pattern-based, ruled-based approaches tend to have poor recall, even though rules tend to be more flexible. The trade-off recall/precision can be improved using automatic methods for rule creation [Xu et al., 2012].

- **Machine Learning (ML)-based**: usually makes use of large annotated biomedical corpora (supervised learning) to perform RE. These corpora are pre-processed using NLP tools and then used to train classification models. Beyond Distant Supervision and Neural Networks, described in detail in Sections 2.4 and 2.5, respectively, it is possible to categorize ML methods into two main approaches, Feature-based and Kernel-based. **Feature-based approaches** represent each instance (e.g., sentence) as a vector in an n-dimensional space. Support Vector Machines (SVM) classifiers tend to be used to solve problems of binary classification, and are considered *black-boxs* because there is no interference of the user in the classification process. These classifiers can use different features that are meant to represent the data characteristics (e.g., shortest path, bag-of-words

(BOW), and POS tagging) [Kim, 2008]. **Kernel-based approaches** main idea is to quantify the similarity between the different instances in a data set by computing the similarities of their representations [Giuliano et al., 2006]. Kernel-based approaches add the structural representation of instances (e.g., by using parse trees). These methods can use one kernel or a combination of kernels (e.g., graph, sub-tree (ST), and shallow linguistic (SL)).

2.4 Distant Supervision for Relation Extraction

Distant Supervision (or weak supervision) heuristically assigns labels to the data in the training corpus based on a provided knowledge base. This technique considers that a pair of entities in any sentence corresponding to a knowledge base entry is likely to describe a relation between those entities. For instance, in the sentence *the **CRB1** gene is a key target in the fight against **blindness***, the *CRB1* and *blindness* entities correspond to an entry in the gold standard human phenotype-gene relations knowledge base, provided by the HPO, and therefore we assume that these entities participate in a relation. This creates a large number of false positives, because not necessarily all sentences that mention an entity pair express the target relation [Jiang et al., 2018]. Nevertheless, this allows us to use a training corpus of any size, an advantage that we do not have in supervised machine learning approaches, that require an annotated corpus.

Distant supervision is not a viable RE system by its own, but the pseudo-relations inferred using this method can be used to train a classifier through machine learning algorithms [Lamurias et al., 2017].

2.4.1 Multi-instance Learning

Multi-instance learning [Dietterich et al., 1997] can solve some of the limitations of distant supervision. This supervised machine learning method uses labeled *bags* instead of labeled instances. These *bags* contain many instances and are suited for when there is a limited amount of knowledge of the labels. The simplest case of multi-instance learning is binary classification. In this case a *bag* is labeled negative if all the instances in the *bag* are negative and positive if at least one of the instances in the *bag* is positive. Then, these labeled *bags* are fed to a learning algorithm. The algorithm that is going to be used in this dissertation is the **sparse multi-instance learning (sMIL) algorithm** [Bunescu and Mooney, 2007]. The instance-level (x) classifier $f(\overrightarrow{x};\theta)$, where θ corresponds to the parameters learned by the classifier, is learned by using a set of instances $I = I^+ \cup I^-$ that we can define as follows [Amores, 2013]:

$$\begin{aligned} I^+ &= \{\mu(X) : X \in B^+\} \\ I^- &= \{\overrightarrow{x} : \overrightarrow{x} \in B^-\} \end{aligned} \tag{2.1}$$

where I^+ and I^- are the sets of instances considered positive and negative, respectively. $\mu(X)$ is the average of instances inside X, and B^+ and B^- are the sets of positive and negative *bags*, respectively.

The idea of the sMIL algorithm is to learn a classifier with a relaxed constraint on the classification of the positive instances in I^+. The goal is to avoid forcing the classifier to provide a positive value for all the instances of a positive *bag* but only to at least one of the instances. To achieve this, the sMIL algorithm applies two sets of constraints:

$$f(\overrightarrow{x};\theta) \leq -1 + \xi_-, \quad \forall \overrightarrow{x} \in I^- \tag{2.2}$$

Equation 2.2 forces the $f(\overrightarrow{x};\theta)$ function to provide a negative value when applied to negative instances, allowing for some degree of misclassification with the ξ_- variable.

$$f(\mu(X);\theta) \geq \left(\frac{2}{\mid X \mid} - 1\right) - \xi_+, \quad \forall X \in B^+ \tag{2.3}$$

Equation 2.3 provides a more relaxed condition for positive instances, depending on the size of the *bag* X from where $\mu(X)$ is extracted. If the *bag* X only contains one instance, we use the standard condition $f(\mu(X);\theta) \geq 1 - \xi_+$ (maintaining the slack variable, ξ_+). Else, if the *bag* X contains many instances, the threshold imposed on the classifier is gradually more and more relaxed.

The sMIL algorithm assumes that the positive *bags* are sparse. An abstract may mention each entity in the candidate pair multiple times, but due to the limitation of the number of words the relation is usually stated only once. Non-biomedical RE systems already applied similar techniques to extract Freebase relations from newspaper articles [Riedel et al., 2010], and to reduce the number of incorrectly labeled relations (by distant supervision methods) [Min et al., 2013].

2.5 Neural Networks for Relation Extraction

Artificial neural networks have multiple different architectures implementations and variants. They often use data representations as additional sources of information to perform text mining tasks, and can even use ontologies as external sources of information to enrich the model.

2.5.1 Architectures

Artificial Neural Networks are a parallel combination of small processing units (nodes) which can acquire knowledge from the environment through a learning process and store the knowledge in the connections between the nodes [Haykin, 1998] (represented by direct graphs [Guresen and Kayakutlu, 2011]) (Figure 2.2). The process is inspired by the biological brain function, having each node corresponding to a *neuron* and the connections between the nodes representing the *synapses*.

Recurrent Neural Networks (RNN) is a type of artificial neural network where the connections between the nodes are able to follow a temporal sequence. This means that RNN can use their internal state, or *memory*, to process each input sequence (Figure 2.3). Deep learning techniques, such as RNN, aim to train classification models based on word embeddings, part-of-speech (POS) tagging, and other

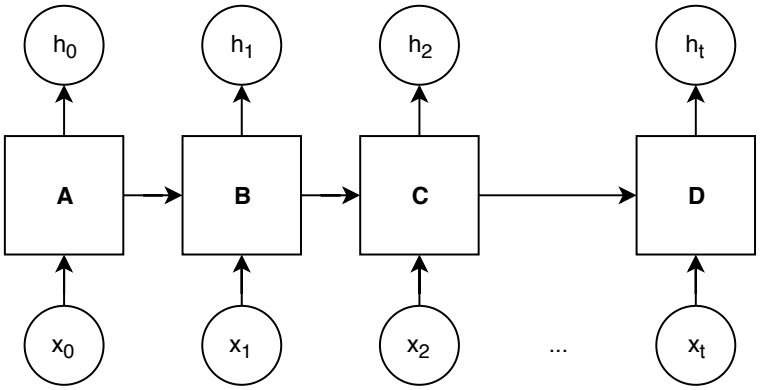

Figure 2.2: Architecture representation of an artificial neural networks model, where $x_{0\text{-}t}$ represents the inputs and $h_{0\text{-}t}$ the respective outputs, for each module from A to D.

features. RNN classifiers have multilayer architectures, where each layer learns a different representation of the input data. This characteristic makes RNN classifiers flexible to multiple text mining tasks, without requiring task-specific feature engineering.

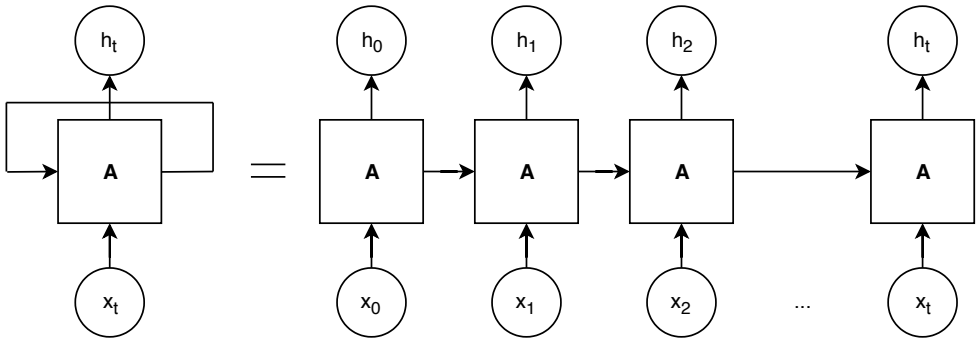

Figure 2.3: Architecture representation of a recurrent neural networks model, where $x_{0\text{-}t}$ represents the inputs and $h_{0\text{-}t}$ the respective outputs, for the repeating module A.

Long Short-Term Memory (LSTM) networks are an alternative to regular RNN [Hochreiter and Schmidhuber, 1997]. LSTMs are a type of RNN that handles long dependencies (e.g., sentences), making this classifier more suitable for the biomedical domain, where sentences are usually long and descriptive (Figure 2.4). In recent years, the use of LSTMs to perform Relation Extraction (RE) tasks has become widespread in various domains, such as semantic relations between nominals [Miwa and Bansal, 2016]. **Bidirectional LSTMs** use two LSTM layers, at each step, one that reads the sentence from right to left, and other that reads from left to right. The combined output of both layers produces a final score for each step. Bidirectional LSTMs have yield better results than traditional LSTMs when applied to the same data sets [Zhang et al., 2015].

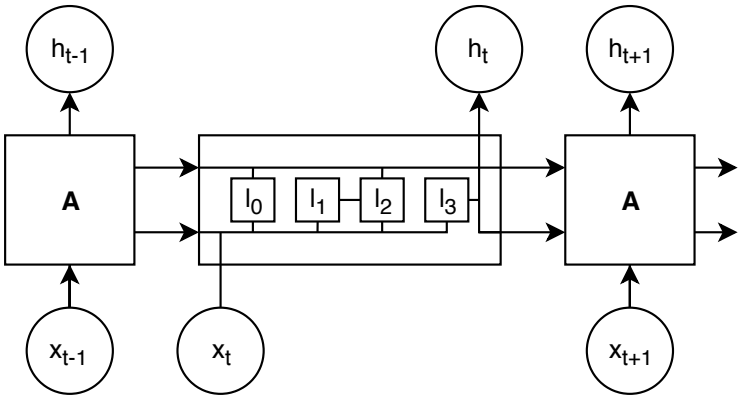

Figure 2.4: Architecture representation of a long-short-term memory networks model, where $x_{0\text{-}t}$ represents the inputs and $h_{0\text{-}t}$ the respective outputs, for the repeating module A, where each repeating module has four interacting layers ($l_{0\text{-}3}$).

2.5.2 Data Representations

The combination of multiple and different language and entity related data representations is vital for the success of neural network models dedicated to RE tasks. Some of these features were already described in Section 2.1, such as POS tagging and parse trees.

Shortest Dependency Path (SDP) is a feature that identifies the words between two entities mentioned in the text, concentrating the most relevant information while decreasing noise [Xu et al., 2015].

Word Embeddings are fixed-sized numerical vectors that aim to capture the syntactic and semantic word relationships. These word vectors models use multiple different pre-training sources, for instance, Word2Vec [Mikolov et al., 2013] uses English Wikipedia, and BERT [Devlin et al., 2018] uses both English Wikipedia and BooksCorpus. Early models, such as Word2Vec, learned one representation per word, but this proved to be problematic due to polysemous and homonymous words. Recently, most systems started to apply one embedding per word sense. One of the reasons why BERT outperforms previous methods is because it uses contextual models, meaning that it generates a unique representation for each word in a sentence. For instance, in the sentences fragments, *they got **engaged***, and *students were very **engaged** in*, the word *engaged* for non-contextual models would have the same meaning. BERT also outperforms other word vector models that take into account the sentence context, such as ELMo [Peters et al., 2018] and ULMFit [Howard and Ruder, 2018], due to being an unsupervised and deeply bidirectional pre-trained language representation.

WordNet Hypernyms are a feature that helps to hierarchize entities, structuring words similar to direct acyclic graphs [Fellbaum and Miller, 1998]. For example, *vegetable* is a hypernym of *tubers*, which in turn constitutes a hyponym of *vegetable*. This feature is comparable to an ontology in the sense that a hierarchy relation is identified, but is missing the identification of the relations between the different terms.

Using different features as information sources feeding individual channels leads to multichannel architecture models. Multichannel approaches were already proven to be effective in RE tasks [Xu et al., 2015].

Regarding biomedical RE, LSTMs were successful in identifying drug-drug interactions [Wang et al., 2017], gene-mutation relations [Song et al., 2018], drug-mutation relations [Peng et al., 2017], among others. Some methods use domain-specific biomedical resources to train features for biomedical tasks. BioBERT [Lee et al., 2019] is a domain specific language representation model pre-trained on large-scale biomedical corpora, based on BERT [Devlin et al., 2018] architecture. BioBERT, using minimal task-specific architecture modifications, significantly outperforms previous biomedical state-of-the-art models in the text mining primary tasks of Named-Entity Recognition, Named-Entity Linking, and RE. The BR-LSTM [Xu et al., 2018] model uses a multichannel approach with pre-trained medical concept embeddings. Using the Unified Medical Language System (UMLS) concepts, BR-LSTM applies a medical concept embedding method developed by De Vine et al. [2014]. BO-LSTM [Lamurias et al., 2019] uses the relations provided by domain-specific ontologies to aid the identification and classification of relations between biomedical entities in biomedical literature.

2.5.3 Ontologies

An ontology is a structured way of providing a common vocabulary in which shared knowledge is represented [Gruber, 1993]. Word embeddings can learn how to detect relations between entities but manifest difficulties in grasping the semantics of each entity and their specific domain. Domain-specific ontologies provide and formalize this knowledge. Biomedical ontologies are usually structured as a directed acyclic graph, where each node corresponds to an entity and the edges correspond to known relations between those entities. Thus, a structured representation of the semantics between entities and their relations, an ontology, allows us to use it as an added feature to a machine learning classifier. Some of the biomedical entities structured in publicly available ontologies are genes properties/attributes (Gene Ontology (GO)), phenotypes (Human Phenotype Ontology (HPO)), diseases (Disease Ontology (DO)), and chemicals (Chemical Entities of Biological Interest (ChEBI)). All of these entities participate in relations with different and same domain type entities. Hence, the information about each entity on a semantic level adds a new layer of knowledge to increase the performance of RE classifiers. For that end, this work uses the HPO and GO ontologies. The **HPO** is responsible for providing a standardized vocabulary of phenotypic abnormalities encountered in human diseases, using biomedical literature [Köhler et al., 2017]. The goal of this ontology is to facilitate medical documents readiness and exchange of medical information between medical professionals and researchers. The HPO entities are often long and descriptive, not following a specific nomenclature, making it hard to identify in unstructured text. The HPO currently contains over 13.000 terms and over 156.000 annotations to hereditary diseases. The **GO** defines a universe of concepts regarding gene functions (GO terms) and their relations [Ashburner et al., 2000]. The GO encompass three categories (sub-ontologies): *molecular function* (11.110 terms),

cellular component (4.206 terms), and *biological process* (29.687 terms), and three types of relations, *is a*, *part of* and *regulates*. As in HPO terms, GO terms are usually long and descriptive. The primary goal of this ontology is to create a dynamic controlled vocabulary that can be applied to all eukaryotes, allowing for inferences regarding gene function by connecting different organisms.

Non-biomedical models using ontologies as an added source of information to neural networks is becoming widespread for several tasks. Li et al. [2016] propose using word sense definitions, provided by the WordNet ontology, to learn one embedding per word sense for word sense disambiguation tasks. Ma et al. [2017] focus their work on semantic relations between ontologies and documents, using the DBpedia ontology. Some researchers explored graph embedding techniques [Goyal and Ferrara, 2018] that convert relations to a low dimensional space which represents the structure and properties of the graph. Other researchers have combined different sources of information, including ontological information, to do multi-label classification [Kong et al., 2013] and used ontology concepts to represent word tokens [Dasigi et al., 2017].

However, few authors have used biomedical ontologies to perform RE. Textpresso [Müller et al., 2004] is a text-mining system that works as a search engine of individual sentences, acquired from the full text of scientific articles, and articles. It integrates biomedical ontological information (e.g., of genes, phenotypes, and proteins) allowing for article and sentence search a query by term. The integration of the ontological information allows for semantic queries. This system helps database curation by automatically extracting biomedical relations. The IICE [Lamurias et al., 2014] system uses kernel-based support vector machines along with an ensemble classifier to identify and classify drug-drug interactions, linking each chemical compound to the ChEBI ontology. Tripodi et al. [2017] system focus on drug-gene/protein interaction discovery to aid database curation, making use of ChEBI and GO ontologies. BO-LSTM [Lamurias et al., 2019] is the only model until now that incorporates ancestry information from biomedical ontologies with deep learning to extract relations from the text, specifically drug-drug interactions and gene-phenotype relations.

2.6 Evaluation Measures

The evaluation of machine learning systems is done by applying the trained models to a gold standard test-set, manually curated or annotated by domain experts and unseen by the system. For a Relation Extraction (RE) task, the gold standard test-set should correspond to the list of pairs of entities (e.g., phenotype-gene or gene-disease pairs) that co-occur in the same sentences and their relation (*Known* or *Unknown*). To any given information extraction system it is necessary to define what constitutes a positive and negative result. In RE tasks the types of results possible are shown in Table 2.1.

The primary goal of a given information retrieval system is to maximize the number of TP and TN. To compare results obtained with different data sets or different tools we have three distinct evaluation metrics: recall, precision and F-measure. Precision represents how often the results are correct, recall the number of correct results identified and F-measure is a combination of both metrics to express overall

Table 2.1: Types of results obtained with an information extraction system for a RE task.

Annotator (Gold Standard)	System	Classification
Relation	Relation	True Positive (TP)
	No Relation	False Negative (FN)
No Relation	Relation	False Positive (FP)
	No Relation	True Negative (TN)

performance, being the harmonic mean of precision and recall:

$$Recall = \frac{TP}{TP + FN} \qquad Precision = \frac{TP}{TP + FP} \qquad F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.4}$$

The performance of the most recent systems dedicated to biomedical RE, described in Section 2.5.2, is shown in Table 2.2. These systems are not comparable, since each system is focused on the relations between different biomedical entities, and even addresses more than binary relations, such as the Graph LSTM (GOLD) system.

Table 2.2: Biomedical RE systems current performance.

System	Precision	Recall	F-Measure
DLSTM [Wang et al., 2017]	0.7253	0.7149	0.7200
Graph LSTM (GOLD) [Song et al., 2018]	0.4330	0.3050	0.3580
BioBERT [Lee et al., 2019]	0.8582	0.8640	0.8604
BR-LSTM [Xu et al., 2018]	0.7152	0.7079	0.7115
BO-LSTM [Lamurias et al., 2019]	0.6572	0.8184	0.7290

For RE tasks a human acceptable performance is usually around 85/90% in F-measure [Aroyo and Welty, 2015]. To facilitate the creation of gold standards we should strive for semi-automation, that is, employ automatic methods for corpora annotation (creating silver standard corpora), and then correct those annotations using domain-specific curators.

The inter-curator agreement metric, that is going to be used in this work to evaluate the quality of the curation of the silver standard corpus annotations, is calculated through the Cohen's Kappa Coefficient (κ) [Cohen, 1960]:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.5}$$

where $P(A)$ corresponds to the percentage of agreement and $P(E)$ the percentage that was expected inter-curators or inter-annotators to agree by chance.

Chapter 3

A Silver Standard Corpus of Phenotype-Gene Relations

The main problem of the systems that perform biomedical Relation Extraction (RE) is a lack of specific high quality annotated corpora, a gold standard corpus, mostly because this task requires not only a considerable amount of manual effort but also specific expertise that is not widely available. A solution to these limitations is to generate the corpus in a fully automated manner, creating a silver standard corpus.

To extract human phenotype-gene relations, both entities, human phenotypes, and genes, have to be recognized. With genes, as a result of lexical features being relatively regular, many systems can successfully identify them in text [Leaman and Gonzalez, 2008]. Even though Named-Entity Recognition (NER) research has significantly improved in the last years, human phenotype identification is still a complex task, only tackled by a handful of systems [Lobo et al., 2017].

Thus, to generate a silver standard for human phenotype-gene relation extraction, we need a pipeline that performs:

- **NER** to recognize genes and human phenotype entities.
- **RE** to classify a relation between human phenotype and gene entities.

There is no corpus available specific to human phenotype-gene relations. This chapter will present the work developed to overcome this issue, by creating a large and versatile silver standard corpus, able to be applied to Machine Learning (ML) tools. To assess the quality of the Phenotype-Gene Relations (PGR) corpus, eight curators manually evaluated a subset of the PGR.

3.1 Methods

The Human Phenotype Ontology (HPO) [Köhler et al., 2017] is responsible for providing a standardized vocabulary of phenotypic abnormalities encountered in human diseases. The developers of the HPO also

made available a knowledge base that links these phenotypic abnormalities to genes. These phenotype-gene relations are regularly extracted from texts in the Online Mendelian Inheritance in Man (OMIM) and Orphanet (ORPHA) databases, where all phenotype terms associated with any disease that is related with a gene are assigned to that gene in the relations knowledge base. In this work, the relations knowledge base created by HPO was used as a gold standard for human phenotype-gene relations.

The first step was retrieving abstracts from PubMed, using the genes involved in phenotype-gene relations and *homo sapiens* as keywords, and the Entrez Programming Utilities (E-utilities) web service[1], retrieving one abstract per gene (Query 1)[2] (Example 3.1).

Later, the keyword *disease* and a filter for abstracts in English were added (Query 2)[3]. Query 2 represents a more focused search of the type of abstracts to retrieve, such as abstracts regarding diseases, their associated phenotypes and genes.

For each gene, the system selected the most recent abstract (Query 1) the two most recent abstracts (Query 2).

Example 3.1 PubMed query 1 result example.

- **Keywords:** *NF2* and *homo sapiens*
- **Abstract Identifier:** 30194202
- **Abstract Title:** Demographical Profile and Spectrum of Multiple Malignancies in Children and Adults with Neurocutaneous Disorders

The query searched by gene name and not human phenotype or the combination of both terms because this approach was the one that retrieved abstracts with the higher number of gene and human phenotype annotations, in the following NER and RE phases. The abstracts that did not check the conditions of being written in English, with a correct XML format and content, were removed. The final number of abstracts was 1712 for Query 1 and 2657 for Query 2 as presented in Table 3.1.

The next step was to use the Minimal Named-Entity Regonition (MER) tool [Couto and Lamurias, 2018] for the annotation of the genes and the IHP framework for the annotation of human phenotype terms.

3.1.1 Gene Extraction

MER is a dictionary-based NER tool which given any lexicon or ontology (e.g., an OWL file) and an input text returns a list of recognized entities, their location, and links to their respective classes.

To annotate genes with MER it is necessary to provide a file of gene names and their respective identifiers. To this end, the system used a list created by the HUGO Gene Nomenclature Committee

[1] https://www.ncbi.nlm.nih.gov/books/NBK25501/
[2] Query 1, corresponds to the *10/12/2018* release of PGR
[3] Query 2, corresponds to the *11/03/2019* release of PGR

Table 3.1: The final number of abstracts retrieved, number of phenotype and gene annotations extracted and the number of known, unknown and total of relations extracted between phenotype and genes, for Query 1 and 2.

Query	Abstracts	Annotations		Relations		
		Phenotype	Gene	Known	Unknown	Total
1 (10/12/2018)	1712	5676	13835	1510	2773	4283
2 (11/03/2019)	2657	9553	23786	2480	5483	7963

(HGNC) at the European Bioinformatics Institute[4]. The HGNC is responsible for approving unique symbols and names for human loci, including protein-coding genes, ncRNA genes, and pseudogenes, with the goal of promoting clear scientific communication. Considering that the goal was not only to map the genes to their names but also their Entrez Gene[5] identifiers, the system used the API from MyGene[6] with the keyword *human* in species. The MyGene API provides several gene characteristics, including the confidence score for several possible genes that match the query. For this work, the choice was the Entrez Gene identifier with the higher confidence score. The first option was the Entrez Gene identifiers because of their widespread use in the biomedical research field.

After corresponding all gene names to their respective identifiers, there were three genes that did not have identifiers (*CXorf36*, *OR4Q2*, and *SCYGR9*). For the first two genes (*CXorf36* and *OR4Q2*), a simple search in Entrez Gene allowed us to match them to their identifiers. The last gene (*SCYGR9*) did not have an Entrez Gene identifier, so the second option was to use the HGNC identifier for that gene instead.

To the original gene list, were added gene synonyms using a synonyms list file[7] (expanding the original list almost 3-fold). These synonyms were matched to their identifiers and filtered according to their length to exclude one character length synonyms and avoid a fair amount of false positives. The number of genes in the original gene list was 19194, and by including their synonyms that number increased to 56670, representing a total gain of 37476 gene names.

At last, some missed gene annotations were identified using regular expressions. These missed gene annotations were next to forward/back slash and dashes characters (Example 3.2).

Example 3.2 Missed gene annotation because of forward slash.

- **Gene:** *BAX*
- **Gene Identifier:** 581

[4] http://www.genenames.org/
[5] www.ncbi.nlm.nih.gov/gene/
[6] http://mygene.info/
[7] https://github.com/macarthur-lab/gene_lists

- **Abstract Identifier:** 30273005
- **Sentence:** According to the morphological observations and DNA fragmentation assay, the MPS compound induced apoptosis in both cell lines, and also cause a significant increase in the expression of **Bax**/Bcl-2.

3.1.2 Phenotype Extraction

IHP is a Machine Learning-based NER tool, specifically created to recognize HPO entities in unstructured text. It uses Stanford CoreNLP [Manning et al., 2014] for text processing and applies Conditional Random Fields trained with a rich feature set, combined with hand-crafted validation rules and a dictionary to improve the recognition of phenotypes.

The IHP system was updated for the most recent version[8] of the HPO ontology. The annotations that originated from the IHP system were matched to their HPO identifier. There were a total of 7478 annotations for Query 1 and 10973 annotations for Query 2 that did not match any HPO identifier. These annotations were gathered to be confirmed or discarded manually, as some of them are incorrectly identified entities but others are parts of adjacent entities that can be combined for an accurate annotation.

The MER system was not used for phenotype extraction mainly because a more efficient tool for this task was available without the limitations of a dictionary-based NER tool for complex terms as phenotypes are.

3.1.3 Relation Extraction

After filtering abstracts that did not have annotations of both types, gene, and phenotype, the total of abstracts for Query 1 was 1712 and for Query 2 was 2656 abstracts as presented in Table 3.1. The abstracts retrieved by Query 1 were not specific enough for human phenotype-gene relations and therefore about half of them did not contain entities from both types, which was addressed in Query 2, increasing from an average of 2.5 relations per abstract to about 3.0 relations per abstract.

Using a distant supervision approach, with the HPO knowledge base that links phenotypic abnormalities to genes, it was possible to classify a relation with *Known* or *Unknown*. The *Known* relations are relations that are in the knowledge base and the *Unknown* relations are relations that are not yet identified or that do not exist. For this end, the system extracted pairs of entities, of gene and human phenotype, by co-occurrence in the same sentence (Example 3.3). The final number of both *Known* and *Unknown* annotations is also presented in Table 3.1.

Example 3.3 Relation extraction.

- **Abstract Identifier:** 23669344

[8] *09/10/2018* release

- **Sentence:** A homozygous mutation of **SERPINB6**, a gene encoding an intracellular protease inhibitor, has recently been associated with post-lingual, autosomal-recessive, nonsyndromic **hearing loss** in humans (DFNB91).
- **Gene:** *SERPINB6*
- **Gene Identifier:** 5269
- **Phenotype:** *hearing loss*
- **Phenotype Identifier:** HP_0000365
- **Relation: Known**

3.2 Evaluation

To evaluate the quality of the classifier, 260 relations were randomly selected from Query 1 to be reviewed by eight curators (50 relations each, with an overlap of 20 relations). Curators were researchers work in the areas of Biology and Biochemistry. These curators had to evaluate the correctness of the classifier by attributing to each sentence one of the following options: *C* (correct), *I* (incorrect) or *U* (uncertain). The *U* option was given to identify cases of ambiguity and possible errors in the NER phase. A true positive (TP) was a *Known* relation that was marked *C* by the curator, a false positive (FP) was a *Known* relation marked *I*, a false negative (FN) was a *Unknown* relation marked *I* and a true negative (TN) was a *Unknown* relation marked *C*.

3.3 Results and Discussion

The final results are presented in Table 3.2. The inter-curator agreement score, calculated from a total of 20 relations classified by eight curators, was 87.58%. Besides the fact that there were a few incorrectly extracted relations due to errors in the NER phase, that were discarded, the inter-curator agreement is not higher due to the complexity of the sentences where the relations between entities were identified. Even with highly knowledgeable curators in the fields of Biology and Biochemistry, most of them expressed difficulties in deciding which mark to choose on complex sentences that did not necessarily imply a relation between the identified entities (Example 3.4).

Example 3.4 Relation marked with *U* (Uncertain).

- **Abstract Identifier:** 27666346
- **Sentence: FRMD4A** antibodies were used to probe 78 paraffin-embedded specimens of **tongue squamous cell carcinoma** and 15 normal tongue tissues, which served as controls.

- **Mark:** ***U***

The most relevant metric for a silver standard corpus, directed towards ML tools, is precision. ML tools depend on correct examples to create effective models that can detect new cases, afterwards, being able to deal with small amounts of noise in the assigned labels.

The precision obtained from the test-set (about 6% of the total of relations), was 87.01%. Although it is not possible to state that this test-set is representative of the overall data-set, it is still strong evidence of the effectiveness of the RE corpus creation pipeline, especially between human phenotypes and genes, and other domains with a gold standard relations knowledge base. The lower recall is mostly due to incorrectly retrieved human phenotype annotations by IHP, that can be manually confirmed in a future optimized version of the PGR corpus, as some of them are parts of adjacent entities that can be combined for an accurate annotation.

Table 3.2: The number of *Known* and *Unknown* relations selected, the number of true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN), and the evaluation metrics for the *Known* relations.

Relations		**Marked Relations**				**Metrics**		
Known	**Unknown**	**TP**	**FN**	**FP**	**TN**	**Precision**	**Recall**	**F-Measure**
77	143	67	86	10	57	0.8701	0.4379	0.5826

The PGR corpus was made publicly available to the research community.[9]

[9] https://github.com/lasigeBioTM/PGR

Chapter 4

Extracting Phenotype-Gene Relations

Human phenotype-gene relations described in biomedical literature can be used to annotate specialized databases and provide a deeper understanding of the origin of some phenotypic abnormalities and their associated diseases. Most Relation Extraction (RE) approaches use supervised machine learning methods that require annotated data sets. The work developed in Chapter 3 successfully tackles the lack of data sets with the creation of the PGR corpus, using a distant supervision approach to extract the human phenotype-gene relations. However, biomedical RE machine learning systems are not prepared to deal with the specifics of human phenotype and gene entities, and their relations. This chapter proposes two different learning approaches (**Distantly Supervised Multi-instance Learning** and **Deep Learning**) for human phenotype-gene RE. These methods were developed based on existing biomedical RE systems that were modified to accommodate human phenotype-gene RE:

- **IBRel** [Lamurias et al., 2017] is a biomedical RE system, based on distantly supervised multi-instance learning, developed to extract miRNA-gene relations in the text. This system uses the TransmiR database [Qiu et al., 2009] as a knowledge base for gold standard miRNA-gene relations. The system combines distant supervision with a multi-instance learning approach, based on the sparse multi-instance learning (sMIL) algorithm [Bunescu and Mooney, 2007], to filter negative candidate pairs.
- **BO-LSTM** [Lamurias et al., 2019] is the first biomedical RE system to incorporate semantic and ancestry information from biomedical ontologies along with deep learning techniques. BO-LSTM is a system that was developed to detect and classify drug-drug interactions in text. The system represents each entity as the sequence of its ancestors in an ontology, using the Chemical Entities of Biological Interest (ChEBI) ontology [Hastings et al., 2015]. In addition to ontologies, it uses word embeddings and WordNet [Fellbaum and Miller, 1998] data representations to perform biomedical RE.

This chapter will present a brief overview of the general methods used to create the IBRel and the BO-LSTM classifiers, and a detailed description of the necessary system adjustments to accommodate human

phenotype-gene RE. These system modules, dedicated to human phenotype-gene RE, will be evaluated using the PGR corpus developed in Chapter 3. To further discuss the benefits of each approach, it is relevant to employ other RE approaches, namely, a co-occurrence (or all-true) baseline method, the state-of-the-art BioBERT application [Lee et al., 2019], and a bootstrap theoretical approach that leverages of both developed modules, also using the PGR corpus. Finally, the chapter will end with the presentation of the results of all the implementations and a comprehensive discussion of the benefits and downsides of each approach.

4.1 Methods

The main goal of all of the models described in this chapter is to extract relations from unstructured or highly heterogeneous text. However, each system expects different data format inputs (instances) and uses resources of distinct external sources of knowledge plus the training data to build them. This section will provide an overview of these differences, and the necessary steps to successfully perform human phenotype-gene RE using these learning methods.

4.1.1 Distantly Supervised Multi-instance Learning Module

The resources needed to perform distantly supervised multi-instance learning (the input text and a knowledge base) were already presented in Chapter 3. The knowledge base for the application of this method was also the HPO gold standard relations knowledge base that links phenotypic abnormalities to genes, with the added synonyms for the gene entities. Therefore, this section will focus on the *bag-of-instances* representations, the model itself, and the necessary changes needed to perform human phenotype-gene RE. Figure 4.1 shows the IBRel system general workflow, and a model simplification in bold.

The input data is used to generate instances to be classified by the model. Each instance represents a candidate pair of entities and consists of multiple relevant data representations besides the entities offsets. For example, using word context windows (of sizes 1, 3, and 5), each word, in each sentence fragment that encompasses a candidate relation, goes through lemmatization and part-of-speech (POS) tagging. Then, these instances are converted into *bag-of-instances* representations, using the scikit-learn library [Pedregosa et al., 2011].

4.1.1.1 *Bag-of-Instances* Representations and Model

Multi-instance learning is a particular case of a supervised machine learning method since it uses a training-set composed of labeled *bags* instead of labeled instances. The multi-instance learning approach performed is based on the sMIL algorithm. The sMIL algorithm assumes that the data is sparse, implicating that for each *bag* only a few candidate pairs are positive. The general assumption is that if two entities preserve a relation in a knowledge base, at least one sentence that mentions the entity pair expresses the

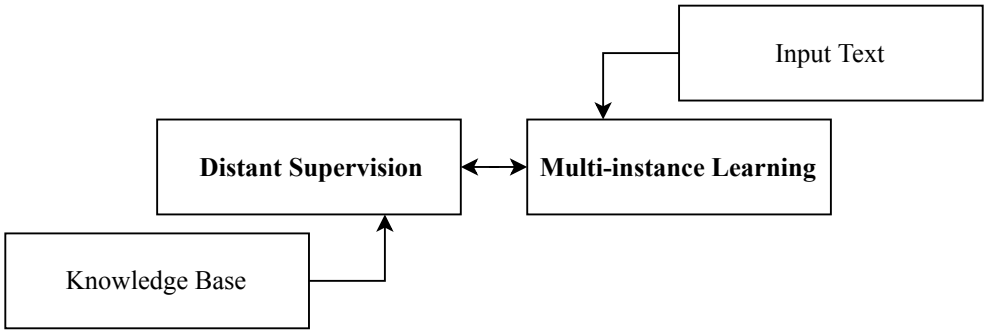

Figure 4.1: IBRel workflow. The input text corresponds to sentences retrieved from PubMed abstracts, and the knowledge base corresponds to the HPO gold standard relations knowledge base. The double arrow represents a dependency between the multi-instance learning step and the distant supervision step. Each *bag* created by the multi-instance learning step is labeled positive if a pair exists in the reference knowledge base, and negative otherwise.

relation [Surdeanu et al., 2012]. Thus, it is necessary to decide how to represent human phenotype-gene relations in the form of *bags*.

In this model, each *bag* is an instance that can contain multiple entries corresponding to the same relation candidate, for the entire corpus. Figure 4.2 presents an example sentence retrieved from the PGR corpus, with five entity annotations. Taking into account that entity E1 and entity E5 are the same, in this sentence, it is possible to extract four unique candidate pairs, that correspond to four different *bags* with the distinct text mentions. A *bag* is labeled as positive (label 1) if the candidate pair exists in the reference knowledge base, and negative (label 0) if the candidate pair does not correspond to an entry in the knowledge base.

After creating the *bags*, the sMIL algorithm was used to generate the classification model, using the miSVM package[1], with the default values.

The original IBRel system source code was in Python 2.7. For the new human phenotype-gene RE model the source code was updated to Python 3.6[2]. Most of the original packages were incompatible with this new version of the system but there were updated versions available. One of the most relevant packages for this system is the miSVM that did not have an updated version. Therefore, it was necessary to develop a Python 3 miSVM compatible version to apply to the model.

4.1.2 Deep Learning Module

This section describes the BO-LSTM model with an emphasis on the modifications to allow human phenotype-gene RE integration. Figure 4.3 shows a simplification of the overall model architecture.

[1] https://github.com/garydoranjr/misvm
[2] https://github.com/lasigeBioTM/IBRel/tree/IBRel_Python3.6

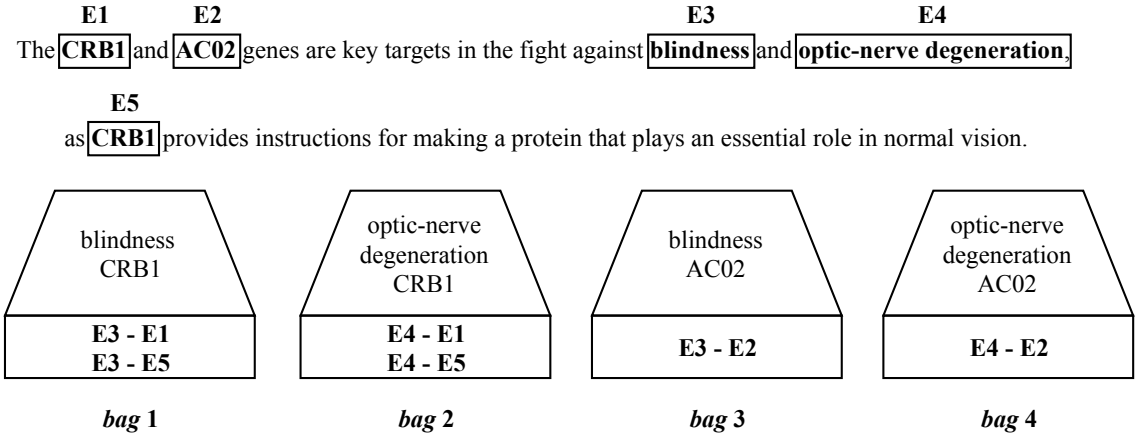

Figure 4.2: Multi-instance learning *bags*. Each *bag* represents one instance, a representation of a candidate human phenotype-gene relation, for the corpus. The *bags* 1 and 4 are positive (labeled 1), and the *bags* 2 and 3 are negative (labeled 0).

4.1.2.1 Data Representations

The BO-LSTM model uses a combination of different language and entity related data representations, that feed individual channels creating a multichannel architecture. The input data is used to generate instances to be classified by the model. Each instance corresponds to a candidate pair of entities in a sentence. To each instance, the model is going to assign a positive or negative class. In this case study, a positive class corresponds to an identified relation between a human phenotype and a gene entity, where the nature of this relation is always of *causality*, and a negative class implies no relation between the different entities.

An instance should condense all relevant information to classify a candidate pair. To create an instance the BO-LSTM model relies on three primary data information layers (Figure 4.3). After sentence tokenization, these layers are:

- **Shortest Dependency Path (SDP)** between the entities of the candidate pair. For instance, in the sentence *The **CRB1** gene is a key target in the fight against **blindness***, the shortest path between the entities would be **CRB1** - gene - is - target - in - fight - against - **blindness**, using the spaCy software library[3]. Every word is replaced by a generic string to minimize the impact of specific words in the model. The model uses pre-trained word embeddings vectors trained on abstracts and full documents from PubMed (more than 29 million) [Pyysalo et al., 2013], using the Word2Vec algorithm [Mikolov et al., 2013]. These vectors are more relevant for biomedical tasks than vectors trained on a generic corpus.
- **WordNet Classes**, using the tool developed by Ciaramita and Altun [2006], matches each element

[3] https://spacy.io/

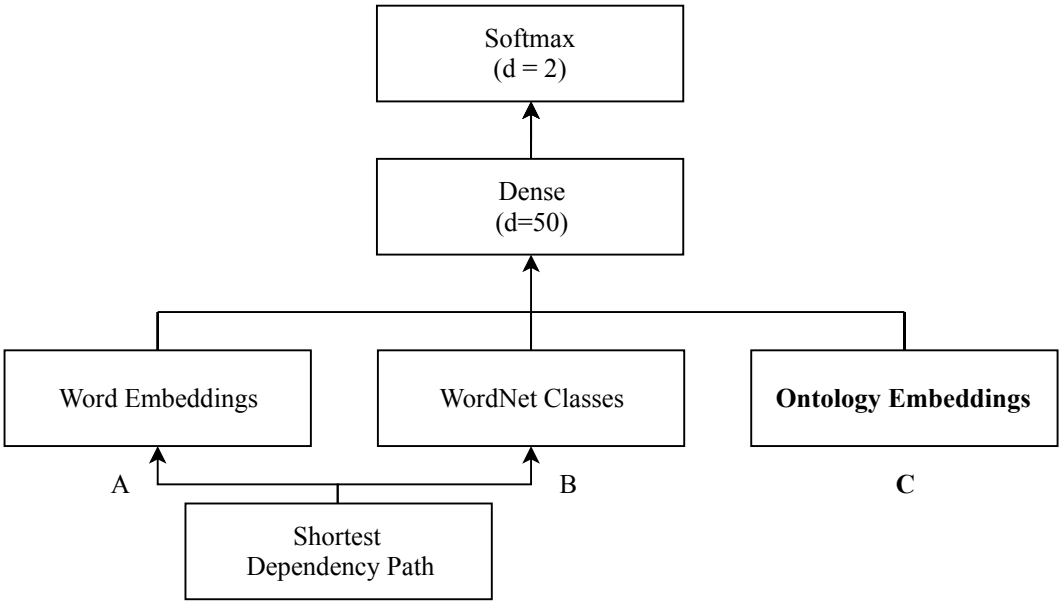

Figure 4.3: BO-LSTM model architecture simplification. **(A)**, **(B)** and **(C)** represent the three primary data information layers that are fed to the model. *d* is the dimensionality of each embedding layer, **(A)** corresponds to the Word Embeddings, **(B)** to the WordNet Channel and **(C)** the Ontology Embeddings.

in the SDP to a WordNet hypernym class. For instance, using the previous sentence, it would be **CRB1**$_{\text{noun.e1}}$ - gene$_{\text{noun}}$ - is$_{\text{verb}}$ - target$_{\text{noun}}$ - in$_{\text{adverb}}$ - fight$_{\text{noun}}$ - 0 - blindness$_{\text{noun.e2}}$.

- **Ontology Embeddings** represents the relations between the ancestors for each ontology concept corresponding to an entity (Figure 4.4).

The model assumes that the input data already has the offsets of the relevant entities identified and their respective concept ID, the Named-Entity Recognition and Linking tasks. However, while human phenotype entities identifiers already corresponded to an ontology concept ID (Human Phenotype Ontology (HPO) [Köhler et al., 2017]), gene entities had only a National Center for Biotechnology Information (NCBI) identifier[4]. Genes have several designated ontologies, for example the Gene Ontology [Ashburner et al., 2000] that describes gene properties/attributes. The Gene Ontology provides a computational representation of our current scientific knowledge about the functions of genes and it is widely used to support scientific research.

The solution to overcome this problem was to match each gene to their most representative GO concept within the *biological process* category (sub-ontology). Each gene has a corresponding set of GO terms inferred from different sources and with different degrees of confidence. NCBI provides a list[5] of genes to GO relations and their evidence codes. There are twenty-six different evidence codes, that fall

[4] https://www.ncbi.nlm.nih.gov/
[5] https://ftp.ncbi.nlm.nih.gov/gene/DATA/

into six general categories: experimental evidence codes (**EXP**, **IDA**, **IPI**, **IMP**, **IGI**, **IEP**, **HTP**, **HDA**, **HMP**, **HGI**, and **HEP**), phylogenetically-inferred annotations (**IBA**, **IBD**, **IKR**, and **IRD**), computational analysis evidence codes (**ISS**, **ISO**, **ISA**, **ISM**, **IGC**, and **RCA**), author statement evidence codes (**TAS**, and **NAS**), curator statement evidence codes (**IC**, and **ND**), and electronic annotation evidence code (**IEA**) (Table 4.1). To match the gene to their most representative GO term the priority was given to concepts inferred from experiments, for having a more sustained background and usually be more descriptive (Example 4.1). For tie-breaking, if we have several GO terms inferred from experiments, the choice is the one term that is the most specific with the longer ancestry line.

Example 4.1 Selection of the most representative GO term for the CRB1 gene, organized by priority order of the evidence code.

- **Gene:** *CRB1*
- **Gene NCBI Identifier:** 23418
- **Biological Process Annotations:**
  - GO:0007009 - IEA - plasma membrane organization **(4th)**
  - **GO:0007157 - EXP - heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules (1st)**
  - GO:0007163 - TAS - establishment or maintenance of cell polarity **(3rd)**
  - GO:0045197 - IBA - establishment or maintenance of epithelial cell apical/basal polarity **(2nd)**

4.1.2.2 Model

The most relevant part of this work is the implementation of the ontology embeddings in the model. An ontology describes a formal definition of concepts related to a specific subject and can be represented by a tuple $< C, R >$, where C represents the set of concepts in an ontology and R the set of relations between the concepts of the same ontology. The type of ontology relations considered were subsumption relations, **is-a** due to its transitive aspect. For instance, if we have $(c_1, c_2) \in R$, and $(c_2, c_3) \in R$, we assume that (c_1, c_3) is a valid relation within the ontology. The ancestors of each concept c are given by:

$$Anc(c) = a : (c, a) \in T \tag{4.1}$$

where T is the transitive closure of R. A relation between different ontology concepts can be represented by (p_1, g_1), where $p_1 \in P$ and P represents the set of concepts in the HPO, and $g_1 \in G$ and G represents the set of concepts in the Gene Ontology. For instance, if we have $(p_2, g_2) \in RA$, where RA is the set of relations between ancestors, and p_2 **is-a** p_1, and g_2 **is-a** g_1, then we can assume that (p_1, g_1)

Table 4.1: Gene Ontology (GO) evidence codes.

Category	Evidence Codes
Experimental	Inferred from Experiment (EXP)
	Inferred from Direct Assay (IDA)
	Inferred from Physical Interaction (IPI)
	Inferred from Mutant Phenotype (IMP)
	Inferred from Genetic Interaction (IGI)
	Inferred from Expression Pattern (IEP)
	Inferred from High Throughput Experiment (HTP)
	Inferred from High Throughput Direct Assay (HDA)
	Inferred from High Throughput Mutant Phenotype (HMP)
	Inferred from High Throughput Genetic Interaction (HGI)
	Inferred from High Throughput Expression Pattern (HEP)
Phylogenetically-inferred	Inferred from Biological aspect of Ancestor (IBA)
	Inferred from Biological aspect of Descendant (IBD)
	Inferred from Key Residues (IKR)
	Inferred from Rapid Divergence (IRD)
Computational Analysis	Inferred from Sequence or structural Similarity (ISS)
	Inferred from Sequence Orthology (ISO)
	Inferred from Sequence Alignment (ISA)
	Inferred from Sequence Model (ISM)
	Inferred from Genomic Context (IGC)
	Inferred from Reviewed Computational Analysis (RCA)
Author Statement	Traceable Author Statement (TAS)
	Non-traceable Author Statement (NAS)
Curator Statement	Inferred by Curator (IC)
	No biological Data available (ND)
Electronic Annotation	Inferred from Electronic Annotation (IEA)

is a valid relation. The concatenation of the relations between the ancestors of concepts p_2 and g_2 can be defined using:

$$ConcRA(p_2, g_2) = Anc(p_2) \bowtie Anc(g_2) \tag{4.2}$$

Figure 4.4 shows an overview of a representation for a candidate pair based on the ancestry of its elements, using the most representative GO term as shown in Example 4.1.

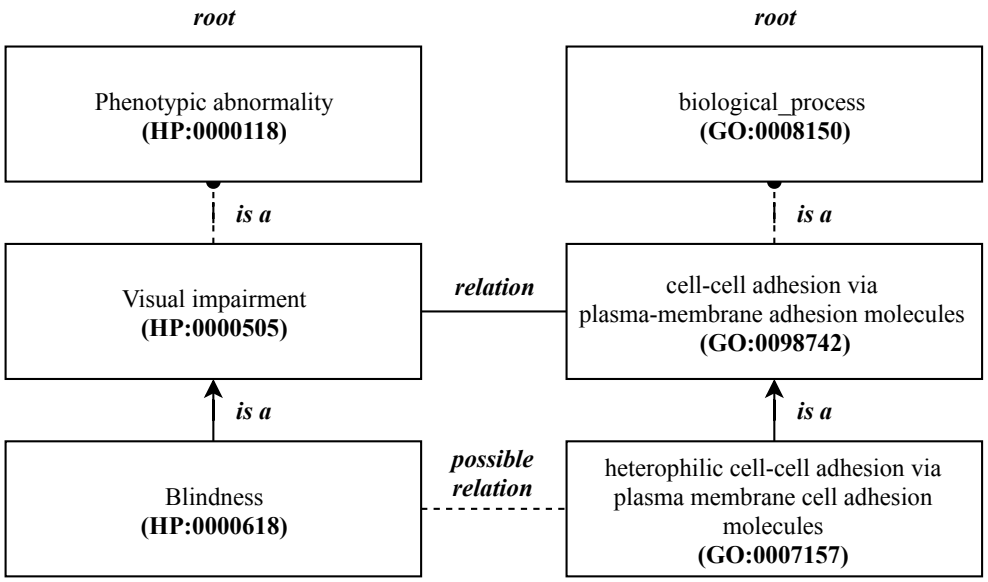

Figure 4.4: BO-LSTM ontology embedding illustration based on the HPO and the Gene Ontology, for the candidate relation between the human phenotype *blindness* and the gene *CRB1* (represented by the GO term *heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules*)

Figure 4.5 presents the detailed workflow of the Ontology Embeddings in Figure 4.3. Each concatenation of relations between ancestors corresponds to one-hot vector v_c, a vector of zeros except for the position that corresponds to the ID of the concepts. An embedding matrix $M \in \mathbb{R}^{D \times C}$ transforms these sparse vectors into dense vectors, where D is the dimensionality of the embedding layer and C is the number of concepts of the ontologies. Then, the output of the embedding layer is given by:

$$f(c) = M \cdot v_c \tag{4.3}$$

Formerly, the ontology embedding layer, with a dimensionality of 50, initializes its values randomly so that they could later be tuned, through back-propagation. This size was the one that performed the best after testing with the sizes 50, 100, and 150, as suggest by Xu et al. [2015]. Then, the sequence of vectors representing the relations between the ancestors of the terms is fed into the LSTM layer, showed in detail in Figure 4.5. Finally, the system uses a max pool layer, which is fed into a dense layer through a sigmoid activation function, and a softmax layer outputs the probability for each class.

Deep learning models map a set of inputs to a set of outputs from the training data. It is not feasible to calculate the perfect weights for a neural network, because it is not a linear problem. Thus, the issue of learning assumes the form of an optimization problem. To approach this optimization problem, we use different optimization algorithms to try to enhance our predictions. In this work, the model was

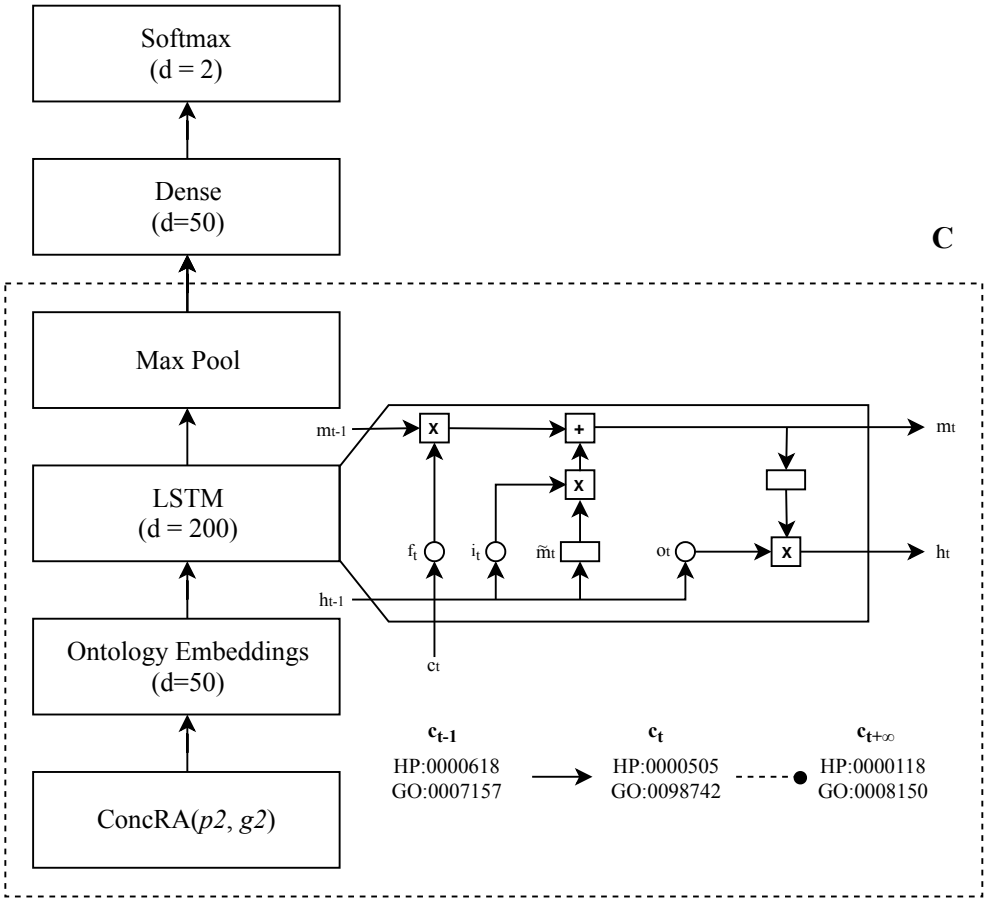

Figure 4.5: Ontology embedding workflow, using the sequence in Figure 4.4 as an example. $\circ$ refers to sigmoid function, $\square$ to tanh, x to element-wise multiplication, and $+$ to element-wise addition. h is the hidden unit, $\tilde{m}$ the candidate memory cell, m a memory cell, i the input gate, f the forget gate, and o the output gate.

trained using a stochastic gradient descent optimization algorithm where weights were updated using the back-propagation of error algorithm. At each iteration, the model with a given set of weights creates predictions and computes the error for those predictions. The optimization algorithm seeks to alter the weights to reduce that error in the next evaluation. The relevant configurations of this model are:

- **Mini-batch Gradient Descent Optimization Algorithm**: RMSprop.
- **Learning Rate**: 0.001 (default value for RMSprop).
- **Loss Function**: Categorical Crossentropy.
- **Dropout Rate**: 0.500 (every layer except the penultimate and output layers).

The dropout strategy adopted [Hinton et al., 2012] reduced the overfitting on the trained embeddings and weights.

4.2 Evaluation

The corpus used to evaluate each module was the PGR corpus, described in Chapter 3. The measures used to evaluate the performance of the RE modules were precision, recall, and F-measure.

To further assess the quality of the developed implementations, it was relevant to employ other RE approaches, namely a co-occurrence baseline method, the state-of-the-art application BioBERT, and a bootstrap theoretical approach that leverages both developed modules. These approaches were also evaluated using the PGR corpus.

4.2.1 Co-occurrence Baseline Method

The applied co-occurrence approach consists of considering every human phenotype-gene pair in the same sentence as positive. This approach produces a high number of false positives and results in a recall of 1. For even distributions of positive/negative pairs, and specifically for abstracts, the co-occurrence method can achieve, for some data sets, almost state-of-the-art results [Dai et al., 2019].

4.2.2 BioBERT Application

The BioBERT system is a new pre-trained biomedical language representation model for biomedical text mining based on the BERT [Devlin et al., 2018] architecture. This system, trained on large-scale biomedical corpora, can perform diverse biomedical text mining tasks, namely Named-Entity Recognition (NER), RE and Question Answering (QA). The novelty of the architecture is that these systems, BioBERT and BERT, are designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. This feature allows easy adaption to several tasks without loss in performance.

The PGR corpus was trained and tested using the available pre-trained weights of the BioBERT model. It was necessary to anonymize the test-set entities, using the pre-defined tags *@GENE* for gene entities and *@DISEASE* for human phenotype entities, for being the tags available closest to the case study of this thesis (human phenotype-gene relations).

4.2.3 Bootstrap Approach (Theoretical)

The bootstrap approach implemented combines the results obtained from the first system (Distantly Supervised Multi-instance Learning Module, Section 4.1.1) with the second system (Deep Learning Module, Section 4.1.2). Thus, for each candidate pair, if the classifiers disagree on the classification, and one of

them classifies the candidate pair with the correct label, that would be the chosen label. Since this approach requires knowledge of the true labels of the test-set, it cannot be used as a classifier for unlabeled data. This approach is relevant because it allows us to know where each system fails, i.e., both systems classify the same sentence in the same way (true positive, false negative, false positive, or true negative) or if their classification differs, and how much it differs, and why.

4.3 Results and Discussion

Each implementation was evaluated using the PGR corpus, developed in Chapter 3. However, before presenting the final results it is necessary to mention some evaluation limitations. It is not possible to dissociate this evaluation from the quality of the NER, Named-Entity Linking (NEL) and RE tasks performed in the previous chapter. If some entities were poorly identified, not identified at all, or not linked to the right identifier, then this will reflect on the performance of each different implementation, when using the PGR corpus. Also, the RE task was performed using a distant supervision approach resorting to a gold standard knowledge base of relations that is still evolving, growing, and therefore missing some relations. Thus, it is relevant to keep in mind the silver standard aspect of the PGR corpus, when interpreting the results. The fact that the classifiers are trained using a silver standard, and not a gold standard, damages the final performance for each of the implementations.

Table 4.2 presents the human phenotype-gene relation extraction results for each implementation. Comparing the methods in terms of F-measure, the distantly supervised multi-instance learning module obtained the best score, while the BioBERT application had the best deep learning performance. In terms of precision, the deep learning module (without ontology embeddings) obtained a slightly better score than the deep learning module using the ontology embeddings, while the BioBERT application outperformed all other implementations, even though all the deep learning-based applications had similar results. Regarding the recall, the distantly supervised multi-instance learning module clearly outperforms all other implementations.

The distantly supervised multi-instance learning module relies on the premise that only a few candidate pairs in each instance *bag* are positive. Due to the number of words restriction, is to be expected that each abstract only mentions a relation one or two times, even if it has more identified domain entities. The developed system takes into account the sparsity of the positive pairs, by implementing a sparse algorithm (sMIL). Nevertheless, the model slightly overestimates the number of positive pairs, resulting in a lower precision when compared to the other implementations. This lower precision/ higher recall is typical for systems that resort to distant supervision methods, that tend to produce noisy data leading to a higher number of false positive instances. Not necessarily all sentences that mention an entity pair express the target relation [Jiang et al., 2018]. When comparing this module with the co-occurrence baseline method, the difference in precision is 0.3386 in favor of the distantly supervised multi-instance learning module. The PGR corpus has a lower number of possible candidate pairs per entity when compared to other gold standards, boosting the performance of distantly supervised multi-instance approaches. In more focused

Table 4.2: Human phenotype-gene relations extraction results for each implementation, the distantly supervised multi-instance learning module, and the deep learning module (without ontology embeddings, and using ontology embeddings). Also, for comparison, the co-occurrence baseline method, the state-of-the-art BioBERT application, and the bootstrap approach (theoretical).

Implementations	**Metrics**		
	Precision	**Recall**	**F-Measure**
Distantly Supervised Multi-instance Learning Module	0.6886	**0.7877**	**0.7348**
Deep Learning Module (Without Ontology Embeddings)	**0.7564**	0.3933	0.5175
Deep Learning Module (With Ontology Embeddings)	0.7333	0.4400	0.5500
Co-occurrence	0.3500	1.000	0.5185
BioBERT Application	0.7895	0.5844	0.6716
Bootstrap Approach (Theoretical)	0.8472	0.8714	0.8592

gold standard data sets, manually annotated, where the manually selected documents are always relevant for the type of relations annotated, the deep learning approaches usually perform better than for corpora like the PGR. The original IBRel model best performance was on the corpus that had fewer relations per entity (more sparse data), also proving this association.

The deep learning module (without ontology embeddings, and using ontology embeddings) was implemented in Keras, a Python-based deep learning library, using the TensorFlow backend. The model used the data representations layers discussed in Section 4.1.2.1), with the hyperparameters tuned using the reference values provided by other authors that applied LSTMs to similar data sets [Sahu and Anand, 2018]. At first, the model was implemented using only the word embeddings of the SDP and the WordNet classes of each candidate pair and then using these two in addition to the ontology embeddings. With the ontology embeddings, there was an improvement of 0.0325 of the F-measure when comparing with the model that does not use ontologies. The most relevant contribution for this metric was an increase in recall, showing that applying ontology embeddings contributes to the identification of more correct relations.

The co-occurrence baseline method obtained the highest recall because it classified every human phenotype-gene pair in a sentence as a true relation. However, this resulted in a very low precision (0.3500), although the corresponding F-measure is comparable to the other implementations. Regarding the BioBERT application, it significantly outperformed all the other deep learning-based implementations with an F-measure of 0.6716, proving that is indeed a viable language representation model for biomedical text mining.

The bootstrap approach results show us that each system evaluated most of the test-set sentences differently. Table 4.3 presents one example sentence for each differently marked candidate pair between the distantly supervised multi-instance learning module and the deep learning module, as well as the

number of annotations for each combination of classifications. Sentences 2 and 3 were slightly condensed to improve their readability.

Table 4.3: Performance comparison for the distantly supervised multi-instance learning (DS) and the deep learning (DL) modules (different classifications). One example for each differently marked candidate pair between the DS model and the DL model, and the number of occurrences for each type of different classifications combination (true positive (TP), false negative (FN), false positive (FP), and true negative (TN)).

ID	Example Sentences	Module	Mark	Occurrences
1	**HDAC4**, and RUNX2 expression is suspected to be involved in the epigenetic regulations behind the **mandibular prognathism** phenotype.	DS	**FN**	12
		DL	**TP**	
2	Collectively, our study uncovers a protein complex, which consists of FIGNL1 and KIAA0146/**SPIDR**, in DNA repair and provides potential directions for **cancer** diagnosis and therapy.	DS	**FP**	28
		DL	**TN**	
3	Single nucleotide polymorphism in infant genes in the folate (**MTHFS**), and transsulfuration (GSTP1 and MGST1) pathways are associated with an increased risk of **congenital heart defects**.	DS	**TP**	64
		DL	**FN**	
4	These findings support that UNC119 is a regulator of the **RASSF6** and functions as a **tumor** suppressor.	DS	**TN**	2
		DL	**FP**	

The number of each type of combination pair of classifications reinforces the strengths and weaknesses of each classifier. The distantly supervised multi-instance learning module tends to extract a higher number of false relations, and the deep learning module a lower number of true relations. The difficulty for both approaches seems to be in longer, more complex sentences, that do not express an immediate clear relation, as we can see by the second and third sentences in Table 4.3. The distant supervision multi-instance learning-based method has this difficulty due to the overestimation of positive pairs explained previously. In what concerns to the deep learning method, the difficulty in classifying longer sentences is even more evident with the 0.4400 in the recall. Long short-term memory (LSTM) networks are a type of recurrent neural networks (RNN) supposedly more suitable for long dependencies, such as the referred sentences, but they still fail on the classification of the candidate pairs of most of those sentences. The BioBERT approach performs better in classifying these sentences, by pre-training deep bidirectional representations by jointly conditioning on both left and right context in all layers. Table 4.4 shows us three sentences where BioBERT was able to detect the positive candidate pairs, that were missed by the deep learning module. Sentence 1 corresponds to the sentence 3 in table 4.3.

Table 4.5 presents examples of sentences where both developed modules chose the same classification. Sentence 1 was condensed to improve its readability.

The numbers of wrongly identified annotations in common, 18 for false negatives and 20 for false

Table 4.4: Performance comparison for the BioBERT application and the deep learning module. Three example sentences that BioBERT was able to detect, and that were missed by the deep learning module.

ID	Example Sentences
1	Single nucleotide polymorphism in infant genes in the folate (**MTHFS**), and transsulfuration (GSTP1 and MGST1) pathways are associated with an increased risk of **congenital heart defects**.
2	Based upon the development-dependent onsets of these psychotomimetic effects, by using a DNA microarray technique, we identified the WD repeat domain 3 (WDR3) and chitobiosyldiphosphodolichol beta-mannosyltransferase (**ALG1**) genes as novel candidates for **schizophrenia**-related molecules.
3	Variants of WNK1 (lysine deficient protein kinase 1), ADRB2 (b2 adrenergic receptor), **NEDD4L** (ubiquitin-protein ligase NEDD4-like), KLK1 (kallikrein 1) contribute to **hypertension**, and AKR1C3 (aldo-keto reductase family1 member C3), is associated with preeclampsia.

Table 4.5: Performance comparison for the distantly supervised multi-instance learning and the deep learning modules (equal classifications). One example for each equally marked candidate pair by both models, and the number of occurrences for each classification combination (true positive (TP), false negative (FN), false positive (FP), and true negative (TN)).

ID	Example Sentences	Mark	Occurrences
1	**TREM2** protein levels were also negatively correlated with the severity of symptoms in humans affected by **autism**.	**FN**	18
2	In particular, MYH9 mutations result in congenital macrothrombocytopenia and predispose to **hearing loss**, and cataracts, whereas thrombocytopenias caused by **ANKRD26**, and ETV6 mutations are characterized by predisposition to hematological malignancies.	**FP**	20
3	Altogether these data demonstrate that mutations in **INPP5K** cause a **congenital muscular dystrophy** syndrome with short stature, cataracts, and intellectual disability.	**TP**	46
4	In T2D patients, **PAX4** Arg192His was associated with earlier age at diagnosis, and GLP1R Arg131Gln was associated with decreased risk of **cardiovascular disease**.	**TN**	5

positives, for the developed modules were reasonably balanced. Regarding the true negatives, the lower number of classifications in common indicates that the systems diverge in the way they classify sentences with negative candidate pairs, which ultimately means that each system identifies different sets of true negatives. As discussed above, the deep learning module has difficulties in classifying candidate pairs

in longer sentences. Thus, the same way this system is missing true positives it can also be more easily catching true negatives in longer sentences, due to the same justification, the inability to accurately classify relations in those sentences. In the distantly supervised multi-instance learning module, the true negatives correspond to straightforward, smaller sentences, with less grammatical complexity. Overall the developed modules tend to classify candidate pairs more differently than equally, as we can see by Tables 4.3 and 4.5. Leveraging on this information it could be possible to integrate both systems. For instance, one could divide the training-set, train a distantly supervised multi-instance learning model with part of that training-set, then use the remaining of the training-set as the test-set, and use the positive candidate pairs identified to train a deep learning model, filtering the negative candidate pairs, and providing the deep learning model with a more accurate training-set.

Example 4.2 presents one sentence example for a relation correctly identified by both developed modules, that was not present in the knowledge base. This relation is one of the 25 relations that are not in the current HPO gold standard knowledge base of human phenotype-gene relations, due to the knowledge base being relatively recent, and is still in construction and updated frequently[6].

Example 4.2 Sentence example for a relation identified by the developed modules, that was not in the knowledge base.

- **Abstract Identifier:** 26701950
- **Sentence:** Germline mutations in KCNJ5 and CACNA1H cause FH-III and FH-IV, respectively, while germline mutations in **CACNA1D** cause the rare PASNA syndrome, featuring primary aldosteronism **seizures** and neurological abnormalities.
- **Gene:** *CACNA1D*
- **Gene Identifier:** 8912
- **Phenotype:** *seizures*
- **Phenotype Identifier:** HP_0001250

These systems can be used to effectively populate the HPO knowledge base or others within the same domain. Also, these approaches, can help reinforce or discard relations between human phenotypes and genes, and learn more about the origin of some phenotypes and their associated diseases. Ultimately, they can lead to the validation of the results of new research, and the proposal of new experimental hypotheses.

Future directions to outperform BioBERT could be adding to this model an ontological data representation layer. Table 4.2 already demonstrated the effectiveness of this additional layer to a deep learning module. The same impact could be achieved with a BioBERT plus ontology embeddings implementation. Further improvements could be achieved if we could also incorporate the distantly supervised multi-instance module and semantic similarity measures.

[6]last update was the *15/04/2019* release

Chapter 5

Conclusion

The main way we communicate scientific knowledge is through scientific literature. At the current rate of document growth, the only way to process this amount of information is by using computational methods. The information obtained through these methods can lead to a better understanding of biological systems. However, as most learning models require a large amount of training data, applying these learning algorithms to biomedical text mining is often unsuccessful due to the lack of training data in biomedical fields. This work made an important contribution to overcame this issue by creating a large and versatile silver standard corpus, the Phenotype-Gene Relations (PGR) corpus.

When creating biomedical text mining systems, it is essential to take into account the specific characteristics of biomedical literature. Biological information follows different nomenclatures and levels of complexity. The distantly supervised multi-instance and deep learning modules, developed in this work, were successfully built to accommodate the specificities of human phenotype and gene entities. Thus, this work accomplished the initial objectives (Section 1.2) with highly promising results, fulfilling the initial hypothesis (Section 1.2.1).

Following the growing tendency of systems targeting different biomedical relations, there is an increasing need for more domain-specific corpora, that can only be accomplished by automated corpus creation. The PGR corpus consists of 1712 abstracts, 5676 human phenotype annotations, 13835 gene annotations, and 4283 relations[1]. Using Named-Entity Recognition tools and a distantly supervised approach it was possible to effectively identify and extract human phenotype and gene entities and their relations. These results were partially evaluated by eight curators, obtaining a precision of 87.01%, with an inter-curator agreement of 87.58%. The PGR corpus was made publicly available to the research community.[2]

Automatic biomedical Relation Extraction (RE) still has a long way to go to achieve human-level performance scores. Over recent years, some innovative systems have successively achieved better results by making use of multiple knowledge sources and data representations. These systems not only

[1] Query 1, corresponds to the *10/12/2018* release of PGR

[2] https://github.com/lasigeBioTM/PGR

rely on the training data but make use of different language and entity related features, to create models that identify relations in highly heterogeneous text. Although, even with an optimal combination of features and the ideal features to perform biomedical Relation Extraction (RE) tasks are still far from human level performance. Nevertheless, the results achieved by the distantly supervised multi-instance and deep learning modules developed in this dissertation, were respectively, 73.48% and 55.00% in F-measure. These modules were able to detect new gold standard relations that were not present in the reference knowledge base.

This work showed that the knowledge encoded in biomedical ontologies and gold standard knowledge bases plays a vital part in the development of learning systems, providing semantic and ancestry information for entities, such as genes, proteins, phenotypes, and diseases. Also, it produced one freely available silver standard corpus of human phenotype-gene relations; a high-performance distantly supervised multi-instance learning module that can effectively extract human phenotype-gene relations from text; and one deep learning module with an ontological data representation layer (Section 1.4).

Integrating different knowledge sources instead of relying solely on the training data for creating classification models will allow us not only to find relevant information for a particular problem quicker, but also to validate the results of recent research, and propose new experimental hypotheses.

This work produced three publications including a book chapter about neural networks, a journal paper describing the ontologies applications to deep learning systems, and a conference paper describing the creation of the PGR corpus.

5.1 Future Work

For Chapter 3, future work can include manually correcting the human phenotype annotations that did not match any HPO identifier, with the potential of expanding the number of human phenotype annotations almost 2-fold and increasing the overall recall. Also, to expand the corpus by identifying more missed gene annotations using pattern matching, which is possible due to the approach being fully automated. Another possibility is the expansion of the test-set for more accurate capture of the variance in the corpus. For example, we can select a subset of annotated documents in which two curators could work to grasp the complexity of manually annotating some of these abstracts. Further, it is possible to use semantic similarity to validate the human phenotype-gene relations. Semantic similarity has been used to compare different types of biomedical entities [Lamurias and Couto, 2019a] and it is a measure of closeness based on their biological role. For example, if the *BRCA1* gene is semantically similar to the *BRAF* gene and the *BRCA1* has an established relation with the *tumor* phenotype, it could be possible to infer that *BRAF* gene also has a relation with the *tumor* phenotype, even if that is not evident by the training data. Finally, the effect of different NER systems applied to the pipeline should be studied.

Regarding the distantly supervised multi-instance learning module, the parameters of the miSVM package could be optimized using cross-validation on the PGR corpus, and different algorithms implemented (besides the sparse multi-instance learning (sMIL) algorithm [Bunescu and Mooney, 2007]).

For the deep learning module, it is possible to integrate the ontological information in different ways. For instance, one could consider only the relations between the ancestors with the highest information content (more relevant for the candidate pair they characterize). The information content could be inferred from the probability of each term in each ontology or resorting to an external data-set. Also, the already mentioned semantic similarity measurement could account for non-transitive relations (within the same ontology).

Future work may also consist in outperforming the BioBERT application by using their model along with a data representation layer of biomedical ontologies, given that this work already proved to improve the recall when comparing with an identical model that did not resort to ontological information.

Lastly, combining the techniques developed and presented throughout Chapters 3 and 4, it would be useful to develop a software tool in which we could annotate documents with human phenotype and gene entities and their relations. More than that, to employ and adapt these techniques to other combinations of biomedical entities to further expand our knowledge about biological systems.

References

Aho, A. V., Sethi, R., , and Ullman, J. D. (1986). *Compilers: Principles, techniques, & tools*. Addison-Wesley. 10

Alves, C. H. and Wijnholds, J. (2018). *AAV Gene Augmentation Therapy for CRB1-Associated Retinitis Pigmentosa*, pages 135–151. Springer New York, New York, NY. 12

Amores, J. (2013). Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81 – 105. 14

Aroyo, L. and Welty, C. A. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36:15–24. 20

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature genetics*, 25(1):25—29. 2, 18, 31

Bunescu, R., Mooney, R., Ramani, A., and Marcotte, E. (2006). Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from Medline. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 49–56, New York, New York. Association for Computational Linguistics. 11

Bunescu, R. C. and Mooney, R. J. (2007). Multiple instance learning for sparse positive bags. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 105–112, New York, NY, USA. ACM. 2, 14, 27, 44

Ciaramita, M. and Altun, Y. (2006). 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*, EMNLP '06, pages 594–602, Stroudsburg, PA, USA. Association for Computational Linguistics. 30

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46. 20

Couto, F. M. and Lamurias, A. (2018). MER: a shell script and annotation server for minimal named entity recognition and linking. *Journal of Cheminformatics*, 10(58):1–10. 4, 10, 22

Dai, H.-J., Wang, C.-K., Chang, N.-W., Huang, M.-S., Jonnagaddala, J., Wang, F.-D., and Hsu, W.-L. (2019). Statistical principle-based approach for recognizing and normalizing microRNAs described in scientific literature. *Database*, 2019. 36

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *arXiv e-prints*, page arXiv:1901.02860. 2

Dasigi, P., Ammar, W., Dyer, C., and Hovy, E. (2017). 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)*, pages 2089–2098, Vancouver, Canada. Association for Computational Linguistics. 19

De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., and Bruza, P. (2014). Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1819–1822, New York, NY, USA. ACM. 18

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 2, 17, 18, 36

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1-2):31–71. 14

Fellbaum, C. and Miller, G. (1998). *WordNet: An Electronic Lexical Database*. Language, speech, and communication. MIT Press. 17, 27

Fleuren, W. W. and Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods*, 74:97 – 106. 10

Giuliano, C., Lavelli, A., and Romano, L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *In Proc. EACL 2006*. 14

Goyal, P. and Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78 – 94. 19

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5:199–220. 18

Guresen, E. and Kayakutlu, G. (2011). Definition of artificial neural networks with comparison to other networks. *Procedia Computer Science*, 3:426 – 433. World Conference on Information Technology. 15

Hao, Y., Zhu, X., Huang, M., and Li, M. (2005). Discovering patterns to extract protein–protein interactions from the literature: Part II. *Bioinformatics*, 21(15):3294–3300. 13

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., and Steinbeck, C. (2015). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, 44(D1):D1214–D1219. 2, 27

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition. 15

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. 1

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580. 36

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780. 2, 16

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics. 17

Jiang, T., Liu, J., Lin, C.-Y., and Sui, Z. (2018). Revisiting distant supervision for relation extraction. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association. 14, 37

Kim, J., Kim, J.-j., and Lee, H. (2017). An analysis of disease-gene relationship from Medline abstracts by DigSee. *Scientific Reports*, 7:40154. 2

Kim, Y. (2008). Detection of gene interactions based on syntactic relations. *Journal of Biomedicine and Biotechnology*, 2008:1821 – 1827. 14

Koike, A., Niwa, Y., and Takagi, T. (2004). Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics*, 21(7):1227–1236. 13

Kong, X., Cao, B., and Yu, P. S. (2013). 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, pages 614–622, New York, NY, USA. ACM. 19

Köhler, S., Vasilevsky, N., Engelstad, M., Foster, E., et al. (2017). The human phenotype ontology. *2017 Nucleic Acids Research*. 2, 11, 18, 21, 31

Lamurias, A., Clarke, L. A., and Couto, F. M. (2017). Extracting microRNA-gene relations from biomedical literature using distant supervision. *PLOS ONE*, 12(3):1–20. 2, 3, 11, 14, 27

Lamurias, A. and Couto, F. M. (2019a). Semantic similarity definition. In *Encyclopedia of Bioinformatics and Computtational Biology*, volume 1, pages 870–876. Oxford: Elsevier. 44

Lamurias, A. and Couto, F. M. (2019b). Text mining for bioinformatics using biomedical literature. In *Encyclopedia of Bioinformatics and Computtational Biology*, volume 1, pages 602–611. Oxford: Elsevier. 1

Lamurias, A., Ferreira, J. D., and Couto, F. M. (2014). Identifying interactions between chemical entities in biomedical text. *Journal of integrative bioinformatics*, 11 3:247. 19

Lamurias, A., Sousa, D., Clarke, L. A., and Couto, F. M. (2019). BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies. *BMC Bioinformatics*, 20(1):10. 2, 4, 6, 18, 19, 20, 27

Leaman, R. and Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing*, pages 652–663. 21

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *arXiv e-prints*, page arXiv:1901.08746. 18, 20, 28

Li, Q., Li, T., and Chang, B. (2016). Learning word sense embeddings from word sense definitions. In Lin, C.-Y., Xue, N., Zhao, D., Huang, X., and Feng, Y., editors, *Natural Language Understanding and Intelligent Applications*, pages 224–235, Cham. Springer International Publishing. 19

Liu, H., Komandur, R., and Verspoor, K. (2011). From graphs to events: A subgraph matching approach for information extraction from biomedical text. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, pages 164–172, Stroudsburg, PA, USA. Association for Computational Linguistics. 13

Lobo, M., Lamurias, A., and Couto, F. M. (2017). Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, 7. 4, 10, 21

Luo, Y., Uzuner, Ö., and Szolovits, P. (2017). Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations. *Briefings in Bioinformatics*, 18(4):722–722. 2

Ma, N., Zheng, H.-T., and Xiao, X. (2017). An ontology-based latent semantic indexing approach using long short-term memory networks. In Chen, L., Jensen, C. S., Shahabi, C., Yang, X., and Lian, X., editors, *Web and Big Data*, pages 185–199, Cham. Springer International Publishing. 19

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA. 9

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford coreNLP natural language processing toolkit. In *ACL*. 10, 24

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). 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, pages 3111–3119, USA. Curran Associates Inc. 2, 17, 30

Min, B., Grishman, R., Wan, L., Wang, C., and Gondek, D. (2013). Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782, Atlanta, Georgia. Association for Computational Linguistics. 15

Miwa, M. and Bansal, M. (2016). 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)*, pages 1105–1116, Berlin, Germany. Association for Computational Linguistics. 16

Müller, H.-M., Kenny, E. E., and Sternberg, P. W. (2004). Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLOS Biology*, 2(11). 19

Nguyen, Q. L., Tikk, D., and Leser, U. (2010). Simple tricks for improving pattern-based information extraction from the biomedical literature. *Journal of Biomedical Semantics*, 1(1):9. 13

Papanikolaou, N., Pavlopoulos, G. A., Theodosiou, T., and Iliopoulos, I. (2015). Protein-protein interaction predictions using text mining methods. *Methods*, 74:47–53. 2

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. 28

Peng, N., Poon, H., Quirk, C., Toutanova, K., and Yih, W.-T. (2017). Cross-sentence n-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115. 18

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *CoRR*, abs/1802.05365. 2, 17

Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., and Ananiadou, S. (2013). Distributional semantics resources for biomedical text processing. In *Proceedings of LBM 2013*, pages 39–44. 30

Qiu, C., Wang, J., Lu, M., and Cui, Q. (2009). Transmir: a transcription factor–microrna regulation database. *Nucleic Acids Research*, 38(1):D119–D122. 27

Radford, A., Jeffrey, W., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *Pre-print*. 2

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *Pre-print*. 2

Riedel, S., Yao, L., and McCallum, A. (2010). Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin, Heidelberg. Springer Berlin Heidelberg. 15

Rinaldi, F., Schneider, G., Kaljurand, K., Hess, M., Andronis, C., Konstandi, O., and Persidis, A. (2007). Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artif. Intell. Med.*, 39(2):127–136. 13

Sahu, S. K. and Anand, A. (2018). Drug-drug interaction extraction from biomedical texts using long short-term memory network. *Journal of Biomedical Informatics*, 86:15 – 24. 38

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909. 9

Singhal, A., Simmons, M., and Lu, Z. (2016). Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine. In *PLoS Computational Biology*. 11

Song, L., Zhang, Y., Wang, Z., and Gildea, D. (2018). N-ary relation extraction using graph-state LSTM. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2226–2235, Brussels, Belgium. Association for Computational Linguistics. 18, 20

Sousa, D., Lamurias, A., and Couto, F. M. (2019a). A silver standard corpus of human phenotype-gene relations. *CoRR*, abs/1903.10728. 6

Sousa, D., Lamurias, A., and Couto, F. M. (2019b). Using neural networks for relation extraction from biomedical literature. *CoRR*, abs/1905.11391. 5

Surdeanu, M., Tibshirani, J., Nallapati, R., and Manning, C. D. (2012). Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 455–465, Stroudsburg, PA, USA. Association for Computational Linguistics. 29

Tripodi, I. J., Boguslav, M., Hailu, N., and Hunter, L. E. (2017). Knowledge-base-enriched relation extraction. In *Proceedings of the Sixth BioCreative Challenge Evaluation Workshop*. 19

Wang, H.-C., Chen, Y.-H., Kao, H.-Y., and Tsai, S.-J. (2011). Inference of transcriptional regulatory network by bootstrapping patterns. *Bioinformatics*, 27(10):1422–1428. 13

Wang, W., Yang, X., Yang, C., Guo, X., Zhang, X., and Wu, C. (2017). Dependency-based long short term memory network for drug-drug interaction extraction. *BMC Bioinformatics*, 18(16):578. 18, 20

Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen, L. J., and Brunak, S. (2018). A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLOS Computational Biology*, 14:1–16. 10

Xu, B., Shi, X., Zhao, Z., and Zheng, W. (2018). Leveraging biomedical resources in bi-lstm for drug-drug interaction extraction. *IEEE Access*, 6:33432–33439. 18, 20

Xu, Y., Hong, K., Tsujii, J., and Chang, E. I.-C. (2012). Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association*, 19(5):824–832. 13

Xu, Y., Mou, L., Li, G., Chen, Y., Peng, H., and Jin, Z. (2015). 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*, pages 1785–1794, Lisbon, Portugal. Association for Computational Linguistics. 17, 18, 34

Zhang, S., Zheng, D., Hu, X., and Yang, M. (2015). Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78, Shanghai, China. 16

Zhou, D., He, Y., and Kwoh, C. K. (2008). *From Biomedical Literature to Knowledge: Mining Protein-Protein Interactions*, pages 397–421. Springer Berlin Heidelberg, Berlin, Heidelberg. 12

Zweigenbaum, P., Demner-Fushman, D., Yu, H., and Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, 8(5):358–375. 11