UNED Escuela Internacional de Doctorado EIDUNED

TESIS DOCTORAL

2021

Biomedical Information Extraction: Exploring new entities and relationships

Hermenegildo Fabregat Marcos

M.Sc. in Computer Science and Technology

PROGRAMA DOCTORADO EN SISTEMAS INTELIGENTES

Dra. Lourdes Araujo Serna

Dr. Juan Martínez Romo

UNED Escuela Internacional de Doctorado EIDUNED

TESIS DOCTORAL

2021

Biomedical Information Extraction: Exploring new entities and relationships

Hermenegildo Fabregat Marcos

M.Sc. in Computer Science and Technology

PROGRAMA DOCTORADO EN SISTEMAS INTELIGENTES

Dra. Lourdes Araujo Serna

Dr. Juan Martínez Romo

Agradecimientos

Esta tesis pone punto y final a un capitulo de mi vida el cual no habría podido cerrar de no ser por el apoyo de las personas que han estado a mi lado este tiempo. Me gustaría aprovechar este espacio para darle a todos ellos las gracias.

A mis directores, Juan Martinez y Lourdes Araujo. Su cercanía e implicación conmigo y esta tesis hacen que me resulte difícil no estar agradecido. Llegué a la UNED pensando que anotaría un par de documentos y me marcharía, o que si hacia una estancia el mundo se me iba a caer encima. Hoy les agradezco su paciencia y los consejos que me han dado durante todo este tiempo. Sonrío al pensar que tenían razón, y me alegro de, al final y tras alguna discusión, haberles hecho caso. En definitiva, no puedo decir nada mas que hacer equipo con ellos ha sido un placer y una de las mejores decisiones que he tomado hasta la fecha. Gracias.

A mis compañeros de LSI. Quiero agradecer el haber encontrado en todos los miembros del departamento una extensión de la cercanía mostrada por mis directores. En especial, me gustaría agradecer la ayuda, el tiempo y el apoyo de Andrés Duque, Alvaro Rodrigo, Roberto Centeno y Raquel Martínez. Además, considero una suerte el haber coincidido con Jorge Carrillo, Mario Almagro, Javier Rodríguez, Agustín Delgado y Alicia Lara, entre otros. Por último, quiero agradecer a Julio Gonzalo la implicación en el desarrollo de las diferentes "charlas de los martes", las cuales han supuesto para mí una oportunidad de formación adicional.

A Ahmet Aker, quien ejerció como mi supervisor durante la estancia en la universidad de Duissburg-Essen. Gracias por la acogida, por la amabilidad y por la oportunidad de colaborar contigo y con Rassan Masood. Gracias a todos mis compañeros del departamento de *Informationssysteme*, Alfred Sliwa, Dr.-Ing. Norbert Fuhr, Ioannis Karatassis, Firas Sabbah y Vu Tran. Me fui a Alemania con nervios, y gracias a vosotros, volví a España deseando regresar algún día.

Por último, **a mi familia y amigos.** Llevo varios años fuera y nunca han dejado de apoyarme, y de recordarme que el único freno es uno mismo y que allá donde llegue, ellos estarán. Gracias.

Por el apoyo, por el consejo... en definitiva, por el tiempo. Gracias.
Hermenegildo Fabregat Marcos

Abstract

The different processes of digitization and dissemination of information that the society is currently experiencing have led to an increase of the available information, especially in the biomedical domain. Due to the effort required to process this volume of information, a research line that has been notably active in the last decade is the exploration of natural language processing and machine learning techniques for the extraction of information from unstructured documents. These techniques represent major milestones in the biomedical domain, especially in some information extraction tasks such as named entity recognition and relation extraction.

In this thesis we present a research focused on the automatic analysis of biomedical documents, deepening in the processing of documents about disabilities and functional impairments. These disorders have a significant impact on the social impact, since they affect to the daily life of a large part of the population, leading in some cases to serious limitations on the autonomy of the affected people. In addition, several rare diseases are associated with a wide range of disabilities, so they are frequently used to define them and they can represent very useful features for the diagnosis of these diseases, for which, and due to their nature, not much information is usually available.

The main objective of this thesis is the exploration of documents from the biomedical domain for the recognition of mentions to disabilities and the identification of their relationships with rare diseases. The processing of these entities involves specific difficulties, such as the lack of formal concretions for the definition of disability, and the wide range of ways to express the same disability.

In order to address this objective, it was necessary to collect and annotate different datasets, including documents written in different languages. After the generation of these resources, we proceeded with the exploration of entity recognition systems for the identification of mentions of rare diseases and disabilities, and with the study

of systems for the extraction of relationships between disabilities and rare diseases. Deepening in the analysis of these entities, we advanced on the exploration of the challenges for the generation of automatic systems oriented to the recognition of disabilities by proposing an evaluation task.

The different lessons learned during the evaluation task were used for the development and enhancement of an automatic system for disability recognition based on deep learning techniques. The developed system is based on the mixed use of different types of recurrent networks and it presented improvements over current state-of-the-art systems. At the same time, this system served as an initial architecture for the exploration of joint entity recognition and relation extraction systems. The study of the synergy between both tasks led to significant improvements.

Finally, in order to explore the effects of negation on information extraction systems, we analyzed several approaches for the automatic processing of negation in Spanish and English documents. During this analysis we examined the performance of proposals for the detection of negation triggers and their scopes, obtaining performance improvements over state-of-the-art proposals for the processing of Spanish documents. The results obtained for negation processing also led to interesting improvements on relation extraction and entity recognition.

Resumen

En la actualidad, los diferentes procesos de digitalización y difusión de información en los que está inmersa la sociedad han dado lugar a un incremento de la información disponible, sobre todo en el dominio biomédico. Debido al esfuerzo requerido para procesar tales cantidades de información, una línea de investigación notablemente activa en la última década es la exploración de técnicas de procesamiento de lenguaje natural y aprendizaje automático para la extracción de información de documentos no estructurados. Estas técnicas están suponiendo grandes hitos en el dominio biomédico, en especial en algunas tareas de extracción de información como el reconocimiento de entidades nombradas y la extracción de relaciones.

En esta tesis presentamos una investigación centrada en el análisis automático de documentos de este dominio, profundizando en el procesamiento de documentos acerca de discapacidades y limitaciones funcionales. Este tipo de patologías tienen un alto impacto social ya que afectan al día a día de una gran parte de la población, conllevando en algunos casos serios impedimentos sobre la autonomía de las personas afectadas. Además, muchas enfermedades raras tienen asociadas diversas discapacidades, por lo que frecuentemente se usan para caracterizarlas y pueden ser rasgos de gran utilidad en el diagnóstico de estas enfermedades, para las que por su naturaleza se suele contar con poca información.

El objetivo principal de esta tesis es la exploración de documentos del dominio biomédico para el reconocimiento de menciones a discapacidades y la identificación de sus relaciones con enfermedades raras. La detección de estas entidades presenta dificultades específicas, que van desde la falta de concreciones formales para la definición de discapacidad, hasta la necesidad de considerar el gran número de formas diferentes de expresar una misma discapacidad.

Con el fin de abordar este objetivo, resultó necesaria la recolección y anotación de diferentes colecciones de datos, incluyendo documentos en diferentes idiomas. Tras

la generación de las diferentes colecciones de datos, proseguimos con la exploración de sistemas de reconocimiento de entidades para la identificación de menciones a enfermedades raras y discapacidades, y con el estudio de sistemas para la extracción de relaciones entre discapacidades y enfermedades raras. Profundizando en el análisis de este tipo de entidades, extendimos la exploración de las dificultades para la generación de sistemas automáticos orientados al reconocimiento de discapacidades mediante la proposición de una tarea de evaluación.

Las diferentes lecciones aprendidas durante la tarea de evaluación propuesta nos sirvieron para el desarrollo y refinamiento de un sistema automático basado en *deep learning* para el reconocimiento de discapacidades. El sistema desarrollado se basó en el uso mixto de diferentes tipos de redes recurrentes y planteó mejoras sobre sistemas actuales del estado del arte. Al mismo tiempo, este sistema nos sirvió de base para la exploración de sistemas de reconocimiento de entidades y extracción de relaciones de forma conjunta. El estudio de la sinergia existente entre ambas tareas supuso la obtención de mejoras significativas.

Por ultimo y con el objetivo de explorar los efectos de la negación sobre sistemas de extracción de información, analizamos el rendimiento de enfoques para el procesamiento automático de la negación en documentos en español e inglés. Durante este análisis comprobamos el rendimiento de diferentes propuestas basadas en *deep learning* para la detección de disparadores de negación y sus alcances, obteniendo mejoras sobre propuestas del estado del arte para el procesamiento de documentos en español. Los resultados obtenidos durante el procesamiento de la negación supusieron además interesantes mejoras en la extracción de relaciones y en el reconocimiento de entidades.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Scope of the Thesis and Motivation

During the last decades, computer science has been acting as a catalyst for changes in many areas of the society. Currently, there are many quantifiable results derived from the rise of the Internet and the constant interest to achieve a better understanding of the information generated every day. Different sectors of society have turned automatic information processing studies into relevant actors. This fact is understandable considering that much of the information generated every day has an unstructured format. This sort of information presents different challenges for automatic processing using conventional analytical tools. Some of these challenges include the use of a free language and the lack of speech control.

This reality is particularly important in domains of high social interest, such as the biomedical domain, where we find examples of unstructured information in the analysis of documents such as research papers, among others. The social impact of the advances in the biomedical domain is considerable, so it is crucial to invest efforts in improving the state of the art. On the other hand, considering the digitization process experienced in different areas of the society, the biomedical domain is again of particular relevance due to the wide range of documents related to it. In summary, the number and variety of documents available in the biomedical domain make automatic information processing a crucial task. The application and natural language processing techniques (NLP) was found useful by different researches to reduce the gap between structured and unstructured information. Analyzing the literature about researches supported by NLP techniques in the biomedical domain, Figure 1-1 illustrates the trend of related publications. Although this figure represents a partial view of the total number of publications, it highlights the growing interest generated by this domain. As a result of these advances, the processes of information retrieval and knowledge inference improved, resulting in a better understanding of the challenges presented in the biomedical domain. This domain represents an interesting research field where the application of NLP techniques has facilitated important advances, in reducing the difficulties inherent to it. The identification of adverse reactions between drugs, the extraction of information for indexing electronic documents, among others, are some examples of applications where NLP techniques have provided important

advances.

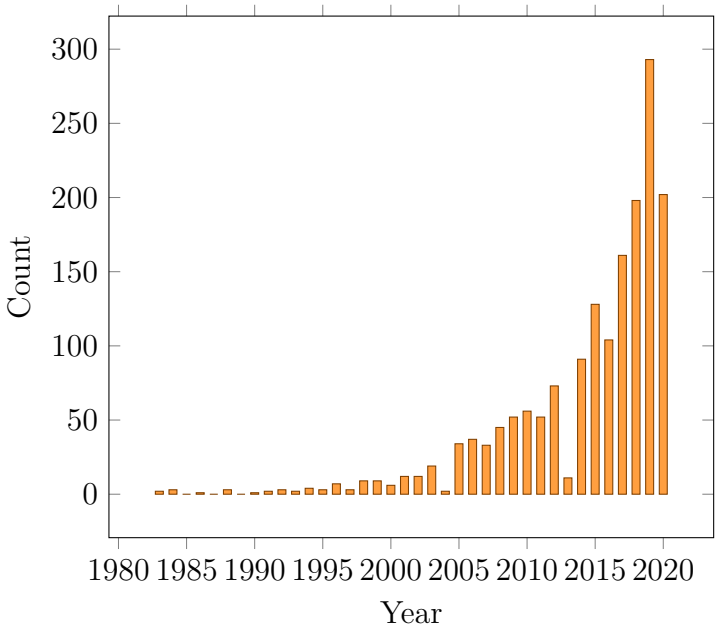


Figure 1-1: Distribution of documents retrieved from PUBMED related to the application of NLP techniques in the biomedical domain. PUBMED Search query: ((BIOMEDICAL) OR (MEDICAL) OR (HEALTH) OR (CLINICAL)) AND (NLP).

Although the current literature covers many challenges related to the biomedical domain, problems such as automatic processing of information about disabilities and rare diseases have not been addressed. The World Health Organization (WHO), an agency of the United Nations specialized in managing global health policies for prevention, promotion and intervention, in a report published in 2011, estimates that about 15% of the world's population suffers from some form of severe or moderate disability[1]. Works such as Schieppati et al. [141] comment on the diagnosis of disabilities in patients suffering from rare diseases. According to the European legislation, a disease or disorder is defined as "rare" when it affects fewer than 5 in

[1]World Health Organization report: `https://apps.who.int/iris/bitstream/handle/10665/70670/WHO_NMH_VIP_11.01_eng.pdf` - Last visited: October 10, 2020

2000 people[2]. The number of rare diseases that are registered in official agencies is huge. According to the Centers for Disease Control data, in 2013, 400 million people worldwide were living with a rare disease. In this thesis we will focus our attention on new challenges in the processing of biomedical documents related to disabilities and rare diseases using Information Extraction (IE) techniques.

Information extraction is a core task of NLP given the massive amount of information digitized that is presented in an unstructured format. IE means the processing of documents for the extraction of structured or partially structured information from them. To process heterogeneous information sources and to integrate the extracted facts into knowledge bases are crucial aspects in IE. Information extraction covers different NLP tasks, and under its umbrella, Named Entity Recognition (NER) and Relationships Extraction (RE) are paramount tasks.

In this work we will cover both the recognition of named entities and the extraction of relationships. These tasks are samples of information extraction processes with a great impact within this domain. Given the different applications that these tasks have in the biomedical domain, both are frequently visited by the literature. These processes have given the chance to analyze cause-effect relationships between entities such as drugs and disorders, and to improve the indexation of medical documents for the generation of reports and analytical studies. Although literature gathers many advances on this research area, the complexity that surrounds it, makes many aspects of it still not addressed. We will study different kinds of biomedical entities, although this research will specially focus on the detection of disabilities and the extraction of potential relationships between disabilities and rare diseases.

[2]Communication about rare diseases: `https://ec.europa.eu/health/non_communicable_diseases/rare_diseases_es` - Last visited: September 29, 2020

Named entity recognition and Relationship extraction

NER task involves the identification and categorization of the mentioned entities in a document. While identification of entities means the recognition of all the terms that comprise any given entity, the named entity categorization implies the identification of the semantic category for each previously recognized entity. The wide variety of entity types to consider and the ambiguity associated with this task, are, among others, reasons why this task is usually addressed for a small set of semantic categories simultaneously. In Example 1.1, a NER system focused on the identification of "companies" and "persons" would recognize the term "Ford Motor Company" as a reference to a company and the expression "Henry" + "Ford" as a reference to a person. Finally, Example 1.2 illustrates why each mention of an entity needs to be contextualized. Although a NER system may be supported by a knowledge base, this system must be able to contextualize the mention of "Alicante" as "Medical center" and not as "City".

Analogous to NER, relationship extraction (RE) consists of two processes: the identification of relationships between named entities and the semantic categorization of the extracted relationships. The relationships extracted from a document by a RE system can be expressed using triples, (Entity_X, link_statement, Entity_Y). In Example 1.1, a RE system would extract the relationship ("Henry", "is the founder", "Ford Motor Company").

Example 1.1: Henry Ford is the founder of Ford Motor Company.

Example 1.2: There is a medical center called Alicante in Fuenlabrada.

This research focuses mainly on the exploration of four research questions and on the study of the difficulties involved in processing specific kinds of biomedical entities.

Research Question 1. *Considering the difficulties inherent to the biomedical domain, do disabilities and rare diseases present additional difficulties for information extraction?*

Processing biomedical documents, involves a series of challenges to be addressed, including the use of a specific language and the unstructured format used to present the data. Within the inherent difficulties of this domain, Research Question 1 seeks to cover the analysis of the added difficulties involved in processing specific types of biomedical entities. Due to the requirement of annotated data inherent to the development and evaluation of some automatic methods, one objective of this thesis is the generation of collections of annotated documents with information about disabilities and rare diseases. Following the generation of these annotated resources, different approaches based on machine learning and NLP will be explored. In order to analyze under a wide scope the performance of these algorithms for identifying disabilities and rare diseases, additional types of biomedical entities will also be covered.

Research Question 2. *Which is the impact of changing the language on the performance of systems for automatic detection of biomedical entities?*

Transposing the above aspects to a multilingual scenario, Research Question 2 covers the use of resources and algorithms to analyze datasets in different languages.

Research Question 3. *Analyzing small datasets, how useful is the use of multitasking systems for entity recognition and relationship extraction?*

Studying in depth the core tasks of this thesis, and given the existing relationship between them, Research question 3 seeks to analyze potential reinforcement considering both tasks simultaneously. In order to explore a multi-objective approach for the recognition of named entities and the extraction of relationships, approaches based on lessons learned during the exploration of both tasks will be analyzed.

Research Question 4. *Analyzing transfer learning between different tasks, what is the effect of negation processing approaches on systems for relation extraction?*

Negation has significant effects on the speech which are interesting for the development of relation extraction systems. Research question 4 requires the study of a system for negation processing and the exploration of transfer learning mechanisms for studying the effects of negation on relation extraction processes.

In summary, the main purpose of this research is the study of the different challenges involved in the understanding of biomedical entities. Narrowing the research to the detection of named entities and the extraction of relationships, this thesis explores different state-of-the-art approaches to process these entities. In order to explore the proposed research questions, we compiled them under a main research objective supported by different partial objectives.

Main Research Objective To study named entity recognition and relationship extraction for the processing of documents related to disabilities and rare diseases. To analyze the performance of automatic systems dealing with documents of different languages and to study the effect of jointly exploring both entity detection and relation extraction. To explore the impact of linguistic aspects such as negation, on the extraction of relationships.

Objective 1: *Generation of a biomedical corpus with annotations of entities and relationships between them.*

Objective 2: *To perform an exhaustive analysis of state-of-the-art systems oriented to named entity recognition.*

Objective 3: *Exploration of the application of different natural language processing techniques, machine learning and, deep learning for the identification of disabilities.*

Objective 4: *Evaluation of the achieved performance processing other kinds of entities using the methods proposed for the detection of disabilities.*

Objective 5: *To perform an exhaustive evaluation and comparison of state-of-the-art systems oriented to relationship extraction.*

Objective 6: *To generate a system using state-of-the-art techniques for extracting relationships.*

Objective 7: *To analyze the performance obtained by a system that simultaneously deals with the detection of named entities as well as the extraction of relationships.*

Objective 8: *Exploration of techniques for detecting negation and its effects on the relationships explored during this thesis.*

1.2 Methodology

Several NLP tasks are addressed as part of this thesis. In this section we detail the methodological scheme followed in the development of this research.

1.2.1 Analysis of Previous Work

In Chapter 2, we analyzed some works published to date for each task covered by this thesis. All these NLP tasks have been frequently discussed in view of their importance in all kinds of information extraction processes. Since we propose a study on the processing of a kind of entities not covered by the literature to date, we analyzed several works on the processing of some similar entities such as drugs, diseases, etc. These entities share certain features inherent to the biomedical domain. This analysis allowed us to generate our own resources following an extended methodology and, at the same time, to explore the different proposed approaches. In summary, through a better understanding of the used techniques, the main objectives of this chapter are the identification of aspects to be improved and/or considered and the contextualization of the proposed approaches.

1.2.2 Sources of Information

Since this thesis covered the analysis of different NLP tasks, the use of different sources of information and corpora was necessary. For the study of entities such as disabilities and rare diseases, two collections of documents have been gathered and annotated (RDD and DIANN). We carried out this effort since there are no collections explicitly oriented to the study of these entities. In order to develop these corpora, we generated several annotation guidelines detailing the criteria used during the annotation. We validated these guidelines thanks to the support of expert medical personnel. To generate these corpora, we used the knowledge base developed and managed by Orphanet [157] to support the searching of documents. Orphanet is currently collecting information to improve the knowledge and visibility of disabilities associated with rare diseases in order to provide tools to help affected people. We obtained the documents using this knowledge base and the PUBMED database,

managed by the National Center for Biotechnology Information. As part of this thesis, we studied other kinds of biomedical entities and contexts using corpora provided by eHealth-KD 2019 [127] and MEDDOCAN [103] workshop organizers.

Regarding the study of negation processing, we supported our research using corpora in different languages (English and Spanish). For English, we explored this task using the BioScope corpus [153]. This resource contains different classes of biomedical documents. Finally, to explore the negation for the Spanish language, we used the SFU Review SP-NEG corpus [162] that contains online comments and opinions about different products.

1.2.3 Approaches

During this thesis, we studied different tasks frequently visited by the NLP research community. On the one hand, we explored the recognition of named entities using both machine learning and deep learning approaches, as well as architectures based on the exploration of terminologies and the generation of terminological variants. We detail the explored approaches on entity recognition in Chapter 4. On the other hand, we covered the study of the extraction of relationships using supervised approaches based on deep learning and machine learning techniques. We describe in detail the approaches explored during this thesis in Chapter 5.

Finally, we studied algorithms based on deep learning for negation processing. We first addressed the analysis of systems for the recognition of negation triggers in Spanish documents. Following this first study and considering documents in different languages, we extended the scope of the research by jointly analyzing the identification of negation triggers and negation scopes. We describe the details of the explored systems in Chapter 6.

1.2.4 Evaluation

Since there is no corpus with annotations related to disabilities and rare diseases, we evaluated the experiments carried out using the corpora developed by us (DIANN and RDD corpus). The absence of any previous results requires the definition of a well-defined evaluation framework around the developed resources. Although an

isolated study of our proposals provides us interesting insights and conclusions about their performance, it is necessary to put into context state-of-the-art approaches oriented to similar tasks and compare their performance with the proposed approaches. The generation of evaluation campaigns provides us with an exciting opportunity to analyze different proposals under a common evaluation framework at the same time. Additionally, participating in similar evaluation tasks can support our reached conclusions. Finally, the evaluation of the proposals oriented to negation processing was carried out considering different benchmarks available in the literature.

We used classic evaluation measures to analyze the performance of each proposed approach, e.g., *precision*, *recall*, and *f-measure*. Although all tasks have been evaluated using an exact match criterion, we considered an additional evaluation criterion based on partial matching to evaluate named entity recognition systems. Whereas strict evaluation criteria are used to measure the performance of a system in the faithful recognition of each component of a named entity, partial evaluation criteria are used to perform a more lenient evaluation, allowing the highlighting of difficulties in the treatment of certain entities. We detail both criteria in Chapter 3.

1.3 Structure of the Thesis

The structure of this document is as follows:

- Chapter 1, Introduction: This chapter details the NLP tasks under analysis in this thesis and the main problem we want to tackle. Some facets of the work to be carried out are also explained. Finally, different aspects of the methodology followed during the development of this thesis are detailed.
- Chapter 2, Related Work: The analysis of the most relevant works on the considered NLP tasks is detailed in this chapter. We also analyze the different aspects studied in the literature regarding the generation of biomedical annotated corpus.
- Chapter 3, Case Study: Disabilities and Rare Diseases: We analyze in this chapter the proposed case of study and related difficulties. Finally, developed corpora and the used annotation criteria are reviewed.

- Chapter 4, Named entity recognition: The work carried out on the detection of named entities is shown in this chapter. The efforts made on processing both disabilities and other types of named entities are presented here.

- Chapter 5, Relationship extraction: The main subject of analysis in this thesis associated with the extraction of relationships is covered in this chapter. We detail the work carried out in the detection of relationships between disabilities and rare diseases. Other related works are also detailed here.

- Chapter 6, Negation processing: We discuss in this chapter the research carried out on the processing of negation.

- Chapter 7, Conclusions and Future Work: A summary of the main contributions and future lines of research is detailed here. Finally, we list the papers published during the development of this thesis, where we discussed some reached conclusions.

Chapter 2

Related Work

Contents

In 1992, Carnegie Group developed one of the first commercialized information extraction systems oriented to the identification of facts from press news [6]. Nowadays, companies of all natures use information extraction techniques to process an endless number of types of information. As information has become more and more accessible, the changing needs of companies over the years have become more evident. Identifying all available information sources is an arduous task, and it is much more difficult to efficiently handle the different types texts. Research in information extraction techniques originates from the demand to summarize unstructured or partially structured textual information in simple databases or templates [158].

Contrary to the idea of a full-text understanding, the extraction of information is dependent on a concrete specification of the target to be analyzed. Among the core tasks of information extraction are both template filling and knowledge base population tasks. In summary, whereas template filling focuses on the generation of a template-based report, knowledge base population means a blind search of information inspired only by the classes of entities and relationships to be studied.

As a preamble to the work carried out, this chapter discusses the main research lines of this thesis, focusing the attention on two well-known knowledge base population tasks: the recognition of named entities and the extraction of relationships. First, in Sections 2.1 and 2.2, we briefly analyze the tasks of named entity recognition and relation extraction. During this analysis we try to focus the discussion on some items explored during this thesis. In addition, and narrowing the scope of this research, Sections 2.1.3 and 2.2.3 detail recent efforts oriented to the processing of biomedical domain information. Finally, and due to the implication of negation in information extraction processes, in Section 2.3 we analyze the research field of negation processing. In this section we focus the analysis on exploring the identification of negation triggers and scopes. Both tasks are considered essential in the processing and comprehension of the speech. Analyzing the approaches explored for negation processing, and being the negation a relevant aspect in the study of the biomedical domain, we discuss in Section 2.3.2 different advances in negation detection applied mainly to the processing of medical-clinical data.

2.1 Named Entity Recognition

In 1995, the sixth series of "Message Understanding Conferences" (MUC-6) was one of the firsts meetings focused on developing the task of named entity recognition (NER) [36, 63]. MUC-6 organizers defined this task as the identification of all persons, organizations and geographic locations mentioned in a given text. Nowadays, these three kinds of entities are considered a classic benchmark in the study of named entity recognition. The results obtained by the NER systems presented in MUC-6 motivated to propose new editions in which different challenges would be covered e.g. processing of other domains and languages (non-English languages). Example 2.1 presents a tagged text extracted from the corpus of the CoNLL-2002 (Conference on Computational Natural Language Learning) shared task, oriented to the recognition of named entities in a multilingual context (Spanish and Dutch) [146].

[**PER** Wolff] *, currently a journalist in* [**LOC** Argentina] *, played with* [**PER** Del Bosque] *in the final years of the seventies in* [**ORG** Real Madrid]

Example 2.1: Named entity recognition sample extracted from CoNLL-2002 Shared task dataset [146].

Challenges

Similar to other information extraction tasks, the recognition of named entities involves the consideration of some aspects that may lead to the consideration of non-typical entities (e.g. drug detection or adverse effects), to the processing of documents from different sources (e.g. clinical reports or scientific documents) and to the consideration of particularities inherent to the analyzed language. The study of systems for the recognition of named entities has sufficient incentives, being this situation a good opportunity for the proliferation of related shared tasks, which are an interesting research tool to analyze different points of view under the same umbrella.

Tasks proposed during the 2002 and 2003 editions of CoNLL are examples of the many NER tasks organized to date [146, 147]. Exploring other languages, GermanEval 2014 NER, collocated at KONVENS 2014 conference, addressed the recognition of entities in German and, IberLef (Iberian Languages Evaluation Forum) 2019 workshop hosted a shared task focused on studying the particularities of Portuguese [12, 34, 29].

In addition to the challenges associated with each language, the recognition of named entities also entails other difficulties more related to the domain to be processed. The biomedical domain requires to analyze some types of entities beyond people, geographical locations and organizations. Gurulingappa et al. [65] generated the ADE corpus oriented to the study of the extraction of reported side effects from drugs in medical reports. Following a similar approach, Herrero-Zazo et al. [70] developed the DDI corpus for the analysis of pharmacological substances and drug–drug interactions. Also, Dogan et al. [40] presented an annotated collection of documents extracted from PUBMED oriented to the identification of diseases in scientific publications. Other works that have explored the analysis of different types of entities can be found in workshops: MADE 1.0 2018 challenge organized to address the detection of medication, indications and adverse effects in English clinical notes and, PharmaCoNER 2019 shared task focused on a similar scope for the exploration of Spanish language documents [76, 61].

2.1.1 Evaluation

Precision, *recall* and *f-measure* (F1) are commonly used in information extraction to evaluate the performance of the approaches under study. These metrics are defined through the following concepts:

True Positive - TP An instance recognized by a system as positive and represented by the gold standard in the same way.

True Negative - TN An instance recognized by a system as negative and represented by the gold standard in the same way.

False positive - FP An instance recognized by a system as positive but represented by the gold standard as negative.

False negative - FN An instance recognized by a system as negative, but represented by the gold standard as positive.

Under the umbrella of these definitions, whereas $TP + TN$ represents the total of correctly recognized instances, $FP + FN$ represents the set of incorrectly classified instances. *Precision*, *recall* and *f-measure* can be defined as a relationship between these concepts.

$$Precision = \frac{TP}{TP + FP} \tag{2.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.3}$$

Precision models how many marked positives are really positive, whereas *recall* represents the number of recovered positives. *F-measure* is a harmonic mean between *precision* and *recall* representing an overall perspective.

Although the evaluation of entity recognition systems has certain parallels with tasks such as Part-of-Speech tagging, the main difference of this task comes from the consideration of "entity" as the minimum unit of interest, i.e. it is necessary to correctly identify the class to which each term belongs and its relationship with the rest of the terms. For example, a system that labeled "American" but not "American Airlines" as an organization would cause two errors, a false negative for "Airlines" and a false positive for "American".

Matching criteria

Given the vast number of ways to express the same entity and the need to correctly detect all the elements, Tzong-Han Tsai et al. [149] and Nejadgholi et al. [121] study different types of relaxation on the matching criteria. The study of different types of matching involves the redefinition of the concepts *True Positive*, *True Negative*,

False positive and *False negative*, opening the door to the analysis of the performance reached by one system using relaxed evaluation criteria. Analyzing the expression "Severe mental retardation" annotated in a dataset as "Disability", an exact matching evaluation would require identifying all the terms as part of the same entity. More relaxed criteria, such as core-term matching [53] or left matching, would allow us to analyze the performance of NER systems on a more relaxed way, considering as *True Positive* the annotation of "mental retardation".

2.1.2 Methods

Machine learning systems, with a special mention to novel deep learning systems, have been used very frequently due to their high performance in different areas. However, the use of terminology resources and rule-based systems are also paramount to understand the current state of the art. Analyzing the use of external resources, although the recognition of named entities can be understood as a list-based search, the limitations in the use of these lists were analyzed by different works from early stages of the NER task [109, 33, 87]. Mikheev et al. [109] studied the role of gazetteers in NER systems. Simple lookup systems based on the use of these lists suffered from generalization problems detected in MUC-6. Combinations of different types of lists (learned and common) was a great improvement over the individual use of them.

Rules-based systems

In the design of named entity recognition systems, the use of rules has been widely studied since its origins. Rule-based named entity detection systems can be defined as rewriting systems. These approaches show a reasonable performance in processing correctly written texts, especially if we take into account formatting aspects such as the presence of uppercase characters, which is very useful for the characterization of entities such as names, organizations or locations. Chieu and Ng [24] compared the performance obtained by a system trained with only uppercase texts versus another one trained with correctly written texts. They highlighted significant improvements by exploiting formatting aspects. Likewise, other works such as Kim and Woodland

[82] evidenced similar problems studying voice transcription processing. To avoid the cost of manually generating ad-hoc rules, this work reviewed the automatic generation of the said processing structures. The proposal of a hybrid system based on the use of lists and the consideration of different types of matching rules achieved the best results. However, although the performance of rule-based systems is often remarkable and the conclusions drawn from them are still applicable today, they suffer from a rather limited adaptability. More complex approaches based on machine learning and, in particular, deep learning techniques are often applied in the current NER state of the art.

Machine learning approaches

Concerning the machine learning based approaches, the use of Support Vector Machines (SVM), Hidden Markov Models (HMM) and Conditional Random Fields (CRF) stands out. Analyzing the original description provided by Cortes and Vapnik [30], a Support Vector Machine can be seen as a binary classifier ($Y = \{-1, +1\}$) based on the search of an optimal hyperplane that separates both classes by a maximal margin. Given its robustness, SVM has been used in many contexts. Although originally, Cortes and Vapnik [30] formulated this method using a binary classification problem, approaches such as "one against all" allow the use of these algorithms for multi-class classification. Although SVM does not have its own mechanisms to model a sequential problem, the literature includes SVM-based approaches and adaptations applied to NER tasks [78, 74]. These approaches label each term of a sentence through iterative calls to a classifier that uses the context information around each term to label them.

More focused on a strictly sequential processing, the use of probabilistic techniques such as HMMs and CRFs is very common. HMM is an algorithm based on Markov chains, which can be seen as a graph that models a set of states and where each state has a probability of transition to the other states. Analyzing the application of HMM to named entity recognition, we can understand the set of states as the classes to predict. The Viterbi algorithm is used to find the most probable sequence of hidden states. Leaman and Lu [93] apply a variant of HMM, Hidden Semi-Markov models (HSMM), to recognize named entities in biomedical documents. In contrast

to HMM, HSMM includes a temporal dimension to model the conditional probability of transition from one state to the other.

CRF was introduced in Lafferty et al. [88] and, although HMM and CRF are based on similar foundations, the representation capability offered by CRF allows the consideration of features beyond the use of observed words. Considering the mathematical basis of CRF, the probability of a set of labels ($y = [y_0, y_1, ..., y_j]$) given a set of instances ($x = [x_0, x_1, ..., x_j]$), can be defined as $P(y|x) = \frac{1}{Z(x)} \exp(\sum_i \sum_j \lambda_i f_i(y_j, y_{j-1}, x, j))$ where Z is a normalization factor and λ is the weight matrix associated with the feature function f. CRF training consists of adjusting the weight matrix to maximize the conditional log likelihood of annotated sequences for a dataset. The approaches based on HMM and CRF tend to perform better given their conception of the NER problem as a sequential classification problem where the probability of a NER tag has a strong relationship with the previously predicted tag and the current observed word.

Deep learning approaches

The more recent approaches and the advances made in parallel computing provide to the research community a good environment to carry out developments based on neural techniques. On the same principle, the vast amount of information created every day as well as their high availability, has generated the possibility of studying resources and tools such as Word Embeddings [110], allowing the analysis of texts at a semantic level beyond other traditional resources. Based on the study of the co-occurrence of terms in a dataset, a Word Embedding is a vectorial representation of textual information and, it is able to capture features on the similarity between terms, as well as other types of relationships. The vocabulary transposition through these vectorial resources and its consequent reduction of dimensionality has allowed significant advances in the use of neuronal systems. Developed by Mikolov et al. [110], Word2Vec is a very popular technique to generate Word Embeddings. CBOW and Skip-Gram are two well-known algorithms used to generate these vectors. These methods differ in the use of the context associated with each term. Whereas CBOW trains the model considering the context of each word for the prediction of it, Skip-

Gram uses the word to predict its context. In both scenarios, as a result of the training, we obtain a vectorial representation of each term. In addition to Word2Vec, techniques such as Glove [125] or FastText [14] for the transformation of text to vector spaces have emerged in recent years gaining a great importance. Whereas Glove follows a similar approach to Word2Vec by trying to represent the co-occurrence between a term and a context, FastText treats each term as a union of n-grams and studies the co-occurrence of them. Thanks to FastText, we obtain a richer representation of rare words and, the chance to model words not seen in the training corpus.

Word Embeddings appear in the literature in many state-of-the-art approaches of different NLP tasks. Analyzing the fundamentals of the named entity recognition task (entity identification and entity categorization), we can use Word Embeddings to study similarity relations between different expressions in order to perform an unsupervised annotation. Among others, this particular use of Word Embeddings has meant advances in the development of systems for languages with scarce annotated resources to generate supervised approaches which require a large amount of data [35].

The reduction of dimensionality provided by Word Embeddings has led to an ideal environment for the study of NLP neural systems. In the same line that approaches based on CRF, the current state of the art of NER gathers approaches of deep learning based on Long Short-Term Memory networks (LSTM) and hybrid applications of LSTM+CRF [89]. Introduced by Hochreiter and Schmidhuber [72], LSTM is a recurrent neuronal model that integrates self-connection or refreshing mechanisms that allow us to speak, on neuronal models, of sequential processing, in its strictest meaning. This kind of networks efficiently solve memory and computational cost problems related to other types of recurrent networks; LSTM offers a constant process time and complexity in every computational time-step. Since this kind of algorithms transfer the directional relationship derived from the input data itself (e.g. word order in a sentence), it is common to talk about bidirectional recurrent networks, where a sequence is processed in both directions, from the beginning to the end and vice versa (Figure 2-1).

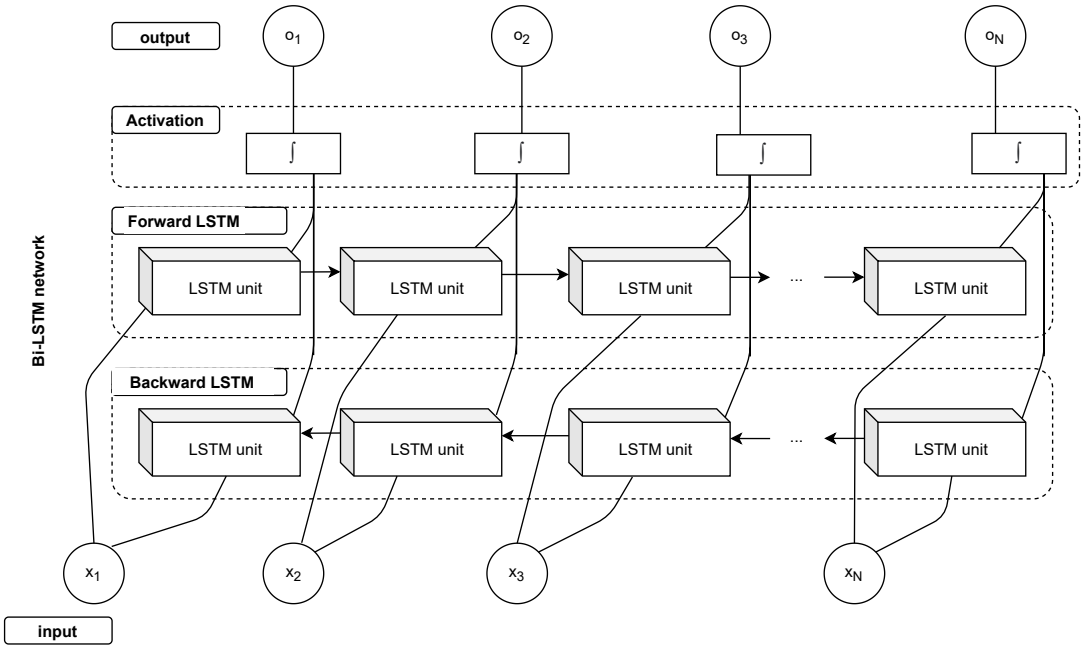


Figure 2-1: Bidirectional Long Short-Term Memory network.

Although LSTMs, in contrast to other recurrent models, can deal with very long contextual dependency relationships, they suffer from a vanishing of "knowledge" in specific contexts. In order to focus the acquisition of knowledge on certain aspects and to minimize the noise generated as a consequence of handling large data structures, different attention models have been recently developed. Although attention mechanisms have been frequently used for automatic translation, works such as Bharadwaj et al. [13] and Rei et al. [132], demonstrated the interest on this aspect in tasks such as NER. Recently and derived from the advances in the development of attention mechanisms, state-of-the-art approaches include the use of architectures based on encoder-decoder and a mechanism of attention. These architectures, known as Transformers, are the basis of language models such as Bidirectional Encoder Representations from Transformer (BERT) [37] and represent benefits over conventional sequential models by modeling more effectively long term dependencies among tokens in a temporal sequence. BERT uses encoders and language masking to reduce generalization problems caused by working with large data sets.

Although these types of architectures require high computational power and deal with high convergence times, they currently outperform state-of-the-art models in different NLP tasks, e.g. NER [66, 145] and relation extraction [160, 4]. We did not explore in depth this technology given the advanced stage of this thesis when transformer-based language models were becoming available to the community. During this thesis, we analyzed the model of attention proposed by Aker et al. [3] for the verification of rumors. The authors studied a model based on the analysis of recurrent terms to weigh the importance of textual information.

Labeling schemes for supervised systems

For a given sentence, the problem of named entity recognition may be seen as a sequential classification task, in which a NER system must assign a label to each word belonging to an entity, depending on the type and position of that word within the entity. In order to evaluate the complexity of this task at different levels and to cover different types of entities and contexts, different annotation schemes were proposed. The IOB (Inside-Outside-Begin) annotation scheme is a very common one that has served as a format to make available to the public a multitude of corpora [131]. Using three types of labels in its simplest version, this scheme tries to represent, in the context of NER's task, the role of each term in a given sentence. Example 2.2 illustrates the representation of a sentence using the IOB scheme. In this example, we can identify two entities (Fragile-X syndrome + mental retardation), as well as the role played by terms like *Fragile-X* and *mental* (Entity beginning) and others like *syndrome* and *retardation* (Part of an entity).

Sentence	**Fragile-X**	**syndrome**	**is**	**an**	**inherited**	**form**	**of**	**mental**	**retardation**	**with**	**a**
IOB Labels	B	I	O	O	O	O	O	B	I	O	O
Sentence	**connective**	**tissue**	**component**	**involving**	**mitral**	**valve**	**prolapse**	**.**			
IOB Labels	O	O	O	O	O	O	O	O			

Example 2.2: Sentence tagged using IOB scheme.

The IOB scheme entails a series of limitations to consider overlapped and nested entities. Some extensions of this scheme cover these aspects e.g. IOBES (also known as BILOU) consider the existence of nested entities by allowing to represent mono-term entities (S) and the end of multi-term entities (E) [150]. The S tag is also used to determine the start or end of nested entities. During this thesis, we have used IOB or IOBES depending on the problem under analysis.

2.1.3 NER applied to the Biomedical domain

The application of NER techniques in different medical-clinical is a well-known field of study. Depending on the requirements of annotated data, we can classify NER techniques into supervised and unsupervised approaches. Although the access and availability of some sources of information is guaranteed in the age of the Internet, the generation of annotated document collections is still a highly time-consuming process. Although the literature on entity recognition systems applied to the biomedical domain is rich, this domain suffers especially from the excessive costs associated with the generation of annotated resources. The small number of annotated collections in this domain is partially justified in view of the need for expert staff to cover the full spectrum of entities and their high degree of specificity. The NLP community has invested over the years much effort in the development of unsupervised systems to mitigate the cost of needing expert personnel in this domain. The complexity and specificity of the used language, suppose two challenges to consider in the development of systems oriented to this domain.

Unsupervised approaches

Due to the non-standardized use of medical terminology and the need to homogenize the language to facilitate the transmission of information, the development of terminological resources such as SNOMED CT [142] and metathesaurus such as UMLS (Unified Medical Language System) [100] aggregating different sources, has been crucial for tasks such as document indexing and named entity recognition. These resources have boosted both corpora annotation and system development oriented

to biomedical applications [81, 9]. Aronson [9] in the development of MetaMap, as well as Kersloot et al. [81] proposing cTakes, are some cases of information extraction systems for English using external resources and focused on medical documents. NER modules of the mentioned systems are based on searches in terminological dictionaries and on rule-based systems for the generation and matching of possible terminological variants (synonyms, abbreviations, expressions affected by modifiers, etc). Both cTakes and MetaMap use the set of categories mapped by the UMLS semantic network to narrow down the number of recognizable named entity types [104]. The hierarchical nature of this network allows a high degree of configuration when exploring the available types of entities. However, although the recall offered by UMLS is not negligible, the specificity degree reached by the described semantic types, in some contexts (e.g. differing between disabilities and diseases), may not be enough. During this thesis, variants of these systems have been used to study unsupervised methods for the detection of named disabilities (details in Section 4.2.1).

Supervised approaches

On the other hand, the state of the art about named entity recognition in the biomedical domain gather supervised approaches based on machine learning and deep learning techniques. The approaches presented under this umbrella can be divided according to how to address this problem is addressed. Named entity recognition can be approached through an entity search using external information, using a word-by-word classification process, or applying a sequential processing where the probability of assigning a label depends on previous steps. Studies such as the one presented by Kazama et al. [80] analyze the use of SVM for named entity recognition using a word-by-word classification approach. This kind of works consider a sentence as a series of instances that model every single word using contextual information. In order to correctly consider certain adjacency links, they modeled next to the information of the "k" term ($k > 0$), information corresponding to the $k - 1$ term. Requiring a modeling strictly dependent on the position of each term in a given sentence, this kind of approaches tend to suffer generalization problems. Other works such as Ju et al. [78] study this problem using the same approach but analyzing more

versatile features, which are more oriented to describe the form of a term and not the term itself, e.g. “the term contains symbols”, “the term contains numbers”, among others.

Analyzing approaches based on sequential processing of a sentence, we find machine learning approaches such as HMM and CRF, and others based on deep learning such as LSTM and variants of these networks. Due to the sequential nature of the NER task, this type of algorithm obtains improvements in comparison with other algorithms. The use of CRF for the detection of medical entities is analyzed by works such as Herwando et al. [71] and Kanimozhi and Manjula [79], where the authors make use of CRF to study the detection of medical entities such as symptoms, diseases, drugs, treatments, among others. This kind of algorithms, although it obtains quite remarkable results, requires a great effort of feature engineering.

Deep learning approaches

Recently, the state of the art of named entity recognition includes approaches based on deep learning algorithms. The different advances in computer acceleration have created an appropriate environment for the study of neural techniques. These algorithms demonstrate a great performance and make the study of features more flexible. In Cho and Lee [25], authors apply different word representations based on vectorial representations to compare the performance of CRFs with LSTM-based approaches. On the other hand, works such as Casillas et al. [20] and Li et al. Li et al. found improvements by analyzing different types of embeddings. Casillas et al. [20] evaluated the performance of LSTM+CRF approaches trained with a variant of CBOW called SkipNG. In SkipNG, the context vector is computed not by averaging over context Word Embeddings but as a weighted sum of individual Word Embeddings. The weight of each Word Embedding is assigned based on the position of each word. Li et al. [97] developed a neural joint model for entity recognition and relation extraction. Using a LSTM-based approach, they explored character embeddings as an auxiliary representation for the analysis of out-of-vocabulary words.

Considering the successful results obtained by state-of-the-art approaches, during this thesis we especially focused our experimentation on the analysis of systems based

on LSTM and CRF.

2.2 Relation Extraction

Given an ordered sequence of terms (S) where two or more entities are mentioned (E_x), the extraction of relationships is the task of predicting the existence of a relationship between a specific pair of entities ($f(S, E_x, E_y) = \{no\_relation, relationship\}$). In the seventh edition of the "Message Understanding Conferences" (MUC-7, Krupka and Hausman [87]), the task of relationship extraction (or Template Relation as it was defined in its beginnings) was mainly analyzed as a natural consequence of exploring the named entity recognition task during previous editions.

Example 2.3: Instance extracted from the corpus provided to participants of the MUC-7.

Extending the extraction of relationship to a multi-class classification problem, along with the extraction of relationships, we can define the sub-task of determining the type of existing relationships, e.g. the identification of three types of relationships was covered during the MUC-7: "LOCATION_OF", "EMPLOYEE_OF" and "PRODUCT_OF". Figure 2.3 shows an example extracted from the dataset designed for this meeting. Although the study of relations extraction techniques may involve more than two entities and multiple sentences, this thesis only covers binary and single line relationships.

Challenges

Similar to many information extraction tasks, the detection of relationships between named entities involves the consideration of both domain-specific knowledge and

language-related aspects. On the one hand, and extending the coverage of meeting such as MUC-7, during SemEval-2010 Task #8 [69] the organizers promoted the extraction of 9 types of semantic relations e.g. cause-effect and producer-consumer. This segmentation represents a major challenge given the possible need to use specific algorithms for each type of relationship. On the other hand, and similar to other information extraction tasks, each language has its own particularities to be taken into account. This is so, either because of the language's inherent peculiarities or because of the scarcity of its own resources. In the context of IberLef 2019 [34], Collovini et al. [29] proposed an evaluation task covering both the detection of entities and the extraction of relationships in Portuguese documents. They studied unspecific relationships between Organizations, Places and People.

Focusing the analysis on the biomedical domain, the ADE [65] and DDI [70] corpora are two well-known resources for the analysis of different types of relationships. Whereas the ADE corpus includes annotations on drug adverse effects, the DDI corpus includes annotations on drug-drug interactions. Furthermore and under the umbrella of BioNLP Open Shared Tasks 2019, works such as Bossy et al. [15] proposed an evaluation task for extracting relationships between microorganisms, phenotypes, etc. They explored both intra-sentence and inter-sentence relationships.

2.2.1 Evaluation

The evaluation of systems for relationship extraction is commonly approached by using *precision*, *recall* and *f-measure* (equations 2.1, 2.2 and 2.3). These metrics use the number of correct answers produced by the system (TP), the number of generated answers not included in the gold standard (FP) and the number of instances included in the gold standard but not identified by the system (FN). Given the need to evaluate systems in a real environment, the set of FP has an additional nuance by needing to consider relationships between entities not correctly recognized and consequently not included in the gold standard (spurious). In contrast to metrics such as Accuracy, measures such as *precision* and *recall* are preferred, since both focus on the return of TP, asking what percentage of correct answers the system has found and how many FP have also been returned by the system.

2.2.2 Methods

Although in the early days of this task, during MUC-7, the extraction of relations was mostly approached using manually generated rule systems [112, 55, 8], the task of relation extraction is nowadays conceived as a classification problem in which the range of approaches described in the literature covers both, machine learning and deep learning approaches.

Machine learning approaches

Analyzing machine learning based approaches, Zelenko et al. [167] and Rink and Harabagiu [134] exploited with great success systems based on SVM. Zelenko et al. [167] analyzed the use of low-dimensional representations versus feature-based algorithms. In this work the authors used kernels for the exploration of representations obtained by shallow parsing systems. They obtained interesting improvements especially when exploring relationships between very distant entities. On the other hand, and making an extensive use of explicit feature-based representations, Rink and Harabagiu [134] explored databases such as TextRunner [161] or NomLex-Plus [108], in addition to lexical and contextual features extracted from resources such as FrameNet [10] or PropBank [85]. This work was presented during SemEval-2010 Task #8 and it obtained state-of-the-art results using this combination of different linguistic resources.

Deep learning approaches

Moving the focus to other types of approaches and after SemEval-2010, approaches based on neural architectures obtained significant improvements for relation extraction by allowing to explore complex representation models with a great versatility. During this thesis, the following works were the basis of the experiments carried out analyzing deep learning techniques for relation extraction.

Wang et al. [156] proposed a convolutional architecture using an attention model for the analysis of the context between entities, i.e. processes that allow focusing the

learning process on specific aspects of the input. Their attention model allowed the identification of relevant tri-grams for each type of relationship e.g. by studying cause-effect relationships between e_1 and e_2 they identified templates such as "e_1 *caused a* e_2" or "e_2 *from* e_1". They developed a system using convolutional networks (CNNs), which, unlike other neural architectures, exploit hierarchical patterns found in the data to extract relevant patterns in a reduced space. Although CNNs have mostly visibility in image processing, different works have explored this type of networks with great success in NLP. Work such as Li and Mao [99] explored relation extraction using convolutional architectures enriched by embeddings of causation words. They generated these embeddings using two publicly available lexical databases, WordNet and FrameNet.

Extending the application of attention mechanisms, Lee et al. [94] proposed an architecture based on Bi-LSTM. They explored a self-attention model inspired by the multi-head attention formulation [151] for the extraction of relevant information from a given sentence, and the use of a second model of attention using a representation closely linked to entity recognition. Evaluating a relationship between two entities (e_1 and e_1) in a given sentence, they used an embedding-based representation to illustrate the relative distance of each term in reference to e_1 and e_2. Since information related to the type or class of each entity (e_1 and e_2) can serve to infer the type of relationship between them, they processed the embeddings of positions and approximate representations of the types of each entity obtained by clustering techniques. They obtained competitive results without using NLP tools or complex information about entities.

Exploring different types of relationships, works such as Lee et al. [94] found clear disconnections between some types of entities. They used attributes related to these entities (e.g. type or class) to support the correct segmentation of relationships. Given two types of entities, they studied the impact on the performance by explicitly indicating the existence or non-existence of relationships between these types of entities. Others works such as Gupta et al. [64] also explored complex representations to analyze latent information of the analyzed entities. In this work the authors investigated the effect on performance of a single joint approach for both entity recognition and relationship extraction. For a sentence of length n, they proposed a

table-filling method where they encoded in a $n(n+1)/2$ space the different objective functions. They placed the labels related to entity recognition in the (i, i) positions and the existing relationships between the w_i term and the w_x term $(x > i)$ in the (i, x) positions. Having modeled the output of the target function using this method, they trained a system based on recurrent networks and on the use of features based on Word Embeddings, part-of-speech, casing and on the study of the interdependence between entity types and relationship types.

2.2.3 Relation Extraction applied to the Biomedical domain

Analyzing approaches such as Giuliano et al. [56] and Bundschus et al. [18], we can highlight that the flexibility of machine learning approaches allows the exploration of a wide range of feature types. On the one hand, Giuliano et al. [56] explored the extraction of proteins and genic interactions using an SVM-based approach. They developed a linear combination of different kernels exploiting several context representations and different linguistic features such as part-of-speech or lemmas. Using this combination of kernels, the authors obtained interesting improvements in all the explored contexts. On the other hand and exploring a CRF-based approach, Bundschus et al. [18] analyzed complex representations extracted from entity recognition systems and incorporated information about different linguistic elements e.g., negation.

Exploring unsupervised approaches for relationship extraction, Quan et al. [130] proposed an approach for the identification of protein-protein interactions or gene-suicide association. They exhaustively analyzed approaches focused on the identification and processing of specific interaction expressions. They explored the performance of a rule system based on dependency parsing and on the analysis of interaction expressions extracted by pattern clustering. For the extraction of interaction words from unlabeled data they applied a clustering approach using the Polynomial Kernel method. Although the authors obtained interesting results using this unsupervised approach, they also explored a variant of this approach using a semi-supervised KNN model obtaining comparable results with supervised state-of-the-art systems.

Addressing the extraction of relationships using deep learning methods, Dewi et al. [38] and Li et al. [97] explored the extraction of drug-drug interactions and adverse

effect recognition. Although both papers used deep learning techniques for relationship extraction, Dewi et al. [38] explored biomedical Word Embedding on convolutional architectures of different sizes and, Li et al. [97] used LSTM-based architectures to support the complete modeling of dependencies between very distant entities or terms. In addition, the authors explored a joint approach for entity recognition and relationship extraction based on parameter sharing, i.e. although they proposed two different models, they explored the training of both models simultaneously by implementing a joint training strategy based on the use of the same set of features. Through this approach, they obtained improvements in both entity recognition and relation extraction.

2.3 Negation processing

Negation is a linguistic phenomenon of great impact in any language and its treatment is essential for the correct development of information extraction systems. According to its definition, this language modulator has the ability to transform a positive message into a negative one, transforming its meaning and implications. Considering negation in information extraction processes is crucial for multitude of NLP tasks, e.g. for medical report processing, negation is important to understand the absence of symptoms or negative results of tests and procedures. The standard two phases of a negation processing system are: identifying the presence of negation markers and determining their scope. According to the classification proposed by Givón [57], negation can be classified according to the type of trigger used: Morphological, where an expression is negated through the use of affixes i.e. i- in illegal; or Syntactic where an expression is negated through the use of a set of words or phrases i.e. no ("no/not"), never, etc. In syntactic negations, negation triggers act as operators on a set of terms and expressions named *scope*, i.e. the most common negation trigger in English is "not" together with its contractions [148]. Since a negation cue does not always act as a trigger, (e.g. in the sentence "You bought the car to use it, didn't you?" the cue "not" is not used as a negation but it is used to reinforce the first part of the sentence) the processing of the negation is not as easy as performing a simple lookup of the most common negation triggers. On the other hand, the fact that a

sentence contains a negation does not necessarily mean that it affects all the elements of the sentence. For this reason, the correct identification of the scope of the negation is a field of study of great importance.

Analyzing the set of resources available for the study of negation, we found interest in different domains and languages. In the biomedical domain, the BioScope corpus is one of the collections that has given most support to the study of negation. This corpus contains medical articles, abstracts and medical reports. In other domains, corpora such as SFU Product Review Corpus or ConanDoyle-NEG represent interesting advances. Whereas SFU Product Review Corpus contains annotated online product reviews, ConanDoyle-NEG collects annotated texts extracted from 4 stories of the adventures of Sherlock Holmes, written by Conan Doyle. Analyzing the complexities of each resource, we found the use of a specific vocabulary, the lack of speech control and the use of complex negation structures. On the other hand, although many studies focused on the coverage of English documents, other works such as CNESP (Chinese Negation and Speculation Corpus) or SFU ReviewSP-NEG facilitated the study of negation in different languages [170, 162].

2.3.1 Evaluation

The tasks of negation trigger detection and scope recognition are commonly understood as sequence labeling tasks and given their similarities with the task of named entity recognition, the methods of evaluation and annotation discussed in Section 2.1, are useful for the evaluation of systems for the detection of triggers and scopes. Analyzing specific evaluation schemes for both tasks, Morante and Blanco [115] proposed different strategies in the exploration of the systems presented during *SEM 2012 Shared Task. In this scheme they studied evaluation methods at scope/cue level and at token level. For the evaluation of scopes, the strict version of the proposed metrics requires a complete identification of the triggers. Exploring the metrics *precision*, *recall* and *f-measure*, they understand as True positive, an exact match between the gold standard and the segments identified as scope and/or negation triggers; and False positive, those segments of the sentence identified as trigger or scope and not matching with any instance of the gold standard, neither exactly nor partially; False negative, those instances of the gold standard that were not identified, neither exactly

nor partially.

2.3.2 Methods

With the aim of processing clinical records, the study of negation processing systems has been closely linked to the biomedical domain since its beginnings [23]. Exploring the literature on approaches for negation detection, we can distinguish three types of approaches: rule-based, machine learning and deep learning.

Rules-based approaches

Chapman et al. [23] presented an algorithm called NegEx based on the use of regular expressions for the detection of negation in discharge Summaries. Although the performance of NegEx in this context was very interesting, Mitchell et al. [113] evaluated NegEx in the annotation of Pathology reports and found inconsistent performance processing different sections. Whereas regular expressions specified in NegEx performed better in sections with simpler linguistic constructs, such as 'Final diagnosis", analyzing more complex sections, such as "Microscopic Description" or "Comments", NegEx performed significantly worse. Nevertheless, NegEx is considered a baseline in many of the works dealing with the automatic study of negation and, tools such as cTAKES [140] or DEEPEN [107], designed for processing medical documents in free text format, use NegEx for the treatment of negation. In addition, although NegEx has been designed for English, it was adapted to different languages such as Swedish, French, German [22] and Spanish [31].

Using a similar philosophy, Mutalik et al. [120] developed NegFinder, a modular system for the recognition of negated patterns in medical documents. This tool consists of a lexical scanner that uses regular expressions to generate a finite state machine and a parser to evaluate the occurrence of UMLS concepts, negation signals and terminators and sentence terminators.

On the other hand, whereas the identification of negation triggers was a recurrent aspect in the study of negation, scope recognition was a more neglected aspect, e.g. NegEx proposes a very simple definition of scope, limiting its size to a maximum of 6

terms. Representing a significant breakthrough in the study of the negation scope, Harkema et al. [67] developed ConText which established that the scope ends with a termination term or at the end of the sentence, regardless of the number of terms. The approach used by ConText worked well for identifying negated, hypothetical and non-patient experiences in different report types.

Machine learning approaches

Although NegEx algorithm shows a high *performance*, an important issue to take into account, is the low *precision* obtained evaluating sentences where the term "no" appears. Goldin and Chapman [60] extend the study of "no" cases exploring two machine learning algorithms (Naive Bayes and Decision Trees). Comparing the performance of these proposals against NegEx, both models showed significant improvements in terms of precision analyzing "no" scenarios.

Goryachev et al. [62] proposed the comparison of the performance of 4 methods for negation detection in outpatient notes: two based on regular expressions and two based on classification algorithms (SVM and Naive Bayes) trained with discharge reports. They trained the classification systems using discharge reports instead of outpatient notes, in order to obtain a "realistic" view of the performance of these systems. Although the algorithms based on regular expressions showed better performance, the authors indicated that the machine learning methods reached a high performance training and testing using only outpatient notes.

Exploring the idea of automatically extracting patterns for negation detection, Rokach et al. [136] analyzed the performance of decision trees trained from regular expressions generated using algorithms such as Teiresias [133]. They compared the results obtained using this tool with those obtained using fixed regular expression schemes (NegEx) and other machine learning algorithms such as CRF and HMM. They obtained significant improvements especially in comparison with the performance demonstrated by HMM.

Analyzing the BioScope corpus [153], Morante and Daelemans [116] developed a multi-step system for the detection of negation triggers and scopes. For the first phase, they used a decision tree to label each term with IOB labels indicating if a

token is at the beginning, inside or outside of a negation signal. In the second phase, they trained a CRF using the output generated by three different classifiers (SVM, k-nearest neighbor classifier and a CRF).

For different domains and dealing with the detection of negation triggers and the recognition of its scope jointly, Li and Lu [98] carried out several experiments using different kinds of conditional random fields (CRF), *linear CRF* [88], *semi-CRF* [139] and *latent variable CRF*. Considering the obtained results, the authors reached a shared opinion [116, 43, 91] regarding the good performance of CRF-based systems in sequence labeling tasks, having obtained remarkable results even after extending the evaluation to languages such as Chinese.

Deep learning approaches

With the application of deep learning algorithms for negation scope detection, works such as Lazib et al. [92] and Fancellu et al. [51] studied the application of different deep learning architectures. Although they analyzed different datasets (Lazib et al. [92]: SFU Review, and Fancellu et al. [51]: ConanDoyle-NEG corpus), both explored models based on bi-LSTM versus others based on simple feed-forward neural networks or CRF. They highlighted the good performance of Bi-LSTM based systems, especially considering the exact matching. Focusing the analysis on specific cases, Fancellu et al. [51] categorized different negated sentences according to the negation trigger (lexical, multi-word, unseen during training, etc.) and evaluated the performance of different configurations of Bi-LSTM-based models. They obtained interesting results processing negations based on multi-term negation triggers, e.g. “by no means of” and “no longer”. Fancellu et al. [52] extended the study to other domains and languages (Chinese), presenting, among others, results for the BioScope corpus and for the SFU corpus [86]. Evaluating the complexity of processing some types of negations, they found it easy to process negations delimited by punctuation signs, being this type of negation predominant in corpora such as BioScope or SFU.

Moving the focus to other types of deep learning techniques, Qian et al. [129] explored the use of CNN for negation and speculation scope recognition on the BioScope corpus. They trained systems using information extracted from the dependency tree

parse and, although they obtained very competitive results processing abstracts and clinical records, among their conclusions they highlighted the difficulty of processing long-distance syntactic dependence.

2.4 Conclusions

Named Entity Recognition and Relation Extraction are two of the main tasks in the research of information extraction techniques. Both tasks are closely related to linguistic aspects such as negation. Currently, although the analysis of negation and these information extraction tasks are significantly different, they share the good performance of deep learning based systems. Whereas, for entity recognition and negation processing approaches based on LSTM+CRF previal, approaches based on convolutional techniques have a better performance for relation extraction. At the same time, the consideration of Word Embeddings allows the modeling of certain semantic links. The high potential of this tool has made its use a welcomed practice. During this thesis we will explore for each task a set of approaches based on different methods but focusing on the study of supervised deep learning approaches.

Chapter 3

Case Study: Disabilities and rare diseases

Contents

This chapter details the main case of study analyzed by this thesis and provides a historical overview of different frameworks covering the study of disabilities and rare diseases. We discuss the relevance of studies about disabilities and we summarize the complexity of the formalization of disabilities as an entity type to be analyzed. In order to analyze the performance of approaches focused on the processing of the topics covered by this thesis, this chapter also discusses the collections of documents that we gathered and annotated. We also include an analysis of the methodology and the annotation criteria applied for the generation of these collections.

3.1 Motivation

International organizations, such as WHO, alert about the functional difficulties and disabilities experienced by a large part of the world's population. Although the study of disabilities is not a pluralistic topic in the society, a large part of the population is aware of the existence of a large number of disabilities. The United States CDC, through the Disability and Health Data System (DHDS) and the Behavioral Risk Factor Surveillance System (BRFSS), estimated that in 2018 about 26% of the U.S. population had some form of disability (Figure 3-1). The integration problems that affected people suffer in their daily lives are undeniable. According to data provided by the European Union through the portal Eurostat, in 2012 it was estimated that 37% of the 70 million people living in the European Union affected by some kind of disability (over 15 years) had some need for assistance to mitigate the effect of this condition in their daily lives[1]. As a result of this situation, many entities and organizations adopted integration plans in their agendas[2]. The DHDS analytical tool is in line with the efforts being made by different organizations to achieve a better understanding of the impact of these disorders on the society.

[1]eurostat - Statistics explained: Disability statistics - need for assistance `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Archive:Disability_statistics_-_need_for_assistance#People_with_disabilities_requiring_assistance` Accessed (October 10th, 2020).

[2]`https://www.un.org/en/content/disabilitystrategy/` Accessed (October 10th 2020).

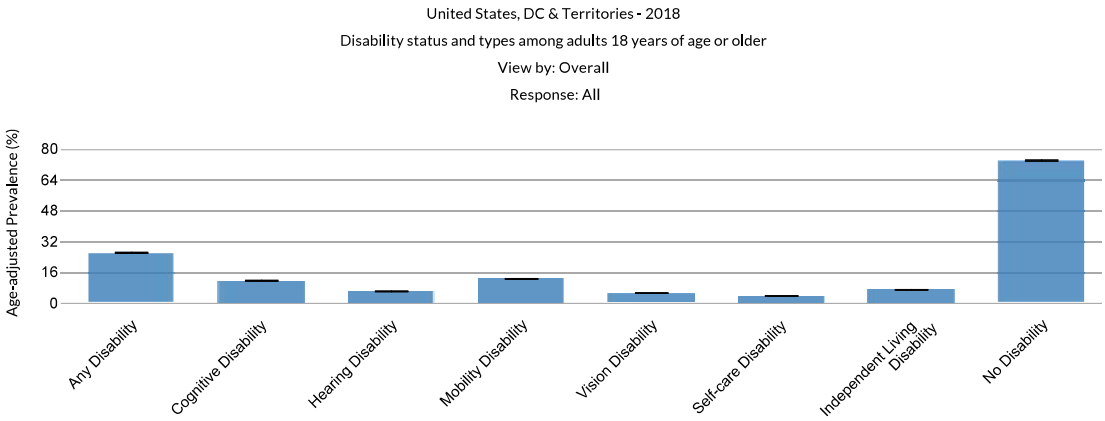


Figure 3-1: Estimation about disability status and types among adults 18 years of age or older. Data source: Behavioral Risk Factor Surveillance System (BRFSS)

Furthermore, many countries include their own legislation to regulate this context and to grant rights and instruments of protection and assistance to people affected by this condition. Considering that all disabilities do not imply the same type of functional impairment, these legal frameworks include the basis of the criteria used by the administrations to regulate the eligibility for subsidies. In Spain, the Ministry of Labor and Social Affairs, by means of *Royal Decree 1971/1999, of December 23, 1999*, on the procedure for the recognition, declaration and qualification of the degree of disability, recognizes up to 5 types of disability depending on the degree of impairment and 13 types depending on the physical/functional aspect affected[3].

Although the efforts made to date on this problem are remarkable, due to the similarities that this condition has with several diseases and/or pathologies, it is difficult to understand the nature of the disabilities and the impact they have on the society. The automatic identification of disabilities and the processing of possible relationships with other medical disorders have been rarely covered. These tasks involve difficulties inherent to the biomedical domain as well as other added challenges.

[3] `https://www.boe.es/eli/es/rd/1999/12/23/1971/` Accessed (October 10th, 2020).

The wide range of ways to express the same disability or the need to identify different aspects of the listed disabilities (e.g. severity or duration) are examples of these challenges. Well-known applications such as MetaMap [9] and cTakes [81] are proposed to detect entities in the biomedical domain, but they don't consider a disability such as a specific semantic type. These tools use the semantic classification provided by UMLS to categorize the identified entities [104]. Analyzing the hierarchical scheme of UMLS classes, there are semantic classes such as "Finding" or "Sign or Symptom" which, by their definition, could fit into a definition of disability.

Finding All that is discovered by direct observation or measurement of an organism attribute or condition, including the clinical history of the patient. The history of the presence of a disease is a 'Finding' and is distinguished from the disease itself.

Sign or Symptom An observable manifestation of a disease or condition based on clinical judgment, or a manifestation of a disease or condition which is experienced by the patient and reported as a subjective observation.

Although these definitions cover identifiable dimensions of disabilities, they also cover concepts unrelated to disabilities, e.g. "deaf-mute" and "body odor" are reported indistinctly in UMLS as "Finding". In short, these tools do not have an ontology that considers disabilities as an independent semantic class. In order to understand the problems derived from the formalization of this new semantic class, it is necessary to review the different frameworks that define the concept of disability.

Disability definition

In 1980, the World Hearth Organization, with the publication of the International Classification of Impairments, Disabilities, and Handicaps (ICIDH), a manual for the classification of the consequences of an illness, defined the term disability as "In the context of health experience, a disability is any restriction or lack (resulting from an impairment) of ability to perform an activity in the manner or within the range considered normal for a human being" [123]. In this context an impairment is

understood as the experience of a loss or abnormality of anatomical, functional or psychological abilities. The ICIDH model of disease consequences can be seen as a preliminary step towards the development of a theory of disability.

Although the conceptual framework established by the ICIDH represented a great advance, this model has different points of criticism, both theoretical and practical, especially due to the overlapping between the definitions of the macro concepts. Given the lack of specificity of these definitions and the need to express the concept of disability under an umbrella more focused on the daily life of the affected people, in 2001, the WHO redefined through the International Classification of Functioning, Disability and Health (ICF) the concept of disability as a relationship between health conditions and contextual factors, where one or more functioning factors are involved [124]. The functioning factors included in this document are classified as Body Functions and Structure, Activity and Participation. ICF tries, among other aspects, to rewrite the scope of the definition provided by the ICIDH in order to improve the focus on the target of the different social benefits provided by several institutions to reduce the impact on the daily life of people suffering this kind of pathology.

Relationships with diseases

The comorbidity associated with several disabilities and diseases, occurring simultaneously or over a lifespan, is a frequently visited topic in the literature [119, 7, 154]. Narrowing the study to specific types of diseases, rare diseases are a particular type of disease characterized by its degree of incidence in the population. In Europe, a disease is considered to be rare when it affects 1 person per 2000. Nowadays, there are registered more than 7000 kinds of rare diseases. The severity of these diseases is very diverse and the limited knowledge available about it causes feelings of helplessness in the affected people.

Orphanet[4], an organization established in France in 1997 and currently constituted as a consortium of 41 countries, both within and outside Europe, aims to promote the

[4]Orphanet: an online database of rare diseases and orphan drugs. Copyright, INSERM 1997. Available at http://www.orpha.net Accessed (November 8th, 2020).

improvement of diagnosis, care and treatment of patients affected by these diseases through the gathering and dissemination of related information. Using an adaption of ICF [122], designed to promote the study of the developing of child and the influence of its surrounding environment, Orphanet provides information on the functional consequences derived from the diagnosis of a rare disease. This organization provides information on the degree of occurrence of different types of disabilities in patients affected by a rare disease. They also include information about the severity and temporality of these limitations according to their functional impact and their duration in time. Table 3.1 shows a set of very frequent functional limitations observed in patients affected by Angelman Syndrome[5]. Although this disease has a low incidence, about one case per 15,000 births [143], it is also related to a large number of serious functional consequences.

Ability	**Temporality**	**Severity**
Acquiring language	Permanent limitation	Severe
Learning to read	Permanent limitation	Severe
Learning to write	Permanent limitation	Severe
Learning to calculate	Permanent limitation	Severe
Reading	Permanent limitation	Severe
Writing	Permanent limitation	Severe
Calculating	Permanent limitation	Severe
Focusing attention	Permanent limitation	Severe
Maintaining head position	Acquisition delay	Low

Table 3.1: Functional limitations very frequent in patients affected by Angelman syndrome.

[5]Complete list available at: `https://www.orpha.net/consor/cgi-bin/Disease_Disability.php?lng=EN&data_id=90` Accessed (October 10th, 2020)

3.2 Corpora generation

At present there are no automatic tools that contemplate disabilities as a specific object of analysis. Likewise, annotated textual sources available in the literature do not include special annotations for these disorders either. Therefore, in order to cover the automatic detection of disabilities in scientific-medical texts and the identification of relationships between disabilities and rare diseases, we compiled and annotated two collections of documents as part of this thesis: RDD (Rare Diseases and Disabilities) and DIANN (Disability annotation on documents from the biomedical domain) corpus. Each collection tries to cover specific aspects of information extraction in biomedical documents.

3.2.1 Methodology

The generation of these collections was carried out by three persons with previous knowledge in the retrieval and annotation of documents and assisted by expert medical personnel. In order to ensure the consistency of the generated collections, we applied an iterative approach, where the developed annotation guidelines were refined in each iteration. Figure 3-2 shows the iterative approach used for the generation of the corpora. This approach was used in the development of similar collections [135].

We used the definition provided by ICF for the term disability and the support of the consulted medical doctors to draw an initial definition of the disability concept, restricting its scope initially to a small set of examples. Using this draft as a common definition framework, we collected a small set of documents using references provided by Orphanet or directly consulting repositories such as PUBMED. Among the considered documents, we selected only those that mention one or more disabilities. Each recovered document was annotated by the three annotators and the compiled knowledge was dumped on the draft of the annotation guidelines. In addition, we annotated some linguistic aspects such as negation and speculation when these affect one or more disabilities. To annotate these linguistic phenomena, we used the criteria published by Vincze et al. [153].

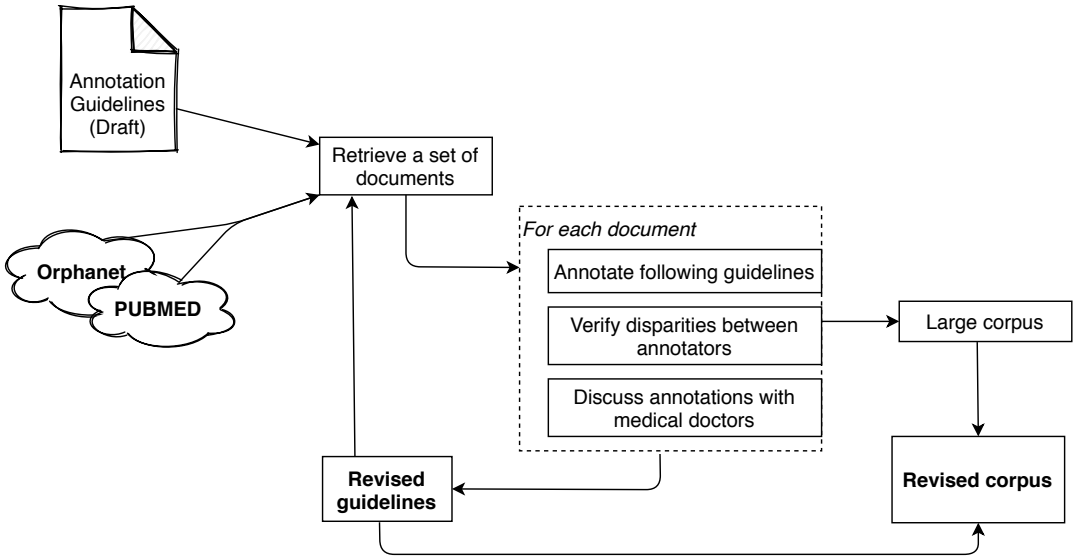


Figure 3-2: Iterative process to generate the annotation guidelines.

After each update of the annotation guidelines, we reviewed the annotations made in previous stages. Differences between annotations generated by each annotator were solved by consulting the medical doctors. The number of documents retrieved and annotated in each iteration was incremental, so that in the first stages the number of documents annotated per iteration was smaller than the number of documents annotated in subsequent iterations. Since after each iteration the developed annotation guidelines addressed more clearly and correctly different aspects of the annotation process, in advanced stages, the verification of the previously annotated documents did not require a great effort.

3.2.2 Annotation guidelines

As a result of the application of the methodology described above, we developed a set of annotation guidelines that cover all aspects involved in the generation of both corpora.

Annotating disabilities: To annotate disabilities we considered that they are described either by a term directly related to a disability, or by a human function

that is absent or limited. In both cases we considered that the disabilities imply a condition that, either because of its duration or because of its severity, implies a disturbance in the normal development of daily life. Table 3.2 provides a list of specific terms that we considered as disabilities. This table includes disabilities in English and their translation into Spanish. We included in this table only one word representing the disability and not all its possible derivative words.

achromatopsia / acromatopsia	dementia / demencia	paraparesis / paraparesis
aphasia / afasia	diplegia / diplegia	paraplegia / paraplegia
apraxia / apraxia	dysarthria / disartria	paresis / paresia
ataxia / ataxia	dysautonomia / disautonomia	quadriplegia / cuadriplegia
autism / autismo	dyskinesia / discinesia	tetraplegia / tetraplegia
blindness / ceguera	hemiparesis / hemiparesia	
deafness / sordera	hyperactivity / hiperactividad	
deaf-mutism / sordo-mudez	paralysis / paralisis	

Table 3.2: Excerpt of specific expressions that refer to a disability, as well as their translation into Spanish.

On the other hand, the definition of disability initially proposed by the ICIDH invites us to consider a disability as the limitation or absence of a physical or mental function. For annotating expressions that represent human functions whose limitation can entail a disability we considered variants of the functions covered by the Orphanet Functioning Thesaurus. This thesaurus is based on ICF and gathers a wide list of physical, mental and intellectual functions. For example, we found the expression “ability to recognize faces”, which can be considered similar to “interacting with other people”, and which is included in the Orphanet Thesaurus. There are two exceptions, “development” and “growth”, which are not included in the Orphanet Thesaurus, but are included in the Children and Youth version of the ICF document (ICF-CY [122]), which is the source of this thesaurus. Since the resource provided by Orphanet is only available in English, we translated the corresponding terms into Spanish. Table 3.3 shows an extract of the list of annotated functions which are involved in some disability expressions. The complete list can be found in Appendix

A (Tables A.1 and A.2). These tables also show the relationship between the annotated functions and those included in the Orphanet thesaurus.

ability to recognize face / reconocer rostros	growth / desarrollo
academic / académico	hearing / audición
activities of daily living / actividades del día a día	intellectual / intelectual
attention / atención	processing speed / agilidad mental
gait / paso	receptive vocabulary / comprensión

Table 3.3: Excerpt of functions used to express a disability which have been identified during the annotation process.

The annotation of disabilities specifies the word that indicates absence or limitation, and the function affected by this. Because a disability can be represented by a negated function, some negation triggers are included in the list of impairment terms. Table 3.4 shows some examples of words that indicate the absence or limitation of a common human function. Additional examples of impairment words can be found in Appendix A, Table A.3.

aberration / aberración	impairment / impedimento
abnormal / anormal	inability / incapacidad
absence / ausencia	limitation / limitación
absent / ausente	limited / limitado
deficit / déficit	loss / pérdida
degradation / degradación	lower / bajo
delay / retraso	no / no
difficulty / dificultad	unable / incapaz
disability / discapacidad	

Table 3.4: Excerpt of impairment words obtained after the annotation process.

Since the literature includes different thresholds or scales related to the same disability, when a disability is mentioned with severity or temporality modifiers, we extended the annotation to cover these aspects. Nevertheless, although the

modifiers affecting a disability are of great importance to assess the needs of affected people, we understand that their inclusion in the annotations makes difficult the generation and evaluation of systems with more complex objectives, e.g. the extraction of cause-effect relationships between different pathologies. For this reason, we defined an additional annotation style based on the concept of core-term matching introduced by Fukuda et al. [54]. Focusing on the annotation of disabilities, using this annotation style we seek to remove from the spotlight terms related to the severity or duration. Thus, if we consider a disability such as "severe inability to walk, with significant progression", the resulting annotation would be focused on the expression "inability to walk". We used this second style of annotation to study in depth the recall of the generated systems by evaluating one of the developed corpora.

*"The hospitalization of the youth took place when, due to the paralysis in his left leg, a severe **inability to walk** with significant progression was observed."*

Full annotation:
The disability is annotated in addition to all possible modifiers that affect it. → "severe inability to walk with significant progression"
Partial Annotation:
Severity and temporality modifiers are omitted. → "inability to walk"

Annotating acronyms and abbreviations: During the annotation process, we found several acronyms referring to a disability. Table 3.5 shows an excerpt of the acronyms and abbreviations found during the annotation process. These acronyms are annotated as disabilities only if the extended form appeared in the same abstract. It is possible that in other contexts or abstracts, the annotated acronyms have other meanings.

Short-form	Extended form
ADCA	autosomal dominant cerebellar ataxia
ADHD	attention deficiency hyperactivity disorder
ARNSHL	autosomal recessive non-syndromic hearing loss
AT	Ataxia-telangiectasia
HL	hearing loss
bvFTD	behavioral variant frontotemporal dementia
CAPD	central auditory processing disorder
DED	depression-executive dysfunction
EA	ear anomalies
FMRL	fragile X mental retardation
FTD	frontotemporal dementia
HSP	Hereditary Spastic Paraplegia
ID	intellectual disability
NDED	non-depressive emotional disturbances

Table 3.5: RDD corpus: List of acronyms and their extended form, identified during the annotation process.

Annotating negation & speculation: To annotate negation and speculation, we used a similar criteria to those used for collections such as BioScope [153]. In both cases, we annotated both the trigger and the scope. We limited the annotation of negation and speculation to those cases where one or more disabilities are affected by these linguistic phenomena, i.e. negative contexts affecting diseases, drugs, etc. are ignored. Some negation triggers annotated were "absence", "absent", "doesn't", "except", "negative", "no", "no evidence", "none", "not", "rather than", "with the exception of", and "without". About speculation, the triggers found during the annotation process include "and/or", "apparent", "appear", "suggest", etc.

For the annotation of negation in non-English documents, we tried to use a similar criterion to that applied to English documents, annotating only those negations that directly affect a disability. In both languages, most annotated

negations begin at the first recognized negation trigger. For Spanish, the considered negation triggers were translations of the triggers being listed in the documentation of BioScope corpus.

Annotating relationships: We extracted explicit mentions to relations between rare diseases and disabilities. The scope of a relationship is restricted to the sentence level. We selected only sentences containing at least a rare disease and a disability. We supported the annotation of rare diseases using the Orphanet list. In addition, we covered the annotation of speculative relationships, i.e. relationships that express a possibility and not a fact with absolute certainty. We divided the relationships into two files, one for positive relationships and a second one for negative relationships. Negative relations are sentences mentioning a rare disease and a disability, but without stating a relation between them. The file of negative relationships also includes negated relations.

3.2.3 Presentation format

Disabilities: Each disability is annotated with the XML tag <dis>: **<dis id=[0 - 9]+> .* </dis>**. This element has an attribute which is an identifier to distinguish different disabilities within the same sentence. In the following example, we can see how the disability "paraparesis" was annotated with the identifier 0 (id=0), since it is the first disability named in the sentence.

Including severity and temporal modifiers

As a cause of the deterioration caused by the disease, the patient showed signs of <dis id=0>severe/moderate **paraparesis**</dis>.

Considering the full annotation criteria, the expression "severe/moderate" was annotated as part of the disability given the relevance of aspects related to disabilities such as severity or duration in time.

The disabilities expressed by the absence or limitation of functions were annotated using the XML tags <dis>, <fun> and <imp>. We used the tag <imp> to annotate terms or expressions that indicate the condition of limitation or absence, and the tag <fun> to indicate the function or functions affected by this altered condition. In the next example, the expression "mental retardation" expresses a disability which involves behavior outside the normal range of mental abilities. This negative altered condition is expressed by the term "retardation".

Disability (Function + Impairment)

A high incidence of <dis id=0><fun>mental</fun> <imp>retardation</imp></dis> has been reported in patients with Angelman syndrome.

If a negation trigger plays the role of an impairment term, we marked this expression using the XML tag <imp> to indicate that this term is affecting a function. If we consider the following instance, the term "no" indicates the absence of a normally assumable skill.

Negation triggers

Patients do <dis id=0><imp>not</imp> have the <fun>ability to stand up for themselves</fun></dis>.

Finally, considering the flexibility of the language and the fact that an expression denoting impairment may affect more than one function (i.e., enumerations), we annotated all functions affected by the same impairment under the same disability tag.

Involving multiple functions

After the accident the youth presented <dis id=0>serious <imp>difficulties</imp> in <fun>walking</fun> and <fun>talking</fun></dis>.

Negation: The XML tag used to annotate a negation trigger is <neg> and for the scope is <scp>. The tag for the scope has an identifier as an attribute to identify the different negated contexts within the same sentence. The scope of a negated disability depends on the syntax, it usually covers the biggest affected phrase. Generally, the scope of negative auxiliaries, adjectives, and adverbs starts right with the negation word and finishes at the end of the phrase, as in the next example:

Annotating negation

Results of the tests carried out show <scp id=0><neg>no evidence</neg> of <dis id=0><fun>cognitive</fun> <imp>impairment</imp></dis></scp>.

Disabilities phrases including a negative keyword are not necessarily annotated for negation, for example, because they may be in a speculative form.

Speculation: The XML tag used to annotate speculative triggers is <spe>. To delimit the scope of a speculative context we used the tag <ssc>. Similar to the annotation of negation, the tag for the scope has an identifier as attribute to recognize the different speculations within the same sentence. We consider the minimal unit expressing doubt for marking the speculative triggers. However, there are cases in which the doubt is expressed by several words together. Then, all of them are annotated as the speculative expression.

Relationships: We used the following format to record the annotated relationships in a separate tabular file:

> **1st column:** PUBMED code of the article to which the abstract belongs.
>
> **2nd column:** The sentence.
>
> **3rd column:** Rare disease affected by the relationship.
>
> **4th and 5th columns:** Offset of the rare disease.
>
> **6th column:** Disability affected by the relationship.
>
> **7th and 8th columns:** Offset of the disability.
>
> **9th column:** Boolean indicating if the relationship is in speculative form.
>
> **10th colum:** Orphanet ID of the rare disease.

Let us consider an example of a positive relationship:

> "21838783 | Coffin–Lowry syndrome is a syndromic form of mental retardation caused by mutations of the Rps6ka3 gene encoding ribosomal s6 kinase (RSK)2. | Coffin–Lowry syndrome | 1 | 22 | mental retardation | 46 | 64 | 0 | 192".

This relation associates "mental retardation" with "Coffin–Lowry syndrome". If there are more than one relationship in the same sentence, all of them are included in the corpus as different entries.

3.2.4 Corpus: RDD

In 2018 we published the RDD corpus with the aim of studying in scientific papers, the recognition of named entities and the extraction of relationships [44]. The

developed corpus gathers a collection of abstracts of scientific articles concerning rare diseases and disabilities. Under the supervision of expert medical staff and using the methodology previously described, this corpus was annotated by three annotators. RDD corpus contains a version of the gathered abstracts, containing annotations on the different mentioned disabilities. Given the importance of different linguistic phenomena in the area of information extraction, we annotated negations and speculations that affect one or more disabilities mentioned in each document. On the other hand, this corpus also contains a file with the relationships between rare diseases and disabilities stated within the documents. So far, this was the largest corpus that explicitly collected annotations on disabilities and rare diseases. The annotation guide used for the creation of this corpus is detailed in the Section 3.2.2. Among the commented criteria for the annotation of disabilities, we only considered the criteria of full annotation for the development of the RDD corpus.

Statistical data

The RDD corpus is composed of 1000 abstracts in English, with an average of approximately 200 words per document. With 9657 sentences, the collected documents cover 578 rare diseases. In total, this corpus contains 3678 annotations expressing a disability. From them, 2792 are expressed as the impairment of a human function and 886 are stated using some disability term. We found that the physical function most often affected by some kind of impairment is hearing. Sight and motor skills are often found impaired too. The most frequently mentioned disability is ataxia, related to motor skills. Deafness appears in second place, whereas dementia, related to problems in cognitive functions, is the third one. Other disabilities such as blindness are also very frequent. Concerning negation, the corpus includes 90 negated disabilities, corresponding to 83 negation annotations, since a negation can include more than one disability. On the other hand, the corpus also includes 194 annotations of speculation, that affect 264 disabilities. We identified 1251 positive relationships and 706 negative. From them, 86 are speculative in the positive set and 8 in the negative set. The identified relationships cover 362 different rare diseases. In 186 cases, a disability tag corresponds to an acronym.

Annotation	Total	Inter-Agreement
Disability	3678	0.87
Negation	83	0.93
Speculation	194	0.67
Relationships	1251	0.77

Table 3.6: RDD corpus: Details of agreement reached between annotators.

Table 3.6 shows details about the generated annotations. This table shows for each type of entity, the total number of annotations and the agreement reached during the annotation process. We do not include annotations that differ in one or two characters in the disagreement counts. Disagreements include omissions as well as inexact matches, in which the spams of the two annotations coincide in some word but not in all of them. The agreement in the annotation of disabilities and negation is high. The speculative annotations seem to be the more difficult ones, since words such as "indicate" are sometimes used to express speculation, but sometimes it seems that they are used as an assertion. The scope of the speculation is also difficult to establish in some cases. The disagreement in the annotation of the relationships only corresponds to the positive relations and it is affected by the disagreement in the annotation of the disabilities. In summary, the results of the agreement suggest that the corpus is robust enough to be used in the study and evaluation of automatic systems.

3.2.5 Corpus: DIANN

After the annotation of RDD corpus, and still focused on the recognition of disabilities in scientific works, we proposed to address the problem in languages other than English. The annotation guidelines generated during the preparation of the RDD corpus facilitated the proposal of new resources to cover the study of this aspect, among others. We published the DIANN corpus in 2018 [45], under the umbrella of the IberEval (Evaluation of Human Language Technologies for Iberian Languages) 2018 conference [137]. This corpus was presented as a common evaluation framework of tools and approaches for the detection of mentions of disabilities in documents

written in Spanish and English. It was used as a benchmark for a homonymous task collocated in IberEval conference. DIANN shared task served us to evaluate, in different languages, the complexity of processing types of named entities such as disabilities.

In total, the DIANN corpus includes 1000 annotated documents, 500 published in English and 500 in Spanish. In contrast to the RDD corpus, which collected only documents in English, DIANN corpus includes abstracts of scientific articles from Elsevier journal papers and published in Spanish and English. We restricted the document search process to documents with the abstract in both English and Spanish languages and at least they contain a mention of disability in both languages. In addition to the generation of resources in multiple languages, during the generation of this corpus, we aimed to allow the study of the identification of named disabilities at different levels of granularity. For this purpose, we established a double annotation criteria. On the one hand, an annotation of the disabilities mentioned in the different documents, including possible temporality and/or severity modifiers. The same annotation criteria were used for the generation of the RDD corpus. On the other hand, to analyze the complexity of the study of this type of modifiers, a second annotation criterion was applied, based on core-term matching. This second annotation criterion does not include in the annotation of a disability any terms beyond the disability itself. The annotation guide used for the creation of this corpus is detailed in the Section 3.2.2.

Statistical data

The DIANN corpus is composed of two sub-collections. On the one hand, the collection of documents in Spanish and on the other hand, the collection in English. Although the Spanish documents are representations of the English documents, they do not correspond to literal translations, i.e. the collected documents are abstracts of research articles distributed in both English and Spanish. As a consequence, it is possible that the number of annotations in the Spanish version of an abstract differs from the number of annotations in the English version, as shown in the statistics provided in Table 3.7.

Annotation	Language	Total	Inter-Agreement
Disability	English	1656	0.89
Disability	Spanish	1555	0.78
Negation	English	63	0.90
Negation	Spanish	62	0.87

Table 3.7: DIANN corpus: Details of agreement reached between annotators for the English and Spanish collections.

For Spanish, this corpus gathers 1555 mentions to disabilities (564 unique mentions) and, for English, 1656 (583 unique mentions). Regarding negation, we annotated 63 instances for English and 62 for Spanish. Although the agreement reached during the generation of this corpus indicates that it is robust enough to be released and proposed as a benchmark, the difference in agreement between languages shows that the set of documents in Spanish was more difficult to annotate.

3.3 Discussion

In order to support the exploration of information extraction techniques for documents related to disabilities and rare diseases, during this thesis we developed two corpora. These corpora contain different types of annotations that cover aspects of interest related to this context. Table 3.8 shows an overview of the aspects covered by each corpus.

		Annotations				Relationships
Corpus	**Language**	**Disability**	**Negation**	**Speculation**	**Matching**	**Rare Diseases and Disabilities**
RDD	English	X	X	X	Exact	X
DIANN	English / Spanish	X	X		Partial / Exact	

Table 3.8: Analysis aspects covered by DIANN and RDD corpora. The annotation criteria (partial and/or exact) is also compared.

On the one hand, the RDD corpus contains documents in English and it includes annotations on disabilities, negations and speculations. In addition, this corpus

includes for each document annotations on relationships between disabilities and rare diseases. On the other hand, the DIANN corpus contains documents in English and Spanish, and includes annotations on disabilities and negations. For the elaboration of this corpus, we used two styles of annotations. The first one is based on a full annotation, which includes each and every one of the terms that mention a disability. And the second one is based on the concept of core-term matching. This style of annotation seeks to focus the attention on small units. In this case, we annotated the disabilities avoiding temporality and/or severity modifiers.

Finally, the agreement reached between annotators during the development of both corpora was high enough to consider both corpora to be valid for the development and validation of tools (Tables 3.6 and 3.7). However, although both corpora collect a considerable number of annotations, the small number of negations and speculations makes it difficult to study these aspects through a machine learning system only considering these corpora.

Chapter 4

Named Entity Recognition

Contents

This chapter describes the work we carried out during this thesis on the detection of disabilities in biomedical documents. Firstly, Section 4.1 presents our first study of NER techniques applied to the recognition of disabilities. We tried to follow a constructive approach, analyzing simple schemes based on different types of techniques. In this Section, we discuss the experiments based on machine learning and deep learning techniques that we carried out for the study of the RDD corpus. Further on, we modified the scope of the study by extending the analysis to a multilingual context. Section 3.2.5 presents DIANN evaluation task, focused on the analysis of disability recognition techniques applied to two different languages. We analyzed the task as well as the different systems proposed by the participating teams. In this section, among the proposals, we present our participation in this task. Inspired on Metamap workflow, we addressed the problem through an unsupervised approach. In Section 4.3 we analyze our participation in two evaluation tasks related to the detection of several medical entities. In an effort to explore different scenarios, and focused on developing a system based on the lessons learned during the DIANN task, these tasks gave us the opportunity to prove in other contexts the conclusions reached for the identification of named entities. We present different experiments carried out in order to explore different aspects of proposals based on deep learning and rule systems. Finally, in Section 4.4 we present a proposal based on the lessons learned during the development of this thesis in the area of NER.

4.1 NER - Preliminary study: Working on RDD corpus

In the process of paving the way for the study of biomedical entities such as disabilities and rare diseases, the development of the RDD corpus was our first work [44]. This work represented the opportunity to evaluate automatic systems strictly dedicated to deal with these entities. Additionally, and analyzing the RDD corpus, we performed a set of experiments covering aspects related to named entity recognition and relationship extraction. Regarding the recognition of named entities, we carried out experiments for the detection of disabilities and rare diseases. The main objective

of these experiments was to establish a preliminary framework for the study of these entities using the RDD corpus. In this section we detail the experiments carried out on named entity recognition. The experiments on relationship extraction exploring this corpus, are detailed in Section 5.1.

Experimental methodology

The RDD corpus includes two separate parts, one focused mainly on the study of the recognition of mentions to disabilities, and the other one focused on covering the analysis of relationships between disabilities and rare diseases. Since the files containing these relationships include information about named entities and stated relationships, we decided to propose an experimental set-up based only on these files. This framework allowed us to analyze under the same umbrella the results obtained for the recognition of named entities and the extraction of relationships. Details about the format of the annotations contained in these files are discussed in Chapter 3.

For each sentence, these files contain as many instances as relationships between disabilities and rare diseases are stated. To adapt the format of these files to one compatible with NER systems, we represented sentences such as "Many neurodevelopmental disorders exhibit syndromic obesity including SMS.", that mentions of "neurodevelopmental disorders" with "syndromic obesity" and "SMS", as follows:

> "Many (O) neurodevelopmental (B-Disability) disorders (I-Disability) exhibit (O) syndromic (B-Disesae) obesity (I-Disesae) including (O) SMS (B-Disesae)."

Example 4.1: RDD corpus: Mentions to disabilities and rare diseases annotated in a sentence included in the corpus.

As shown in this example, we used the IOB notation. To label the first word of an entity we used the B tag. The rest of words from that entity were labeled with the I

tag. The terms that do not belong to any entity were labeled with the O tag. Among the limitations of this representation format is the impossibility of representing nested entities. However, we decided to simplify the task removing relations with nested gold annotations (e.g., the RD "X-linked mental retardation" and the disability "mental retardation"). This adaptation aligns this work with other studies such as Li et al. [97].

4.1.1 First proposed model: Supervised approach

Figure 4-1 illustrates the model of deep learning explored by us. A neural architecture using two inputs based on embeddings and one-hot vectors.

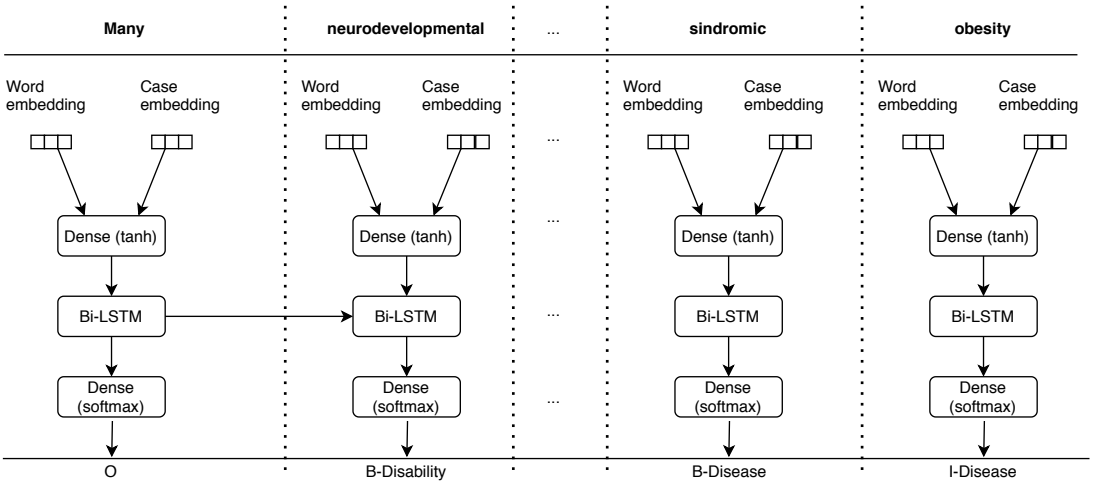


Figure 4-1: RDD Corpus: Deep learning model for disabilities and diseases recognition.

This model used pre-trained Word Embeddings to represent the words of each sentence. We used the Word Embeddings generated by Levy and Goldberg [96]. In order to optimize the size of the generated system, we reduced the considered vocabulary by transforming all words to lowercase. Although this practice is common, the destruction of information that it entails is evident. We used an additional representation to store the information lost due to the process of vocabulary reduction. Specifically, we used a CASE feature to indicate if a word is lowercase, all upper-case,

if it had first letter capital, or if it had at least one non-initial capital letter (Algorithm 4.1). This feature was represented using a one-hot encoding matrix of size 8.

Algorithm 4.1: Algorithm: Generating a casing one-hot vector.

```
function getCasing(word):
    numDigits = 0
    foreach character in word:
        numDigits = numDigits + toint(isdigit(character))
    if isuppercase(word): # All characters are in upper case
        casing = 'allUpper'
    elif isuppercase(word[0]): # initial char upper, then all lower
        casing = 'initialUpper'
    elif islowercase(word): # All lower case
        casing = 'allLower'
    elif isdigit(word): # Is a digit
        casing = 'numeric'
    elif numDigits > 0: # Contains digits
        casing = 'contains_digit'
    return casing
```

On the composition of the proposed architecture, the first layer was a densely connected hidden layer (Dense), with 100 neurons and a hyperbolic tangent (*tanh*) activation function. This layer takes the concatenation of the features as input. Using this layer, we sought to reduce the space generated by the inputs to dimensions more easily manageable by posterior layers.

To analyze each term as part of a context, the next layer of the model was a bi-directional LSTM. In each step, this network processed the captured information, analyzing it in both directions. We set the dimension of each LSTM to 100. The output of this network was the concatenation of the information generated in both directions. Finally, in order to calculate the probabilities of all entity labels, we used another Dense hidden layer with 4 neurons (Labels + Padding Token) and a softmax activation function.

This model was trained on each cross-validation fold up to 150 epochs using an early stop criterion depending on the loss function values obtained at the end of each training epoch. We used AdaGrad optimizer [42] with default parameters. The remaining parameters were configured following the indications of previous works such as [97].

4.1.2 Results & Analysis

We carried out an evaluation of the proposed models using a 10-fold cross-validation. In order to ensure as far as possible the efficiency of this type of evaluation, especially in methods based on deep learning, we established fixed parameters for the initialization of the weights of each part of the proposed models. To evaluate the output, the tags 'B-' and 'I-' were only considered correct if they were in the correct sequence. Table 4.1 shows the results obtained using support vector machines (SVM) and the Bi-LSTM approach. We provide aggregated results for both kinds of entities, and detailed results for each kind of entity.

For the SVM classifier, inspired by other work [39], we used as features the PoS tag of the current word and a vector representation of the word itself and the two following adjacent terms. We utilized the SVM implementation provided by Weka, using default parameters [159]. Although we tried other approaches based on machine learning, SVM offered the best average results.

The results for dealing with both types of entities jointly (RD+DI) indicate that the system based on Bi-LSTM reaches a better performance. However, analyzing these results by type of entity, while the difference between SVM and Bi-LSTM is outstanding for the recognition of disabilities, it is not so obvious for the recognition of rare diseases. This behavior may be due to the fact that the RDD corpus gathers more disabilities than rare diseases. Being the use of annotated data a basic requirement of supervised algorithms, circumstances, such as this one, evidence that deep learning algorithms need a larger amount of data to achieve optimal generalization results, compared to other techniques. On the other hand, some words frequently used to name rare diseases are often vaguely represented in Word Embeddings due to their low frequency of appearance in relation to the context, i.e. the common use of the term "Angelman" is as a last name, while expressions such as "Angelman Syndrome" are common in this corpus. This may affect systems based on embeddings when they try to process these words without the use of additional representations.

	Precision (sd)	Recall (sd)	F-measure
SVM(RD+DI)	0.723 (0.0314)	0.524 (0.0134)	0.608
SVM(RD)	0.621 (0.0476)	0.528 (0.0457)	0.57
SVM(DI)	0.625 (0.0274)	0.523 (0.0344)	0.569
LSTM-W(RD+DI)	0.711 (0.0341)	0.647 (0.0315)	0.677
LSTM-W(RD)	0.583 (0.0333)	0.588 (0.0472)	0.585
LSTM-W(DI)	0.72 (0.039)	0.696 (0.0319)	0.707
LSTM-W+C(RD+DI)	0.767 (0.0242)	0.684 (0.0326)	0.723
LSTM-W+C(RD)	0.632 (0.0281)	0.619 (0.0432)	0.625
LSTM-W+C(DI)	**0.763** (0.039)	**0.749** (0.0471)	**0.755**

Table 4.1: Results of *precision*, *recall* and *f-measure* obtained identifying entities in the RDD corpus with a classic machine learning method, SVM (first frame), and with a LSTM neural network architecture, using only Word Embeddings (-W) (second frame) and using word and case information (-W+C) embeddings (last frame). Standard deviation appears under each value (sd). RD stands for rare disease, DI for disabilities and RD+DI for aggregate data. Best results for detecting disabilities appear in boldface.

Analyzing the *precision/recall* results achieved in both experiments, it is evident that the use of techniques such as Word Embeddings resulted in a greater capacity for generalization and adaptation, in contrast to the contribution obtained by the representation based on bags of words. After examining the recall obtained by the different techniques, a large part of the false negatives produced by SVM refer to entities that were not completely identified. LSTM presents a better performance in the complete recognition of entities independently of their length.

4.1.3 Discussion

This work exploring the RDD corpus was an early approach to the task of recognizing disabilities and rare diseases in biomedical documents. Based on previous works, we proposed an experimentation using a sequential approach, where we assumed that the generated label at the K step ($K > 0$) depends on the information processed in previous steps. In order to explore the RDD corpus, we proposed some preliminary experiments based on deep learning (Bi-LSTM) and machine learning (SVM). The

efforts we made in feature engineering were relatively scarce. With these experiments we only sought to establish an initial framework for the analysis of the RDD corpus.

We observe that the use of Word Embeddings for the task provides high performance for both *precision* and *recall*, reaching an *f-measure* of 79.13% for disability recognition. It is slightly improved by including casing information in the model, reaching then an *f-measure* of 81.11%.

4.2 Exploring other languages: DIANN Shared Task

As part of the 2018 IberEval workshop [137], we proposed the DIANN shared task focused on the detection of disabilities.To evaluate the systems developed by the participants, we made the DIANN corpus available. The details of this corpus were discussed in Sections 3.2.2 and 3.2.5. In contrast to the RDD corpus, the DIANN corpus is exclusively oriented to the study of named disabilities. However, while the RDD corpus gathers only documents in English, the DIANN corpus includes documents in English and Spanish. According to our knowledge, this is the first organized task addressing the recognition of named disabilities in a multilingual context.

To organize this task, we divided DIANN corpus into two parts, one for training and the other one for test. In addition to the training corpus and to contextualize the problem, we provided to the participating teams, an excerpt of identified disabilities in both Spanish and English languages. According to the scheduling specifications of the task, participants had one month to develop their systems since the publication of the training corpus. Then, we released the test set without annotations and the participants had fifteen days to send their results to the task organizers. We indicated to each team that they could present up to three different approaches per language.

Although the DIANN task covered aspects related to negation processing, in this Section we analyze the proposals made only for the disability annotation task.

Evaluation criteria

In addition to the exact matching and due to the freedom with which a disability can be expressed, we used a second evaluation criteria to compare the different systems. This second evaluation criteria, called partial matching, is based on the concept of core-term match [54]. Whereas the exact matching criteria checks if every proposed annotation matches exactly with the ground truth, the partial matching criteria checks if each disability has at least its identified minimum unit or core contained in the ground truth. To carry out the evaluation with the partial match criteria, we used the file provided with the DIANN corpus containing annotations about the core of each mentioned disability. Some examples included in this file can be seen in Fig. 4-2. The motivation for the study of this second evaluation method seeks to make an in-depth analysis of the recall obtained by the presented approaches. For each evaluation criteria, the performance was measured with $F_{\beta=1}$ rate:

$$F_\beta = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \tag{4.1}$$

where *precision* is the percentage of named entities found by the system that are correct or partially correct and *recall* is the percentage of named entities present in the corpus that are found or partially found by the system.

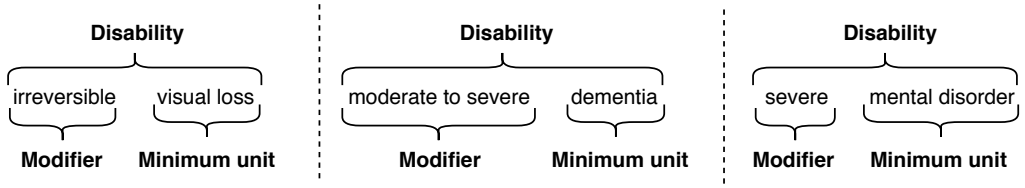


Figure 4-2: DIANN Shared Task: Partial evaluation examples extracted from the ground truth and where the minimum unit considered for each disability is shown.

Summary of participating proposals

A total of 8 teams participated in the task, presenting 18 runs for English and 19 for Spanish. The participating teams presented a wide variety of approaches. The teams proposed the use of different types of resources and suggested both supervised and unsupervised approaches. We proposed an unsupervised system, which we detail in Section 4.2.1. The rest of the proposals are described below.

- **SINAI [102] - Group of Intelligent Systems of Information Access, University of Jaén.** SINAI group proposed an unsupervised system based on the generation of variants using UMLS (Unified Medical Language System) terminology and Word Embeddings. The system used two different biomedical entity extractors, MetaMap for English and a similar tool for Spanish. After identifying potential expressions, the group proposed a filtering process based on two aspects: semantic categories manually identified as relevant and the analysis of the similarity between the candidate expressions and the expression "disability" using Word Embeddings.

- **IxaMed [58] - Ixa Group, University of the Basque Country.** For the detection of named disabilities, this group used a deep learning system consisting of a Bidirectional Long Short Term Memory network and a Conditional Random Field at the top. To process the English documents, IxaMed used Word Embeddings and Brown clusters as inputs [17], while for the Spanish documents they only used Word Embeddings. Although IxaMed used different Word Embeddings depending on the language, both embeddings were generated using documents from the biomedical domain (English: MIMIC-III corpus [77] - Spanish: Electronic health records in Spanish). Finally, they used a rule-based system for the detection of abbreviations.

- **IXA [1] - Ixa Group, University of the Basque Country.** This model deals with the detection of both disabilities and negation triggers using the ixa-pipe-nerc tool [2] and the Perceptron implementation provided by the Apache OpenNLP project. Among the different types of inputs used by IXA were public gazetteers and clusters, such as Brown and Clark [27].

- **GPLSIUA [118] - Natural Language Processing and Information Systems Group, University of Alicante.** Analogous to the proposal submitted by SINAI group, this approach was a system divided into two parts: an expression extractor and a candidate expression filtering process. While the extraction method proposed by SINAI group was based on external biomedical resources, the method proposed by this team was much more generic, dealing only with the extraction of all noun phrases in each sentence. This proposal transferred the responsibility of filtering candidates to a machine learning system known as CARMEN [117], which uses a Random Forest and was trained with several textual features, e.g. suffixes, affixes, etc. This system also used some contextual information on the "relevance" of lemmas of certain terms that appear in a fixed-size contextual window.

- **UPC_3 [106] - TALP research group, Polytechnic University of Catalonia.** Two different proposals were presented by this team: one was based only on CRF and another one based on the tuple Bi-LSTM+CRF. While the CRF model was trained using several features, including lemmas, suffixes, part-of-speech and Word Embeddings, the Bi-LSTM+CRF model was trained only using Word Embeddings. UPC_3 carried out several experiments with different kinds of embeddings. One was trained with generic documents (sources such as Wikipedia among others). Another one was trained only with documents from the biomedical domain (research articles, electronic health records, etc.). In order to avoid over-fitting during the training, this team proposed a semi-supervised methodology using unlabeled documents during the training.

- **UPC_2 [152] - TALP research group, Polytechnic University of Catalonia.** Based only on the use of a CRF for the identification of named disabilities, the system that proposed this team was trained with both syntactic and semantic features. Thus, some studied features include casing information such as capitalization and the use of non-alphanumeric elements, which were considered useful by different teams for the detection of abbreviations. This team also used a list of terms that were extracted from the training set in order to represent if a term found in the test set is of interest or not.

- **UC3M [165] - Human Language and Accessibility Technologies Group, Carlos III University.** This team used a Bi-LSTM+CRF based architecture to deal at the same time with the recognition of named disabilities and the identification of negation. Unlike other teams that used a similar architecture, UC3M trained the system using exclusively the following distributed representations: Word Embeddings for terms representation, character embeddings to represent n-grams of characters and sense embeddings for disambiguation. This team used a LSTM to process the sequence of character embeddings of each word. The output of this LSTM, along with Word Embeddings and sense embeddings, were the input to the Bi-LSTM+CRF model.

4.2.1 Proposed approach - LSI_UNED: System description

There are many benefits associated with unsupervised systems, especially in domains such as biomedical, where the availability of annotated data is relatively limited. For this reason, and focusing on the DIANN Shared Task, we explored an unsupervised system for the detection of disabilities, inspired by MetaMap [9].

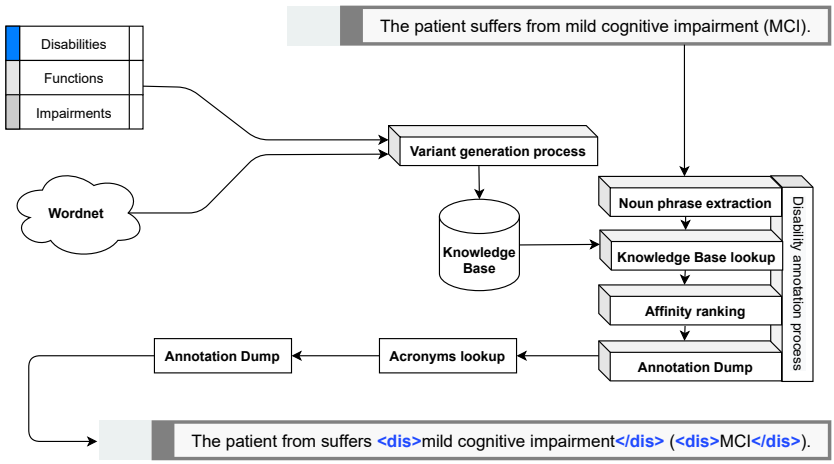


Figure 4-3: DIANN Shared Task: Unsupervised model proposed by LSI_UNED group.

Figure 4-3 summarizes the phases of the generated pipeline. First, we considered a set of manually created terminologies. These lists contained examples of disabilities, body functions, and impairment words. In order to generate a knowledge base using these lists, we applied a variant generation process based on Wordnet (for English) and EuroWordnet (for Spanish) [111, 155]. For each expression included in the lists, we calculated a set of terminological variants. We generated both derivative (e.g. "magnetic" ⇒ "magnetism" or "simply" ⇒ "simple") and synonym-based variants (e.g. "pipe" ⇒ "tube" or "sad" ⇒ "unhappy"). The number of variants generated for each expression depends on the number of tokens that form the expression and are contained in Wordnet.

Once the knowledge base was generated, the disability annotation process involved extracting noun phrases and looking at the knowledge base for candidate expressions with a high degree of affinity. We evaluated the affinity of the extracted noun phrases with the expressions compiled in the knowledge base. Given a noun phrase, and its affinity, we established a ranking between the different candidates in the knowledge base. We used this ranking to sort the terms in the knowledge base according to extracted noun phrases. We defined the affinity function as follows,

$$affinity = \frac{distance + centrality + 2 * coverage + 2 * cohesiveness}{6.0} \tag{4.2}$$

where:

- Centrality: The centrality value is simply 1 if the term in the thesaurus involves the head of the noun phrase and 0 otherwise.
- Distance: The variation value estimates how much the variants in the thesaurus term differ from the corresponding words in the noun phrase.
- Coverage: The coverage value indicates how much of the thesaurus term and the noun phrase are involved in the match.
- Cohesiveness: The cohesiveness value is similar to the coverage value but emphasizes the importance of connected components. A connected component is a maximal sequence of contiguous words participating in the match.

Finally, the detection of acronyms was an additional independent stage since it was not covered by the process of generating variants. The method for the detection of acronyms was based on the use of regular expressions. Using the annotations made previously, our approach used regular expressions such as the ones studied by Montalvo et al. [114]. We used a quite limited set of rules. These rules were based on the premise that all abbreviations are presented in the following format:

Template 1: .* <dis>ENTITY</dis> \((?acronymm'[A-Z0-9]+)\).*

Template 2: .* <dis>ENTITY</dis> \((?acronymm'[A-Z0-9]+)[;:].*\).*

Given an annotated entity, if it was found next to an expression in capital letters and bounded by parentheses, the system would consider this expression as an acronym of the preceding entity (Template 1). Additionally, if the acronym was followed by some kind of symbol (e.g. semicolon), only the part before the symbol would be considered as an acronym (Template 2). Once an acronym is annotated, all its appearances are annotated. After that, the document is processed again from the line where the acronym is found by searching for words containing the initials of the acronym. The main problem with this form of annotation is that it depends on the annotation of the entities because, for the regular expression to work, the entity that precedes the acronym must be correctly annotated.

4.2.2 Participating systems: Comparison & Results

Table 4.2 shows a comparison of the different features and resources used by each team to deal with the recognition of named disabilities. Both supervised and unsupervised systems are included in the comparison. Although the use of embeddings was very common, not all the supervised systems used them. UPC_3 presented runs studying embeddings from different sources (generic and specific domain), and UC3M considered both, character and sense embeddings, as well as Word Embeddings. IxaMed applied calculated embeddings using electronic health reports. Furthermore, amongst other NLP techniques, such as lemmatization or the extraction of suffixes/prefixes, the use of clustering and part-of-speech techniques were a popular practice, especially to reduce

and to label the considered vocabulary. Part-of-speech taggers were considered useful by the participants for the identification of qualifying expressions i.e. expressions which denote temporality or severity. Finally, other proposals introduced by the participants were the use of casing information and some external resources. Our system used both casing information and regular expressions for the identification of abbreviations.

	IXA	IxaMed	UPC_3	LSI_UNED	GPLSIUA	SINAI	UPC_2	UC3M
Embeddings		E	G-E			G		G-C-S
Clustering	✗	✗	✗					
Dictionaries	✗			✗		✗		
Part-of-Speech			✗		✗		✗	
Wordnet				✗				
Char n-grams	✗				✗			
Lemmas			✗		✗		✗	
Casing		✗	✗	✗			✗	

Table 4.2: DIANN Shared Task: Comparison of features and resources used by participating teams. Each column represents one of the participants and each row represents if a feature or resource has been used or not. (G) Generic domain sources - (E) Specific domain sources - (C) Characters - (S) Sense.

Evaluation results

In Appendix B we show the results obtained by the different systems. In order to simplify the presentation and analysis of the results, Table 4.3 shows only the best results achieved by each team for the task of named entity recognition. *Precision*, *recall* and *f-measure* are reported for each language taking into account the different matching criteria.

	SPANISH			ENGLISH		
	Precision	Recall	F1	Precision	Recall	F1
IXA (S)	0.65 (0.72)	0.64 (0.71)	0.64 (0.71)	0.7 (0.76)	0.53 (0.57)	0.60 (0.65)
IxaMed (S)	0.75 (0.82)	**0.81 (0.88)**	**0.78 (0.85)**	0.78 (0.84)	**0.86 (0.92)**	**0.82 (0.88)**
UPC_3 (S)	**0.81 (0.89)**	0.59 (0.65)	0.68 (0.75)	0.79 (0.87)	0.60 (0.66)	0.68 (0.75)
GPLSIUA (S)	0.79 (0.95)	0.17 (0.20)	0.28 (0.33)	**0.88 (0.91)**	0.25 (0.25)	0.39 (0.4)
SINAI (U)	0.18 (0.20)	0.41 (0.46)	0.25 (0.28)	0.22 (0.25)	0.42 (0.48)	0.29 (0.33)
UPC_2 (S)	0.73 (0.82)	0.50 (0.56)	0.59 (0.67)	0.75 (0.82)	0.56 (0.6)	0.64 (0.7)
UC3M (S)	0.80 (0.88)	0.65 (0.71)	0.71 (0.79)	0.77 (0.82)	0.72 (0.76)	0.74 (0.79)
Our system (U)	0.41 (0.84)	0.24 (0.51)	0.31 (0.63)	0.67 (0.85)	0.59 (0.76)	0.63 (0.8)

Table 4.3: DIANN Shared Task: Results of named disability recognition using exact and partial evaluation (in brackets). The results obtained by both, supervised (S) and unsupervised (U) systems are shown. Best results for each language are shown in boldface.

The best systems were presented by IxaMed, UC3M and UPC_3; all of them based on Bi-LSTM and CRF. While UC3M and UPC_3 proposed a strategy based on one classifier, IxaMed opted for a cascade approach considering the annotation of disabilities and the annotation of abbreviations in different phases. Strategies based on the use of rules for the detection of abbreviations showed a high performance. The IxaMed group and we adopted this strategy to detect abbreviations.

Most systems exhibited notable differences in performance comparing partial and exact results. The unsupervised system proposed by SINAI was one of the least affected by the type of evaluation. However, this approach produced a large number of false positives, probably due to the mechanism used to filter candidate expressions according to the distance between embeddings. We also proposed an unsupervised approach divided into a candidate extraction phase and a heuristic-based filtering process. Our proposal obtained competitive results, especially in English and for the partial matching criteria. The versions of Wordnet used for each language may be the reason of the performance differences between languages. In addition, we generated short annotations, often ignoring temporal or severity modifiers. In this respect, approaches using part-of-speech or sequence processing architectures, e.g. LSTM and CRF, had better results. Finally, approaches using cluster-based representation and generalization methods, e.g. IXA and UPC_3, obtained very interesting results using

these methods intensively. To summarize, detected errors were mostly found on the following aspects:

- Temporal and/or severity modifiers were not detected. Many of the detected errors in the exact evaluation were related to this aspect.
- Identified diseases (e.g. Parkinson) or symptoms (e.g. headache) as disabilities. It is more frequently observed in approaches based on external resources.
- Some disabilities described with more than 4 or 5 words were not covered, e.g. "Patient unable to perform activities of daily living autonomously...". Both supervised and unsupervised systems reported those errors.

4.2.3 Discussion

Proposed unsupervised approaches were based on a similar pipeline architecture: candidate expression retrieval + filtering process based on external knowledge. Among the unsupervised proposals, our approach presented the best performance. Analyzing the results obtained by our system, the differences between exact and partial evaluation were particularly remarkable.

Overall, the most successful systems adopted supervised approaches. Although the best results were obtained by systems based on deep learning, other techniques such as CRF stand out. Finally, many of the conclusions reached during the exploration of the RDD corpus, were confirmed in many of the systems proposed in DIANN.

- The systems based on deep learning obtained the best results.
- The use of Word Embeddings, although it amplifies the capacity of generalization, is not sufficient to address the problem, being necessary the use of addictive representations.
- The overall low frequency of certain terms can be a problem working with Word Embeddings.
  - The use of domain-specific embeddings helps to deal with this issue.

- Exploring the identification of disabilities and abbreviations through a mixture of deep learning techniques and rules-based approaches reaches promising results.

4.3 Exploring other entities

In order to study different aspects proposed during the DIANN task in a more generalist context, we participated in two evaluation tasks focused on challenges of the biomedical domain [47, 49]. Both tasks were developed in the context of IberLEF 2019 (Iberian Languages Evaluation Forum) and were focused on processing documents in Spanish.

Since the work previously done gave us a background very focused on the detection of disabilities and rare diseases, our objective with these works was to study more general aspects related to the biomedical domain. We participated in the following tasks:

eHealth Knowledge Discovery (eHealth-KD) Challenge [127]: Focused on the processing of documents in the biomedical domain, this task addressed the identification and classification of keyphrases extracted from biomedical documents and, the extraction of relationships between the extracted keyphrases. Entity types covered by this task were:

Concept: a general category indicating if the key phrase is a relevant term, concept, idea, in the knowledge domain of the sentence.

Action: a concept that indicates a process or modification of other concepts.

Predicate: a category used to represent a function or filter that affects other elements in the text. These expressions can be understood as a semantic tag, e.g. “mayores” (older), and is applied to a concept, such as “personas” (people).

Reference: a textual element that refers to a concept, which can be indicated by textual clues such as “esta”, “aquel”, and similar.

Regarding the extraction of relationships, we present the details of the proposed model and the results obtained in section 5.2.

MEDDOCAN task [103]: This task addressed the anonymization of medical documents in Spanish. In short, this task addressed the identification and categorization of Protected Health Information (PHI), e.g. names, cities, etc. The organizers facilitated in different stages (training, development and test), a collection of medical reports in Spanish. The annotation guidelines used to annotate this corpus includes 29 different types of entities.

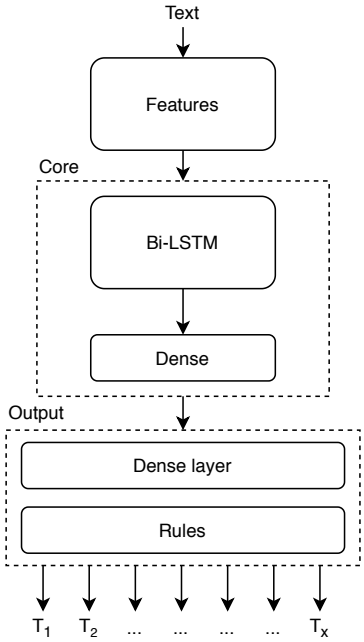


Figure 4-4: Exploring other entities: NER model template.

To address both tasks we proposed the study of supervised approaches based on a common architecture. Figure 4-4 shows the proposed template. Proposed approaches shared a common element (core), composed by Bi-LSTM + Dense tuple. The best systems presented to DIANN task were based on Bi-LSTM models. We proposed a model based on this architecture since these models demonstrated good performance

in many other NLP tasks. Although in DIANN task the use of Bi-LSTM + CRF was considered, we preferred to study the performance of the Bi-LSTM separately. We configured each LSTM with 150 neurons and ReLU as activation function. Seeking to reduce the dimensionality of the space generated by the Bi-LSTM, we included a dense network of 50 dimensions between the Bi-LSTM and the system output. On the other hand, we configured aspects such as the data model and the configuration of the output layer, depending on the task to be addressed.

Exploring different features

Words We used a representation based on the Word Embeddings generated by Cardellino [19] due to the richness of the sources from which they were generated and to their high recall. These vectors have a total of 300 dimensions and gather around 1,000,653 unique tokens. Although in DIANN task, some participants such as Goenaga et al. [58], proposed using Word Embeddings generated with medical health records, we did not have access to this information. For this reason, we used general domain Word Embeddings.

Casing During the experimentation carried out on the RDD corpus, we highlighted the need to represent the information possibly destroyed by the process of conversion to lowercase. With this representation we achieved this purpose, and also ensured an auxiliary representation to deal with elements not recognized by the Word Embeddings. Vecino and Padró [152] and Medina et al. [106], among others, also identified this feature as interesting in DIANN task. We used a single one-hot vector using the Algorithm 4.1. In addition to the conditions covered by this algorithm; for each task, additional casing rules were used to test the usefulness of this representation.

Part-of-speech We used this feature due to its importance in different natural language processing tasks. Using an auxiliary representation based on this feature gave us the opportunity to reduce the vocabulary considered by the Word Embedding, categorizing each term with its corresponding grammatical category. Since many of the errors detected during the DIANN task were

related to the length of the annotations, we considered this feature useful to generate annotations with a correct span. We used the PoS-Tagging model provided by the CoreNLP library for Spanish. We represented this feature using embeddings generated during training. The resulting embeddings consisted of 25 dimensions.

Chars Following the same motivation for the use of casing information, we used an auxiliary representation based on character embeddings. In order to avoid introducing language dependent aspects in previous stages to the learning model, we did not use a complex system of tokenization, i.e. we tokenized only by spaces and ,/./:/; at the end of the word. The use of a character-based representation allowed to mitigate representation problems derived from this tokenization process. On the other hand, given that we used embeddings not specific from the biomedical domain, some terms may benefit from an auxiliary representation. These terms may not be correctly represented in the embeddings, either due to their low frequency of appearance, or due to the contexts in which they appeared in the sources used to generate the embeddings. We used character embeddings generated during the training, applying a convolutional approach based on Zhang et al. [168] to process them.

While for MEDDOCAN task we used words, casing, PoS tags and characters, we decided not to use character embeddings for eHealth-KD task. Using character embeddings, we tried to minimize recall problems derived from processing documents with typing errors. In addition, for the eHealth-KD task we preferred to focus on features such as PoS given their similarity to the entities analyzed by this task.

Exploring output layer

We proposed for both tasks an output based on the tuple Dense + rules system. On the one hand, the number of neurons of the Dense layer depended on the number of classes to be studied. We used the BILOU format to represent the NER annotations. BILOU adds to the IOB annotation scheme a label to represent the last word of a

multi-term annotation (L) and another for mono-term entities (U). We preferred this annotation scheme to the IOB in order to simplify later stages of post-processing.

On the other hand, we applied to the output a post-processing phase based on the application of manually generated rules. This approach covered the correction of systematic errors detected during the development phase and allowed us to establish clear patterns for easily identifiable entity types (e.g. emails). Some considered rules aim to ensure that the final output of the system correctly follows the output BILOU format. Equations 4.3 and 4.4 show examples of the applied rules:

$$T_1(O)\ T_2(B|Action)\ T_3(L|Concept) \Rightarrow T_1(O)\ T_2(B|Action)\ T_3(L|Action) \quad (4.3)$$

$$T_1(O)\ T_2(I)\ T_3(L) \Rightarrow T_1(O)\ T_2(B)\ T_3(L) \quad (4.4)$$

Equation 4.3 shows term T_1, labeled as not belonging to an entity (O), term T_2, labeled as the beginning term (B) of an Action entity, and term T_3 labeled as the last term (L) of a Concept entity. In this case, the applied rule transforms the last entity type from Concept to Action, for it to match the type of entity beginning in T_2.

Equation 4.4, on the other hand, adapts the output to the expected BILOU format: term T_2 is labeled as intermediate term (I) of an entity, while the previous term T_1 is neither an intermediate nor a beginning term. The following term T_3 is the last term of the entity. Hence, for the output to make sense, term T_2 must be relabeled as beginning term, so the final entity is composed of T_2 and T_3.

In addition to the above rules we considered others depending on the task:

Rules applied to eHealth-KD: We studied a set of rules to deal with frequently detected errors related to the scope of identified keyphrases. We modified the scope of these keyphrases using rules based on PoS-tag and casing information.

Rules applied to MEDDOCAN: Some errors detected were related to entity types that are easily handled using regular expressions, for example, telephone numbers (sequence of nine numbers) or e-mails (expression that must contain @). We considered these types of entities using regular expressions.

4.3.1 MEDDOCAN: Results & Analysis

The organizers of MEDDOCAN task proposed two evaluation scenarios. The first scenario (Task 1), more similar to named entity recognition, required the correct identification of an entity (span and type) (e.g. patient names, telephones, addresses, etc.). The second track (Task 2) was focused on the detection of sensitive text more focused on a practical scenario, where the only relevant aspect is the identification of PHI and not its correct classification. For Task 2, the organizers studied two evaluation criteria, one considering a stricter evaluation of the span (Strict) and the other one considering unions of protected health information connected by non-alphanumeric characters (Merged).

A total of 18 teams participated in the track, submitting a total of 63 systems for Task 1 and 61 systems for Task 2. In order to contextualize the results obtained by our system, Table 4.4 shows the results of the best approaches. Although it is not our intention to analyze each proposal in detail, the results obtained and the similarities between our approach and the other ones were useful to validate some aspects considered by our system.

	Task 1			**Task 2 (Strict)**			**Task 2 (Merged)**		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Lange et al. [90]	0.969	0.969	0.969	0.975	0.974	0.974	0.987	0.983	0.985
Hassan et al. [68]	0.969	0.956	0.963	0.975	0.962	0.968	0.981	0.968	0.975
Pérez et al. [126]	0.964	0.956	0.960	0.971	0.964	0.967	0.979	0.972	0.975
León [95]	0.958	0.960	0.959	0.963	0.965	0.964	0.966	0.969	0.968
Jabreel et al. [75]	0.959	0.956	0.958	0.967	0.964	0.966	0.974	0.974	0.974
Our system	0.959	0.928	0.943	0.964	0.934	0.949	0.973	0.942	0.957
Baseline	0.370	0.503	0.426	0.441	0.506	0.471	0.505	0.513	0.509
Mean	0.902	0.893	0.894	0.929	0.910	0.917	0.946	0.924	0.933

Table 4.4: MEDDOCAN: Results obtained by the best systems. F1 stands for *f-measure*, P for *precision* and R for *recall*. The results are sorted by *Task1: F1*.

Analyzing the results of Task 1, our approach obtained similar results to those achieved by the best systems. If we analyze the approaches proposed by the partic-

ipants, we found that many of them reached similar conclusions regarding the use of some approaches. For example, the use of character embeddings to increase the recall obtained by Word Embeddings was identified as useful by several participants. However, while we proposed a convolutional approach to its processing, others such as Lange et al. [90], proposed an approach based on Bi-LSTM. On the other hand, Bi-LSTM also was proposed by several teams as a central processing element; Lange et al. [90] and Jabreel et al. [75] suggested approaches based on the use of Bi-LSTM. In contrast to us, they connected the output of the Bi-LSTM to a CRF for calculating the probability of each NER tag. The use of CRF was a common element in different participating systems, including Hassan et al. [68] and Pérez et al. [126].

Development Set									
	Task 1			**Task 2 (Strict)**			**Task 2 (Merged)**		
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
No Rules	0.933	0.911	0.922	0.938	0.916	0.927	0.948	0.925	0.936
With Rules	0.964	0.930	0.947	0.968	0.934	0.951	0.976	0.941	0.959

Table 4.5: MEDDOCAN: Analysis of the impact of the proposed rules in our system - Evaluation using development set. F1 stands for *f-measure*, P for *precision* and R for *recall*.

Finally, approaches using rules systems were considered by different teams. Studying the development set, Table 4.5 shows the impact detected on the performance considering or not, the rules. In this sense, the approach proposed by León [95] was particularly interesting. They proposed a system based only on the use of external resources and the application of nearly 100 manually developed rules. Analyzing the results obtained by each team in the different tasks, this system achieved very consistent and competitive results.

4.3.2 eHealth-KD challenge: Results & Analysis

Since the task covered both, entity recognition and relationship extraction, the organizers proposed different scenarios to evaluate the participating systems. Only

the first scenario contemplated the evaluation of NER systems in isolation. The second scenario studied the behavior of relationship extraction systems using gold NER annotations. The third scenario studied the whole pipeline.

A total of 10 teams participated in this task and Table 4.6 shows the results of the best proposed approaches. In the same way as the analysis of MEDDOCAN task, it is not our intention to analyze each proposal in detail. The results obtained and the similarities between our approach and the other ones allowed us to validate some considered aspects. In this task, considering all the presented approaches, we detected limitations associated with the method proposed by us to process the characters.

Team	**F1**	**P**	**R**
Medina and Turmo [105]	0.8203	0.8073	0.8336
Bravo et al. [16]	0.8167	0.7997	0.8344
Alvarado et al. [5]	0.8156	0.7999	0.8320
Ruiz-de-laCuadra et al. [138]	0.7903	0.7706	0.8111
Our system	0.7543	0.8069	0.7082
Baseline	0.5466	0.5129	0.5851
Mean	0.7334	0.7270	0.7424

Table 4.6: eHealth-KD challenge 2019 - Scenario 2 (Subtask A): Best results. F1 stands for *f-measure*, P for *precision* and R for *recall*. The results are sorted by *F1*.

Although the approaches proposed in this task were very homogeneous, the proposal presented by Medina and Turmo [105] was particularly interesting. They used a single approach to jointly study keyphrase detection and relationship extraction. In summary, among the best presented approaches, the use of Bi-LSTM and CRF as core processing elements was a common attribute. The use of a wide variety of data models was an outstanding aspect of the analyzed proposals. As we proposed, Alvarado et al. [5] and Ruiz-de-laCuadra et al. [138] used Word Embeddings and characters embeddings. To process the characters, they used Bi-LSTM-based approaches. Alvarado et al. [5] proposed a comparison between the use of a convolutional approach or an approach based on a Bi-LSTM for processing characters (convolutional approach F1: 0.7483 / bi-lstm approach F1: 0.8156). On the other hand, representations based

on contextualized embeddings were used by Medina and Turmo [105] (BERT) and Bravo et al. [16] (ELMO). As a consequence of the use of these embeddings, they had to deal with a high dimensionality of the input space which made it difficult to study other features in depth (BERT: 768 and ELMO: 1024 dimensions).

4.3.3 Discussion

In order to enhance the set of conclusions we drew during the DIANN task by covering more generic aspects of the biomedical domain, we participated in two evaluation tasks focused on the processing of other biomedical documents and entities. Having analyzed the identification of disabilities through different points of view, we thought that an analysis of tasks covering more generic aspects than those addressed in previous experiments would help us to analyze aspects to be proposed in the development of a system oriented to the recognition of disabilities and rare diseases. We address both considered tasks using a common architecture based on a supervised approach. We used some aspects considered as relevant by the participants of the DIANN task. In both tasks, the approaches presented by the other participants presented many similarities with our developed approach. This supported our initial belief about the utility of each feature and aspect considered in the model. Among the participants was quite common the use of Word Embeddings, character embeddings, PoS tag information and rules, as well as the use of Bi-LSTM to process them. Regarding the obtained results, the systems we proposed showed results above the mean, although not surpassing the results obtained by the best participants. In both tasks, the obtained *recall* suggested to reconsider the approach used to represent aspects such as the characters (Convolutional approach ⇒ Bi-LSTM) or the output layer (Dense ⇒ CRF).

4.4 Exploring lessons learned: DIANN Corpus

Having analyzed different collections, in this section we present the experiments carried out for the task of recognizing named disabilities, taking into account the lessons learned. In this chapter we continue the analysis of the DIANN corpus [45].

The experiments described in this section are published in Fabregat et al. [50].

We focused the experiments on supervised approaches. The following techniques and representation methods were studied:

- **Pre-trained Word embeddings (W).** The experiments were carried out using Glove [125] for both languages and working with the 200-dimensions model.

- **Char embeddings (Ch).** Character embeddings proved to be a very interesting tool in the different experiments we carried out. Due to the reduction of dimensionality caused by the use of Word Embeddings, some systems use this representation to reduce the loss of information. We reconsidered the proposed model, and used an LSTM-based approach.

- **Brown cluster (B).** Cluster-based representations are very useful to support generalization capabilities of Word Embeddings. Brown clustering is a form of hierarchical clustering based on the premise that expressions that occur in similar contexts could be semantically related [17]. To group different terms under the same class, Brown clustering uses a binary merging criterion based on the log-probability of a text under a class-based linguistic model. We represented this feature using an one-hot vector.

- **Part-Of-Speech embeddings (P).** We considered Part-of-speech tagging as an element to be analyzed, since most of the complexity of this task consists on identifying each disability with its severity and/or temporal qualifiers.

- **Casing vector (C).** This vector encodes format information, in some cases omitted by Word Embeddings, e.g.:
  Expression: 01234AB $->$ *Casing Label:* mainly numerical expression.
  Expression: Mentally $->$ *Casing Label:* uppercase expression.

- **Attention (A).** We used the attention method proposed by [3] to consider the filtering mechanism suggested by the SINAI team. This attention method weights each term according to its relation with some recurrent terms. Although

we carried out some experiments with different sets of terms, we obtained the best results with the average of the terms “disability” and “handicap”.

Figure 4-5 shows the architecture studied by us. In line with Goenaga et al. [58] and Zavala et al. [165], and considering the experiences we gathered in the MEDDOCAN and e-Health tasks, we analyzed a model based on Bi-LSTM and CRF.

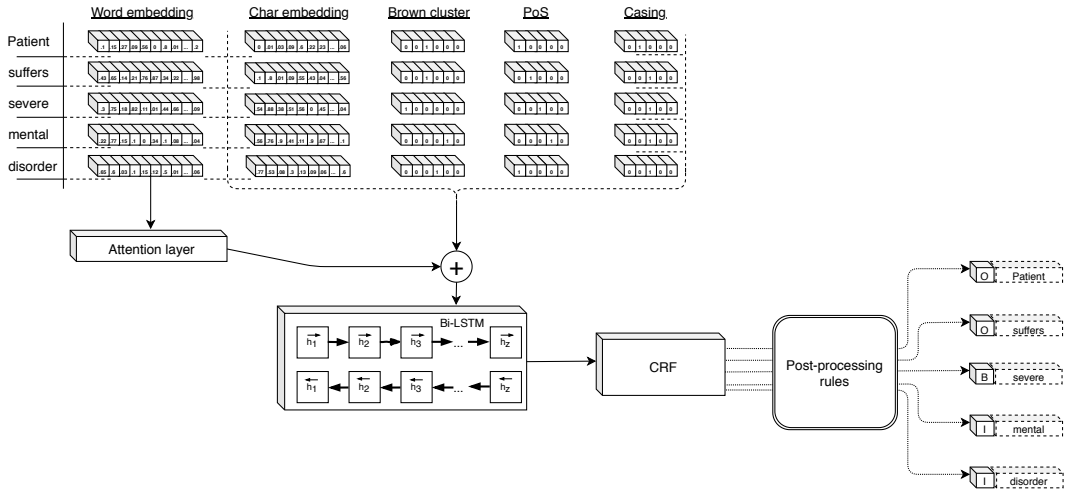


Figure 4-5: Extending DIANN Shared Task: Proposed model including its inputs and layers. The post-processing rules are applied to the output of the deep learning model.

The proposed model used a Bi-LSTM as core element. After analyzing different configurations, the size of the main Bi-LSTM is similar to the size reported by the IxaMed and UC3M teams, 150 neurons in the output layer. To process the characters, instead of using a convolutional approach such as the architecture applied by Fabregat et al. [47], given the good results achieved by Zavala et al. [165] and Lange et al. [90], we opted for a representation based on Bi-LSTM with 50 neurons. Finally, we used a CRF to process the concatenation of the output of both Bi-LSTMs.

Given the small size of the DIANN corpus, we trained the model during 50 epochs using a learning rate of 0.01 and small batches (16). We considered an early stop function based on the loss score. In order to avoid over-fitting we used dropouts of

0.25 between processing layers. In addition, as proposed by Goenaga et al. [58], we revisited the analysis of post-processing rules. Specifically, we implemented a set of rules to detect abbreviations (inspired by [114]) and to process special cases detected in the training set. On the other hand, some of the proposed rules were inspired in Fabregat et al. [47], e.g. the rules to process enumerations:

- If an annotation contains a statement such as: "<dis>cognitive delay/mental disability</dis>", then it is divided into "<dis>cognitive delay</dis>" and "<dis>mental disability</dis>".

- If an annotation contains a statement such as: "severe or <dis>moderate loss of vision</dis>", then it is expanded to "<dis>severe or moderate loss of vision</dis>".

4.4.1 Results

	SPANISH			ENGLISH		
	Precision	Recall	F1	Precision	Recall	F1
W+Ch+C	**0.79 (0.90)**	**0.69 (0.79)**	**0.74 (0.84)**	0.81 (0.89)	0.68 (0.75)	0.74 (0.81)
W+Ch+C+A	0.78 (0.87)	0.67 (0.75)	0.72 (0.80)	0.8 (0.86)	0.74 (0.80)	0.77 (0.83)
W+Ch+P+C+B	0.79 (0.86)	0.70 (0.76)	0.74 (0.81)	0.83 (0.88)	0.74 (0.78)	0.78 (0.83)
W+Ch+P+C+A	0.80 (0.86)	0.67 (0.73)	0.73 (0.79)	**0.86 (0.92)**	**0.74 (0.79)**	**0.79 (0.85)**

Table 4.7: Improving results of DIANN Shared Task: Exact and partial (in brackets) evaluation results before applying post-processing rules. Best results for each language are shown in boldface. (W) Word embeddings - (Ch) Char embeddings - (C) Casing - (B) Brown clusters - (A) Attention.

Table 4.7 shows the obtained results before applying post-processing rules for both languages. We indexed the results according to the used features. These results show a remarkable difference in performance between exact and partial evaluation. In many cases, this difference may be explained as a result of the difficulties faced to identify contextual or modifying elements related to labeled disabilities. For Spanish, the best results were obtained using characters and casing information, whereas for

English the best results were achieved including part-of-speech tags and the attention mechanism. Limitations of the Part-of-Speech model used for Spanish may be the cause of this discrepancy.

	SPANISH			ENGLISH		
	Precision	Recall	F1	Precision	Recall	F1
Goenaga et al. [58]	0.75 (0.82)	0.81 (0.88)	0.78 (0.85)	0.78 (0.84)	0.86 (0.92)	0.82 (0.88)
Zavala et al. [165]	0.8 (0.88)	0.65 (0.71)	0.71 (0.79)	0.77 (0.82)	0.72 (0.76)	0.74 (0.79)
W+Ch+C	**0.83 (0.91)**	**0.79 (0.87)**	**0.81 (0.89)**	0.8 (0.90)	0.77 (0.87)	0.78 (0.89)
W+Ch+C+A	0.78 (0.86)	0.78 (0.86)	0.78 (0.86)	0.78 (0.87)	0.79 (0.88)	0.79 (0.88)
W+Ch+P+C+B	0.78 (0.85)	0.79 (0.86)	0.79 (0.85)	0.81 (0.88)	0.81 (0.89)	0.81 (0.88)
W+Ch+P+C+A	0.79 (0.85)	0.76 (0.82)	0.77 (0.84)	**0.84 (0.92)**	**0.81 (0.89)**	**0.83 (0.90)**

Table 4.8: Improving results of DIANN Shared Task: Results achieved after the application of post-processing rules and comparison with state-of-the-art systems. Exact and partial (in brackets) evaluation results. Best results for each language are shown in boldface. (W) Word embeddings - (Ch) Char embeddings - (C) Casing - (B) Brown clusters - (A) Attention.

After applying the post-processing rules we obtained notable improvements in both languages, being more evident in Spanish (Table 4.8). On the other hand, although the obtained results show clear improvements, the high results of *recall* obtained by Goenaga et al. [58] stand out. These results could be a consequence of the use of specific embeddings from the biomedical domain. In summary, the proposed system improves the results obtained by the best participating system using a set of features suggested during the task.

4.4.2 Discussion

Given the trend in medical documents to use abbreviations, elements focused on the processing of this type of entities are of great interest. We improved the identification of abbreviations considering methods such as casing or character sequence processing. The model used to represent character sequences was quite versatile, allowing both, the representation of n-grams and the representation of terms not included in the Word Embeddings. On the other hand, disabilities can be expressed in countless ways.

While the use of clustering techniques enhanced the identification of semantically similar terms, the use of the attention model contributed to reduce the number of false positives. However, the small size of the corpus affected the effectiveness of this mechanism. Improvements using this method were only achieved in English documents. Finally, the use of post-processing rules, focusing on the improvement of the exact matching evaluation results, was highly effective. Most of the implemented rules were useful to deal with enumerations and abbreviations, improving the previously achieved results of *recall*. Concerning state-of-the-art systems, Goenaga et al. [59] used both custom embeddings (generated with medical reports) and ad-hoc post-processing rules to deal with the recognition of abbreviations, which would justify the achieved results of *recall*.

4.5 Conclusions

During the development of this thesis we explored different approaches and techniques for the recognition of disabilities in documents of the biomedical domain. Analyzing unsupervised approaches, we used DIANN task to analyze our proposals as well as other systems developed and evaluated in the same context by other participating researchers. Considering the benefits of unsupervised approaches in the biomedical domain, the approaches raised during the DIANN task obtained interesting results. Although approaches proposed by other teams using filtering processes based on Word Embeddings did not achieve satisfactory results, other approaches based on variant generation obtained results that suggest not to discard the exploration of unsupervised techniques for this task. Reviewing the annotations generated by our system, we detected several errors derived from the external resources we used, especially for Spanish documents.

On the other hand, among all the supervised models that we studied, the good performance of systems based on deep learning stood out. During our research using the RDD corpus, we found that models based on recurrent neural networks help to consider easily contextual elements in the processing of data sequences. However, although models based on machine learning require an effort to develop and build new features, they provide more flexibility and generate more understandable

schemes. This first study served us to analyze in a preliminary stage the development of supervised systems. Afterwards, with the development and evaluation of the DIANN task, in addition to our unsupervised proposal, we had the opportunity to study the supervised approaches proposed by the participating teams. One of the conclusions we drew from this meeting was the good performance of the systems based on Bi-LSTM+CRF. This is a conclusion widely found in current named entity recognition researches. In addition, approaches using information extracted from characters, n-grams and word clusters obtained results evidencing the usefulness of this information for the study of entity recognition. We also highlighted the need to address the detection of disabilities and acronyms in a two-stage approach, given the good performance obtained with rule-based systems.

After the organization and participation in DIANN task, we addressed the validation of different aspects proposed during the task by participating in two different evaluation tasks. Our intention was to deepen the analysis carried out on the DIANN task, covering general aspects of the biomedical domain and not only those related to disabilities. We proposed a common architecture for both tasks and we studied the validity of a system based on Bi-LSTM and rules. The results that we obtained in both tasks helped us to validate and redesign different configurations of the architecture. On the one hand, having identified the use of character embeddings as a relevant element, we applied a convolutional approach obtaining worse results than other teams that used a Bi-LSTM approach. On the other hand, the use of rules was a necessary element to cover certain entity types and some aspects related to the BILOU annotation scheme.

Finally we addressed the generation of a system that will effectively address the identification of disabilities in English and Spanish documents. We proposed different experiments using the characteristics analyzed during the thesis. Among the aspects that we gathered were the use of clusters and an attention model based on the weighting of each analyzed term according to a reference set of terms. While the use of clusters had similar implications in both languages, the analyzed attention model caused improvements for English and not for Spanish. Since this attention model used the embeddings of the term disability (discapacidad for Spanish) to contrast its similarity with the representations of each term of a sentence, ignoring

embeddings of synonyms and terminological variants, the different performance of this model may be caused by the great variety of commonly used synonyms of the term discapacidad (Spanish) in comparison with the ones of disability (English). In contrast to the clusters, which were generated using information extracted from the training collections, the attention model we proposed was linked to the Word Embeddings. In short, processing Spanish documents, we obtained the best performance using character embeddings, casing information, Word Embeddings and rules. For English, we obtained the best results using Part-of-Speech information and the attention model, plus the features we proposed for Spanish.

Chapter 5

Relationships extraction

Contents

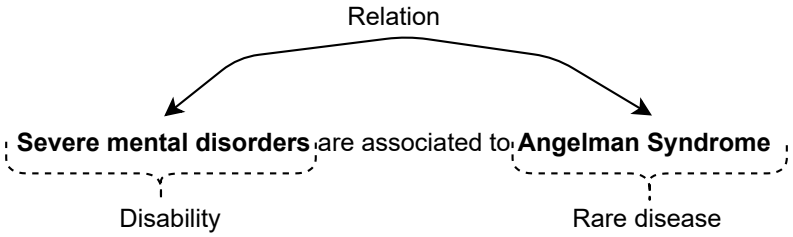
After analyzing in the previous chapter all the experiments on named entity recognition (NER), in this chapter we analyze our experiments on the extraction of relationships (RE) between biomedical entities. Given the links between the tasks of named entity recognition and relationship extraction and, in contrast to the previous chapter in which the detection of named entities was studied as a single and isolated task, this chapter presents a joint exploration of named entity recognition and relationship extraction approaches. On the one hand, in addition to the results obtained on relationship extraction, we explored the performance of the proposed systems within a non-ideal context where the previous detection of entities is not a perfect process. Our objective is to measure the degree of detriment that the imperfection of NER processes can inflict on the performance of subsequent systems (RE). On the other hand, covering the tasks of named entity recognition and relationship extraction under the same umbrella, we want to analyze the possible synergies between both tasks.

About the organization of this chapter, in Section 5.1 we present our experiments about relationship extraction techniques applied to the identification of relationships between rare diseases and disabilities. This section reports on the application of different methodologies for the complete processing of the relationships and entities contained in the RDD corpus. We followed an incremental and comparative approach in order to thoroughly analyze the proposed methods and their improvements and disadvantages with the other proposed alternatives. To conclude this section, we present an analytical comparison of the results reached on the study of the relationship between disabilities and rare diseases. Furthermore, exploiting the opportunity provided by the eHealth-KD Challenge 2019 for the analysis of other types of entities and relationships, in Section 5.1.3 we present our experiments carried out for this evaluation task. In these experiments we tried to explore some alternatives trying to focus on the generalizability of the proposed models.

5.1 Disabilities and Rare diseases relationships

Rare diseases are a specific type of disease characterized by its low population incidence. Throughout Chapter 3 we discussed some details about the potential

relationships between rare diseases and disabilities. Using the RDD corpus, we explored the extraction of relationships as a binary classification of the tuple (Disability + Rare Disease + Sentence) as related or not. We addressed this task as a binary classification because, although the RDD corpus collects different types of relationships (co-occurrence and cause-effect), it does not collect them in a differentiated way (details in Section 3.2.2). The Example 5.1 shows a sentence extracted from the RDD corpus where we can identify a relationship between a disability ("Severe mental disorders") and a rare disease ("Angelman Syndrome").



Example 5.1: RDD corpus: Relationship between a rare disease and a disability.

In order to establish a methodological link between the explored approaches for named entity recognition and relationship extraction, we used for both tasks the files compiling the positive and negative relationships included in the RDD corpus. This framework allowed us to analyze the explored proposals using different methodologies. Firstly, in Subsection 5.1.1, we present a preliminary study of different proposals to process the relations collected in the RDD corpus. With this first study we tried to establish a sufficient basis to motivate the rest of the experiments proposed in this context. After analyzing the results of this first study, in the Subsection 5.1.2 we continue the exploration of deep learning models using more refined approaches. In this section, we explore the performance of the proposed systems in an ideal context and also as part of a natural language processing pipeline. Finally, in Subsection 5.1.3 we explore the performance of a joint approach for named entity recognition and relationship extraction.

5.1.1 RDD Corpus: Relation Extraction - Preliminary study

Our first effort to analyze this task came hand in hand with the publication and release of the RDD corpus [44]. In this work, we tried to establish a benchmark for exploring models based on deep learning and machine learning. On the one hand, we analyzed approaches based on support vector machines. Among other features, we studied named entities using a representation based on lemmatization and bag of words. We also explored Part-of-Speech tags, the presence of negation and/or speculation, the length of each entity and the distance or the possible overlapping between the entities. We configured the explored SVM models using SMAC (Sequential Model-based Algorithm Configuration [73]) and exploring Sigmoid and RBF kernels (Equations 5.1 and 5.2) considering the following spaces: Sigmoid : γ ={0.0001, 1} ; ρ ={0, 5} | RBF : γ ={0.0001, 1}.

$$Sigmoid = \tanh(\gamma(x, x') + \rho) \tag{5.1}$$

$$RBF = \exp(-\gamma \parallel x - x' \parallel^2) \tag{5.2}$$

On the other hand, we analyzed several deep learning architectures based on CNN and LSTM. These approaches have been successfully applied to different NLP tasks in the literature. We used as features the concatenation of lexical information provided by Word Embeddings and sentence level information provided by the position of the entities (disabilities and diseases) within the sentence. For each entity involved in the relationship to be classified, we generated a feature representing the absolute distance of each term with the corresponding entity (Figure 5-1). We encoded this information using embeddings calculated during training. We processed the concatenation of both features using either a two-layer convolutional architecture or one model based on a single LSTM network. For the LSTM model, we used the output of the last neuron to summarize the information of each sentence in a single element. We used a bidirectional LSTM network of 200 neurons (100 per direction). Concerning the convolutional model, we used a max pooling layer to reduce the dimensionality and a linear rectifier as activation function [166]. Finally, in both proposed models we transformed the output of the last layer into the probability of having found (or not)

a relationship in the sentence using a softmax activation layer.

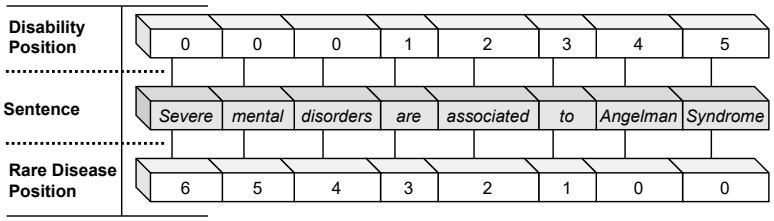


Figure 5-1: Relationship extraction: Generation of vectors capturing the information related to the position of each entity being part of a relationship. In the example, Disability: *"Severe mental disorders"* and Rare Disease: *"Angelman Syndrome"*.

Results & Analysis

We evaluated the experiments carried out using a 10-fold cross-validation. To test the statistical significance of differences in the results achieved by different methods and configurations, we used a paired t-test ($p < \alpha : \alpha = 0.05$). Table 5.1 compares the results of the SVM, and those obtained by the deep learning models. Considering the results of *f-measure* as a summary measure, the deep learning algorithms achieved the best results. We found statistically significant the differences of *precision* results between the CNN-based model and the other ones. Taking into account the results obtained by SVM, we can not rule out these approaches a priori (especially comparing the results obtained by LSTM approaches). However, the considerable feature engineering effort involved in the use of these machine learning methods is a major point to consider. In this respect, approaches based on deep learning offer greater versatility. The performance achieved by deep learning approaches is similar to that obtained during the study of named entity recognition techniques on the RDD Corpus (Section 4.1.1). The use of CNN, in contrast to LSTM, represented interesting improvements. During all the experiments carried out we used a representation of the entities extracted from the gold, i.e. we did not propagate the possible errors of the entity detection phase. Considering approaches based on a pipeline of entity recognition and relationship extraction, the rigidity of this representation could imply a significant loss of effectiveness by establishing a strong dependence of the

performance on a previous scenario.

	Precision (sd)	**Recall (sd)**	**F-measure**
SVM	0.706 (0.0388)	0.709 (0.0813)	0.707
CNN (Words + Pos)	**0.757** (0.0627)	**0.759** (0.0514)	**0.758**
LSTM (Words + Pos)	0.700 (0.057)	0.726 (0.0503)	0.713

Table 5.1: RDD Corpus: Results of *precision*, *recall* and *f-measure* obtained identifying relationships between Rare Diseases and Disabilities (SD: Standard deviation). SVM and Deep Learning models. In bold the best results and underlined statistically significant differences ($p < \alpha : \alpha = 0.05$).

In addition, extending the analysis of the results presented in Table 5.1, we performed a set of additional experiments to test the possibility of classifying the different instances without using gold information about the entities. We transferred the scope of the study from the extraction of relationships between pairs of entities given a sentence (*Disability, Rare Disease, Sentence*) to the identifications of sentences mentioning at least one relationship. We analyzed the performance of different models only using word embedding information. The results we obtained in these experiments seem to indicate a better representation of possible relationships using LSTM-based approaches (CNN *f-measure*: 0.562; LSTM *f-measure*: 0.624).

5.1.2 RDD Corpus: Extending the analysis

The results obtained for the extraction of relationships during the preliminary study [44] encouraged us to extend the exploration of deep learning architectures. As a result of this exploratory process we developed the model illustrated in Figure 5-2.

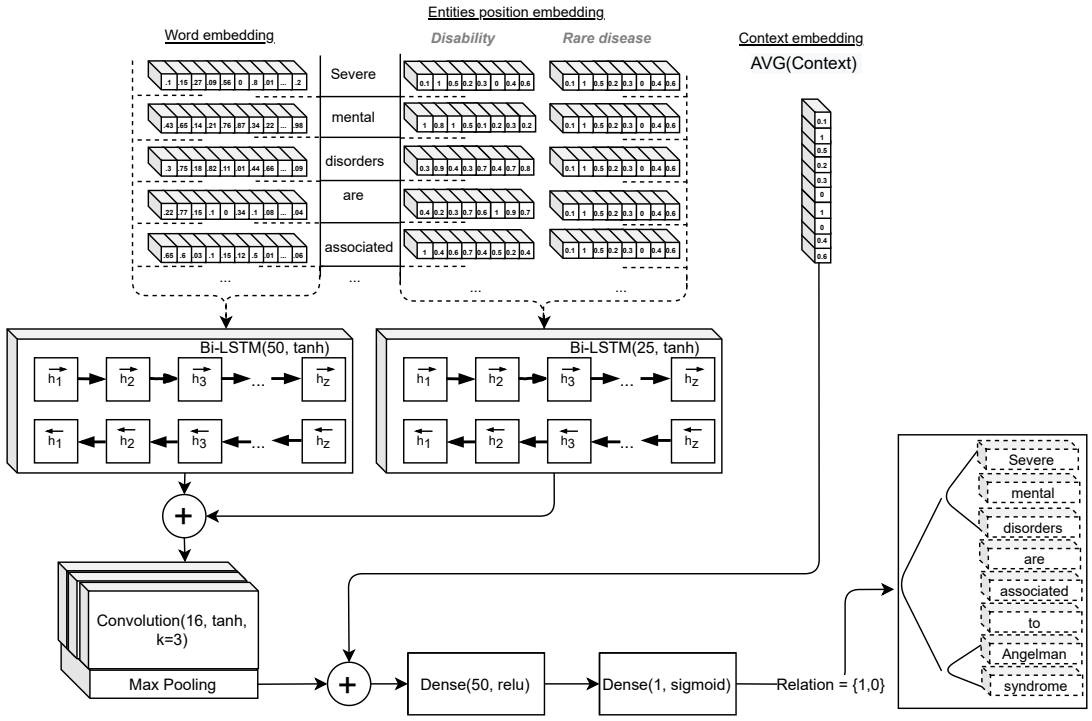


Figure 5-2: RDD Corpus: Deep learning model for relationship extraction. The model includes two neural stacks. (1) two Bi-LSTM (50 and 25 neurons) and a convolutional network (16 neurons, kernel=3) to process the inputs of more than two dimensions (Word Embeddings and position embeddings). (2) two densely connected networks (50 and 1 neuron/s) to process the concatenation of the pooling of (1) with the two-dimensional inputs (summarized context embeddings).

In addition to the features studied in previous experiments, this model added a vector representing the average value of the embeddings of the words between the two studied entities. In Sample 5.1, where it mentions a relationship between "Severe mental disorders" and "Angelman Syndrome", the context is "are associated to", and the feature vector is the average of the embeddings of each term that compose this expression. To model this set of features we considered two neural stacks: The first stack consists of a structure based on the Bi-LSTM+CNN scheme. We used two Bi-LSTMs to process the word embeddings and entity positions in parallel. After the

parallel application of these networks, we used a convolutional network to process the concatenation of the generated spaces. Trying to extract the components of greater weight, we applied a discretization of the generated subspace through the application of a max pooling layer. The second stack consists of two densely connected neural networks that process the concatenation of the output of the first stack with the vector representing the context between the entities under study. The output of this stack corresponds to the output of a binary classifier. Figure 5-2 shows the configurations used for the construction of the model.

Results & Analysis

We trained the proposed model during 20 epochs using Adam optimizer [83] with default parameters and we evaluated its performance using a 10-fold validation. As we did in previous experiments where we evaluated the performance of deep learning algorithms using cross-validation. Table 5.2 shows the results obtained exploring different Word Embeddings and, Table 5.3 shows the results of the experiments carried out exploring the set of features and using the best Word Embedding selected previously. In both cases, the conclusions obtained after using a paired t-test ($p < \alpha : \alpha = 0.05$) are also reported to highlight the statistically significant differences between the best configurations and the other ones.

	Precision (sd)	**Recall (sd)**	**F-measure**
Glove50d	<u>0.756</u> (0.0526)	0.75 (0.0682)	0.753
Glove100d	**0.81** (0.0287)	**0.799** (0.0345)	**0.804**
Glove200d	0.808 (0.0321)	0.789 (0.0328)	0.799
FastText	0.78 (0.0382)	0.761 (0.0532)	0.770

Table 5.2: Analysis of different embeddings: Results achieved by the model illustrated in the Figure 5-2. Metrics: *precision*, *recall* and *f-measure* (SD: Standard deviation). The best results are shown in bold, and the statistically significant differences in respect of the best results are underlined ($p < \alpha : \alpha = 0.05$).

Glove100d	**Precision (sd)**	**Recall (sd)**	**F-measure**
Words+POS+Context	**0.81** (0.0287)	**0.799** (0.0345)	**0.804**
Words+POS	0.719 (0.0633)	0.694 (0.0768)	0.706
POS+Context	0.777 (0.0328)	0.757 (0.0348)	0.767
POS	0.682 (0.0611)	0.678 (0.0735)	0.68

Table 5.3: Analysis of different inputs: Results achieved by exploring the model illustrated in the Figure 5-2 using Glove embeddings of 100 dimensions. Metrics: *precision*, *recall* and *f-measure* (SD: Standard deviation). The best results are shown in bold, and the statistically significant differences in respect of the best results are underlined ($p < \alpha : \alpha = 0.05$).

Firstly, although the results obtained show that the best configuration was obtained using Glove[1] [125] with 100-dimensional embeddings (Glove100d), the differences found when comparing these results with those obtained using Glove200d or FastText [14] are not statistically significant. However, a significant deterioration in *precision* was found when studying the performance of the system using Glove50d embeddings.

Analyzing the performance of the system using Glove100d and switching the set of features, we found that the use of all studied features results in significant improvements in comparison to the use of smaller sets. We found improvements in the detection of relationships depending on the length of the context, defining it as the number of tokens between the entities part of the relationship. Most of the improvements corresponded to relationships with an average context length of 20 tokens.

Analyzing pipeline approach: Detection of relationships using predicted entities

In order to study the particularities of the proposed models in a non-ideal context, we studied a neural information extraction pipeline for named entity detection and

[1]In this chapter we used Glove embeddings based on documents extracted from Wikipedia and Newswire Text Data (Glove6B).

relation extraction. The pipeline used a NER architecture based on the model described in Chapter 4 (Section 4.4 - Figure 4-5). Since the rule-based system used by this NER model does not consider the recognition of rare diseases, we simplified the study of the proposed pipeline by omitting this post-processing phase. Following the detection of named entities, we explored the application of the proposed model for the extraction of relationships (Figure 5-2).

We evaluated the performance of the proposed pipeline using a 10-fold cross-validation process wherein each iteration we trained both systems exploiting the gold-standard annotations. After training both components of the pipeline, we measured the performance using a classical approach, in which we consider that: (1) an entity is correctly identified if all its components are correctly identified; (2) a relationship is correctly identified by the pipeline if it is exactly listed in the gold standard. Thus, by transferring the false positives identified by the entity recognizer to the relation extraction process, this evaluation method establishes an additional penalty on the final results by considering relations in the evaluation process that are not included in the corpus.

Glove100d	**Precision (sd)**	**Recall (sd)**	**F-measure**
NER: Global	0.781 (0.0231)	0.746 (0.0418)	0.763
NER: Disability	0.815 (0.021)	0.806 (0.0344)	0.81
NER: Rare Disease	0.737 (0.0526)	0.669 (0.0861)	0.701
RE: Predicted entities	0.616 (0.0619)	0.495 (0.0693)	0.549
RE: Gold entities	0.81 (0.0287)	0.799 (0.0345)	0.804

Table 5.4: RDD Corpus: Performance of a pipeline for named entity recognition (NER model illustrated in Figure 4-5, without the application of post-processing rules) and relationship extraction (RE model illustrated in Figure 5-2). Metrics: *precision*, *recall* and *f-measure* (SD: Standard deviation).

Table 5.4 shows the results obtained by the NER system, both overall and by type of entity, as well as the performance of the relationship extraction (RE) system using the entities recognized by the NER system. In addition, this table includes

the performance of the system using the entities extracted from the gold standard. Analyzing the disparate performance obtained by recognizing mentions of disabilities and rare diseases, a large part of these differences are caused by the number of annotations that the RDD corpus collects for each entity type. Extending this analysis, the RDD corpus collects a large number of sentences that establish 1 to N relationships between a rare disease and multiple disabilities. As a result, this corpus contains many more disabilities annotated than rare diseases, mainly due to the document retrieval process carried out for the construction of this corpus. Using a list of rare diseases extracted from Orphanet, we performed a document search in scientific repositories and then we selected those documents that also mentioned one or more disabilities. Although this process shifted the role of disabilities to a secondary role, it paradoxically resulted in more annotations on disabilities than on rare diseases. Errors arising from the recognition of these infrequent entities are a significant penalty for the extraction of relationships. Finally, regarding the errors caused by the imperfection of the disability detection process and, as it occurred in previous NER experiments, we evidenced problems in the processing of entities that are excessively long.

5.1.3 RDD Corpus: Joint approach for named entity recognition and relationship extraction

Trying to deepen in different aspects on the detection of relationships between disabilities and rare diseases, we worked with different methodologies. After exploring the performance of different approaches separately, we studied a multi-task approach for named entity recognition and relationship extraction (NER and RE). By studying models based on deep learning, we explored the interconnection of both tasks using a weight-sharing approach between different models [28]. In this context, to analyze an enrichment of the features used to explore the extraction of relationships, we studied the concatenation of intermediate representation spaces obtained by a NER system.

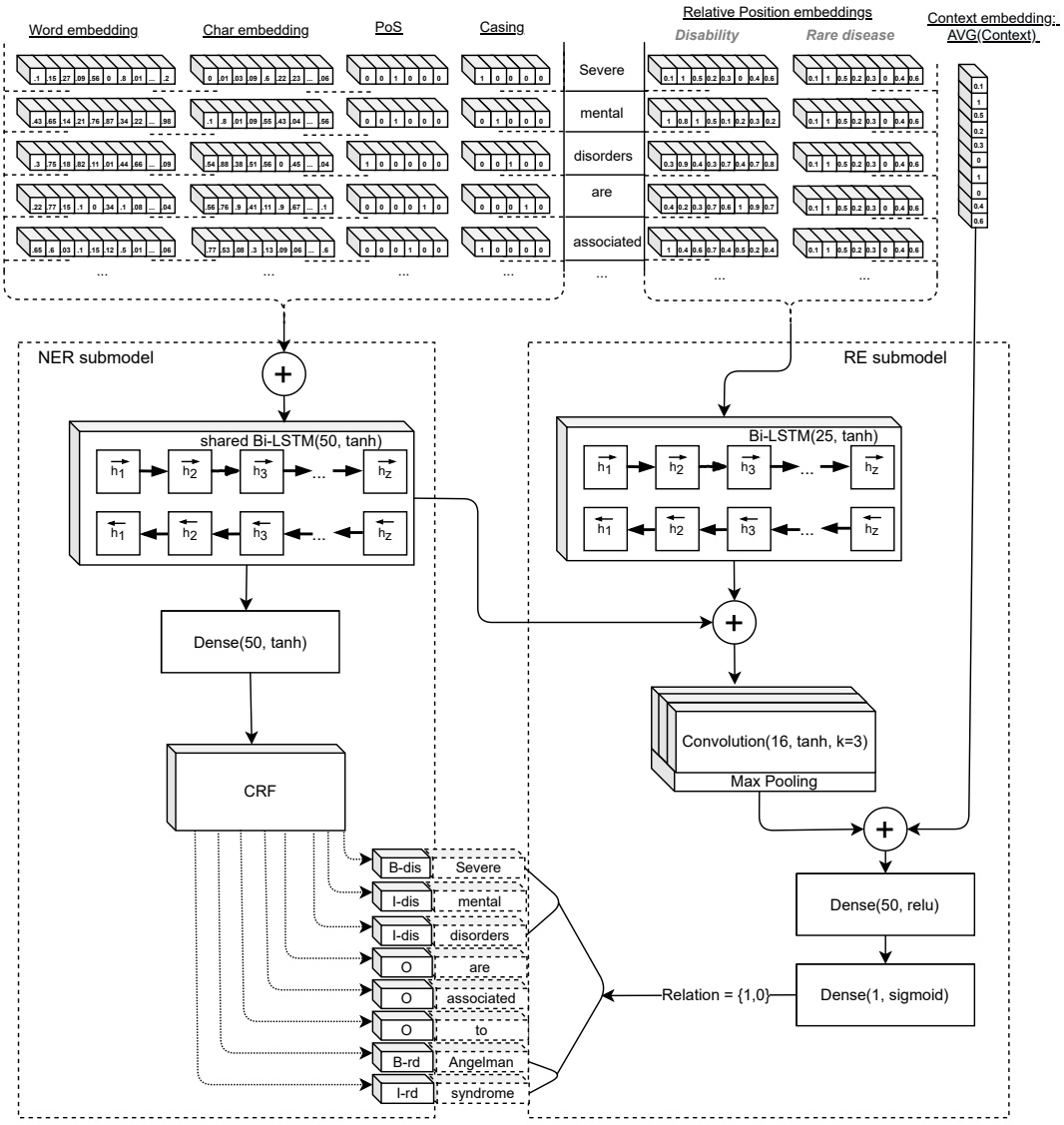


Figure 5-3: RDD Corpus: Deep learning joint model for named entity recognition (NER submodel) and relationship extraction (RE submodel). NER submodel: a Bi-LSTM+CRF based stack to process the concatenation of the inputs (words and characters embeddings, and one-hot vectors representing PoS and casing information). RE submodel: (1) two Bi-LSTMs (one of them shared with NER submodel) and a convolutional network to process the inputs of more than two dimensions. (2) two densely connected networks (50 and 1 neuron/s) to process the concatenation of the pooling of (1) with the two-dimensional inputs (summarized context embeddings).

Figure 5-3 shows the union of the explored subsystems. For the NER model we used an architecture based on Bi-LSTM+CRF and, for the RE model, we used one based on the tuple Bi-LSTM + convolutional network. We connected both submodels by sharing the latent space generated by the Bi-LSTM network of the NER model. We concatenated the output of this layer with the output of the Bi-LSTM of the RE model. Given this nexus and contrasting with other approaches such as the approach proposed in Section 4.4, we included an intermediate transformation of size and range similar to the Bi-LSTM of the NER model. This decision was mainly influenced by the two-stage training proposed to work with this model: (1) NER submodel training; (2) training of the whole model, i.e. including the NER submodel layers trained in (1). Finally, Figure 5-3 also shows the configuration of each layer, i.e. size of the generated space, kernel size and activation function. In addition to this information, Figure 5-3 also shows a list of the features used by each part of the proposed architecture. Firstly, for the NER model we used different types of representations to vectorially map each word of a given sentence. We used pre-trained embeddings to represent words and characters, and one-hot vectors to represent Part-Of-Speech and casing information. During the experimentation we worked with different types and shapes of Word Embeddings. The RE submodel included the processing of other types of features. We provided the definition of these features in Sections 5.1.1 and 5.1.2.

Results & Analysis

We evaluated the performance of the proposed system using a 10-fold cross-validation approach. In each iteration we trained the model and evaluated its performance for each task. Firstly, we trained the NER submodel and then both models together. The results we obtained are shown in Table 5.5. This table shows the different experiments we carried out by changing the model of word embeddings. We explored the embeddings of Glove and FastText. In addition, we carried out several experiments exploring models of Glove of different dimensions. The results presented in the first columns are those obtained in the Step 1 of the training, i.e., the one corresponding to the training of the NER submodel. The results of this process correspond to those discussed in Table 5.4. Secondly, we also present in Table 5.5 the results

obtained at the end of the second training step. We show the results obtained for both entity detection and relationship extraction. We obtained statistically significant improvements on the NER results as a consequence of the re-training of the NER submodel layers at Step 2 of the training. Although the improvements were mostly focused on *recall*, in cases such as FastText we obtained significant improvements in all the analyzed metrics.

The combined study of both tasks does not lead to an improvement in performance in relation extraction using gold entities information, compared to analyzing the tasks independently. Analyzing in detail the detected relationships, we had no improvements in the extraction of relationships with an average context length of more than 20 tokens, however, we identified some improvements in the detection of relationships between long entities. These improvements were possibly a consequence of the improvements obtained in the named entity recognition process.

		Step 1: NER Training			Step 2: Joint Training		
		Prec. (sd)	Rec. (sd)	F-measure	Prec. (sd)	Rec. (sd)	F-measure
Glove50	NER	0.77 (0.0381)	0.729 (0.0613)	0.749	0.769 (0.0197)	*0.778* (0.0249)	0.773
	RE				0.788 (0.0516)	0.779 (0.052)	0.76
Glove100	NER	**0.781** (0.0232)	0.746 (0.0419)	0.763	0.787 (0.0348)	***0.813*** (0.0375)	**0.8**
	RE				0.788 (0.0324)	0.779 (0.0367)	0.783
Glove200	NER	0.768 (0.041)	**0.761** (0.0249)	**0.764**	0.78 (0.0323)	*0.804* (0.0243)	0.792
	RE				0.792 (0.036)	0.777 (0.0378)	0.784
FastText	NER	0.764 (0.0269)	**0.761** (0.029)	0.762	***0.804*** (0.0144)	*0.793* (0.0288)	0.798
	RE				**0.798** (0.0283)	**0.782** (0.0393)	**0.79**

Table 5.5: RDD Corpus: Results obtained by the proposed joint model for relationship extraction (RE) and entity recognition (NER). We report the results obtained using different embeddings. The table shows the results obtained in both training phases: NER sub-model training (Step 1) and full system training (Step 2). Metrics: *precision* (Prec.), *recall* (Rec.) and *f-measure* (SD: Standard deviation). Best results are shown in bold. By column and task (NER/RE) and considering the best results, we report the statistical significant differences underlined ($p < \alpha : \alpha = 0.05$). By row, we report in cursive statistical significant differences found between Step 1 and Step 2 results.

Finally, in order to evaluate the performance of the proposed model in a non-ideal context, we studied the performance of the RE submodel using the entities predicted by the NER submodel. Table 5.6 shows the results obtained using this pipeline and

it compares these results with those obtained using gold-standard information about entities. Although there is a significant loss of performance using this pipeline, this difference of performance is less marked comparing the results of this pipeline with the ones obtained by the pipeline based on stand-alone systems (Table 5.4). In this sense, the improvements obtained in the relationship extraction process were related to the improvement in the detection of long entities.

FastText	**Precision (sd)**	**Recall (sd)**	**F-measure**
NER (Step 2): Global	0.804 (0.0144)	0.793 (0.0288)	0.798
NER (Step 2): Disability	0.845 (0.0196)	0.842 (0.04)	0.844
NER (Step 2): Rare Disease	0.748 (0.03)	0.728 (0.049)	0.738
RE (Step 2): Predicted entities	0.611 (0.0448)	0.536 (0.0341)	0.571
RE (Step 2): Gold entities	0.798 (0.0283)	0.782 (0.0393)	0.79

Table 5.6: RDD Corpus: Results obtained for the task of relationship extraction using the joint model and information about entities not extracted from the gold standard. Metrics: *precision*, *recall* and *f-measure* (SD: Standard deviation).

5.1.4 Discussion

During the development of this thesis we explored different aspects of the extraction of relationships between disabilities and rare diseases. In this study, we evidenced the value of additional representations to model the context between two entities. Although the studied representations have their limitations, they provided significant contributions in the study of relationships between not very distant entities. However, the processing of very long context sequences leads to a distortion of the message to be modeled, since these contextual representations are based on the average of all the terms between two entities. On the other hand, the detection of relations between pairs of named entities implies, by definition, working with the entity recognition output. In this sense, we found a significant deterioration of the results by working with information about entities not extracted from the gold standard. Trying to transfer features of the learning process of the NER system to the learning of the

relation extraction system, we tried to explore a multi-task approach. The combined approach resulted in several improvements in the entity recognition process which were reflected in the complexity of the identified relationships. These results meant an interesting improvement evaluating the performance of a pipeline based on this joint approach.

5.2 Exploring other relationships: eHealth-KD

The eHealth-KD challenge 2019 [127] provided us an opportunity to explore case studies distant from the main line of research of this thesis. This task was developed in the context of IberLEF 2019 (Iberian Languages Evaluation Forum) and it involved the processing of Spanish documents for both named entity recognition and relation extraction. Regarding the proposed entity detection scenarios, we commented on their details and our proposal in Section 4.3. In summary, this task covered the detection of 4 generic types of entities: Concepts, Actions, Predicates and References. After the detection of these entities, the organizers proposed different scenarios to evaluate the performance of relationship extraction systems. In these scenarios, 13 different types of semantic relationships were studied, e.g., in the sentence "El asma afecta a las vías respiratorias." (keyphrases: ["asma" (concept), "afecta" (action) and "vías respiratoria" (concept)]), we can find the following relationships annotated: ("asma", subject, "afecta") and ("afecta", target, "vías respiratorias"). Piad-Morffis et al. [127] provides detailed information on the different types of relationships.

Trying to enrich our study on deep learning techniques applied to relationship extraction, we proposed a model based on a Bi-LSTM with an attention layer. Given its similarities with approaches based on convolutional networks and max pooling, we decided to use the attention model proposed by Zhou et al. [169] for the extraction of the most important semantic information from a sentence. This attention model is based on the use of exponential functions to transform the space generated in previous stages. Similar to the effect studied with the use of pooling mechanisms in Bi-LSTM+CNN-based architectures, the application of this attention mechanism provides a discretizing effect on the temporal dimension of the Bi-LSTM output. Regarding the neural stack configuration, we used a 300-neuron Bi-LSTM (150 per

direction) with *tanh* as activation function. At the output of the model, we studied the space generated by a densely connected network consisting of 13 neurons using a softmax activation function. During the training of the model, we used dropouts of 0.5 as regularization mechanism. Finally, we proposed the use of different features: pretrained Word embeddings, Casing information, Part-of-Speech, Named entity tags and Dependency graph information. As we discussed in the study of entity recognition techniques for Spanish, we used the Word Embeddings released by Cardellino [19] due to the wide variety of sources used to generate them. These vectors have a total of 300 dimensions and gather around 1,000,653 unique tokens. To represent information such as Part-of-Speech or named entity tags, we used embeddings of 25 dimensions generated during training. To represent the casing information or the information extracted from the dependency graph we used one-hot vectors. For a sentence S composed of a maximum of N terms ($w_{x<N}$), we modeled the dependency graph using, for each w, a tuple of vectors (g, t). Whereas g_x contains information about the possible dependency relation of the w term, where w plays the role of governor, the vector t contains information about the relations where w plays the role of target.

5.2.1 Results & Analysis

Team	**Precision**	**Recall**	**F-measure**
Medina and Turmo [105]	0.6667	**0.5915**	**0.6269**
Suárez-Paniagua [144]	0.5892	0.4243	0.4933
Català and Martín [21]	**0.7133**	0.3768	0.4931
Our system	0.6235	0.4665	0.5337
Baseline	*0.4878*	*0.0704*	*0.1231*
Mean	0.5176	0.3234	0.3615

Table 5.7: eHealth-KD challenge 2019 - Subtask B: Best results ordered by F1-Measure. Bold highlights our results and italic marks baseline results.

Similar to our approach for previous evaluations of relationship extraction systems, the eHealth-KD 2019 organizers proposed two different scenarios for the evaluation of the participating systems: (1) evaluation of the relationship extraction systems using the gold annotations of the entities (Subtask B) and (2) evaluation of the participating systems using the entities annotated by the proposed NER systems (Main).

Analyzing the results obtained by the best systems, Table 5.7 shows the results obtained for the Subtask B and, Table 5.8 shows the results obtained for the main scenario. Although the results we obtained show a competitive performance in the evaluation of Subtask B, we observed a loss of effectiveness when we used the annotations predicted by the NER system presented in Section 4.3. This effect was the same effect that we evidenced in the experiments presented in Section 5.1.2, and which seems to be reduced in approaches such as the one presented by Medina and Turmo [105]. They presented a joint approach for keyphrase detection and relation extraction, which during the training was able to transfer to the relation extraction submodel a representation of the imperfections of the keyphrase detection submodel. This approach was based on a Bi-LSTM+CNN stack with two CRF-based outputs. They modeled the extraction of relationships using a table-filling approach. Using this approach they modeled the extraction of relations as a generation of N predictions for each term of a sentence of size N. As a mechanism for modeling each element of a sentence, they used BERT-Cased multilingual. And for the initialization of the weights they did a pre-training of the model where they used data from previous editions. On the other hand, Català and Martín [21] proposed an analysis of the types of relationships according to their complexity and ruled out some types of relationships during training. Although they did not obtain remarkable results of *recall* in Subtask B, they proposed a voting algorithm that obtained very interesting results in the main scenario thanks to the optimization they performed over the training set. The discarding of certain types of relationships due to their complexity or low representativeness in the training set resulted in a refined representation of the majority classes. In this sense, it was to be expected that the attention mechanism we used would be a handicap to represent the minority classes. Finally, Suárez-Paniagua [144] presented an architecture based on Bi-LSTM and pooling. While we

used an attention mechanism for emphasizing the most important features of the latent space generated by the Bi-LSTM, they used a max pooling function for this purpose. Although the effects of both mechanisms were similar, in our case, the use of the attention mechanism endowed the system with greater flexibility generating a trainable intermediate space between the Bi-LSTM output and the final classifier.

	Step 1: NER			Step 2: RE using predicted entities		
Team	Prec. (sd)	Rec. (sd)	F-measure	Prec. (sd)	Rec. (sd)	F-measure
Medina and Turmo [105]	**0.8073**	**0.8336**	**0.8203**	0.6506	**0.6286**	**0.6394**
Català and Martín [21]	0.7986	0.7763	0.7873	**0.7454**	0.5334	0.6218
Suárez-Paniagua [144]	0.5129	0.5851	0.5466	0.4551	0.4056	0.4289
Our system	0.8069	0.7082	0.7543	0.6561	0.4695	0.5473
Baseline	*0.5129*	*0.5851*	*0.5466*	*0.5204*	*0.3677*	*0.4309*
Mean	0.7270	0.7424	0.7334	0.6444	0.4437	0.5201

Table 5.8: eHealth-KD challenge 2019 - Main scenario: Best results ordered by F1-Measure. Bold highlights our results and italic marks baseline results.

5.2.2 Discussion

In an attempt to extend the target of this thesis to case studies more distant from our main line of research, we participated in the eHealth-KD challenge 2019. This opportunity allowed us to explore alternatives to aspects that we studied for the extraction of relationships between disabilities and rare diseases. Since this task covers 13 types of semantic relationships related to the information contained in dependency graphs, we used the graph information to enrich the analysis of certain types of relationships. In addition, we tried to model relationships between entities of different sizes by representing each term of an entity involved in a relationship with its corresponding BIO label (Begin, In, Out). Finally, looking for alternatives to CNN and pooling-based schemes for pattern extraction, we studied the use of an attention function after a Bi-LSTM.

Comparing our results with those obtained by the rest of the participants, although they were competitive, it was clear that there is still room for improvement.

- **Considering the results of *recall*:** Although we considered features focused

on enriching our data model to develop the generalization capabilities of our model; our approach was not as competitive as others that implemented transfer learning techniques or data modeling using contextualized embeddings.

- **Considering the results obtained for the main scenario:** We obtained remarkable results in the Subtask B, however we found an evident loss of efficiency using the information predicted by the NER system. In this sense, approaches based on the elimination of certain classes during training obtained better results.

5.3 Conclusions

We have discussed in this chapter the experiments carried out during this thesis for the study of relationships between different types of biomedical entities, focusing our efforts on the study of relationships between disabilities and rare diseases. Analyzing the relationships between these entities, we explored supervised approaches on the RDD corpus. Although we studied in depth approaches based on deep learning, we also evaluated the performance of systems based on classical machine learning techniques. Highlighting the results obtained using deep learning approaches, in comparison with approaches based on LSTM, we obtained evident improvements using convolutional neural networks, especially exploring features related to the distance between entities. Following these experiments, we jointly studied the identification of relationships and entities in order to identify possible synergies. Using the proposed approach, we obtained significant improvements in both the exploration of relationships and the recognition of named entities. Finally, the improvements obtained in the detection of entities led to a better identification of relationships between entities not extracted from the gold standard.

Furthermore, in the context of eHealth-KD 2019 we extended our study to other types of entities and relationships by proposing a supervised deep learning system. For this workshop and, due to the type of relationships that were addressed, we found useful the use of information extracted from the dependency graph i.e. the relationships studied during this workshop were similar to those described in this

graph. In addition, we incorporated an attention model to focus the learning process and, at the same time, to study the effect of these approaches. Both aspects were promising and provided results comparable to the more advanced approaches based on BERT.

Chapter 6

Negation

Contents

Negation is an aspect of great interest in several domains and natural language processing tasks due to its implications on the speech. This linguistic element performs important functions of discourse polarization, being this aspect especially relevant in sentiment analysis and relationship extraction [32, 26]. During this thesis, through the study of the detection of negation triggers and their scope, we examined the impact of negation in the extraction of relationships between disabilities and rare diseases. Considering Example 6.1, the detection of triggers involves the identification of expressions such as {“No”, “jamas”, “no”} and the detection of the scope involves the identification of the impact of these expressions, e.g. the expression “no” only affects “me gusta”.

> [SCOPE] [**TRIGGER** No] tendré [**TRIGGER** jamas] que aceptar un trabajo que [SCOPE] [**TRIGGER** no] me gusta [/SCOPE] por el dinero [/SCOPE] .

Example 6.1: SFU Review SP-NEG [162] fragment with tag assignment.

We explored negation following an incremental approach in which we defined the following stages: (1) Exploring the detection of triggers; (2) Exploring the joint detection of triggers and scopes; (3) Exploring the contribution of negation in relation extraction and entity recognition systems. First, in Section 6.1 we explore the recognition of negation triggers developing a deep learning system in combination with a post-processing phase based on a rules system. We evaluated the performance of this architecture under the umbrella of the NegEs Workshop i.e. a meeting oriented to the study of systems for the detection of negation triggers in Spanish documents. Next, in Section 6.2 we present a deep learning model for the joint detection of triggers and scopes. By the joint exploration of both elements, we incorporated knowledge about scope recognition in the triggers detection learning. In order to evaluate the adaptability of the proposed system, we evaluated it in different contexts (products reviews and biomedical documents) and languages (Spanish and English). Finally, in Section 6.3 and considering the results obtained by works such as Chowdhury

and Lavelli [26], we extend the exploration presented in Section 5.1.3, including the negation in the pipeline for the joint detection of relations and entities. Due to the small size of the RDD corpus and the small number of negations it contains, we found difficult to develop supervised models for negation detection based on this corpus. For this reason, in this section we explore an approach based on transfer learning between deep learning systems oriented to different tasks in order to analyze the impact on the recognition of named entities and the extraction of relationships.

6.1 Detection of negation triggers

Analyzing the detection of negation triggers, we explored a deep learning model based on a two-layer neural stack (Figure 6-1) and a post-processing phase using regular expressions to correct some frequents errors.

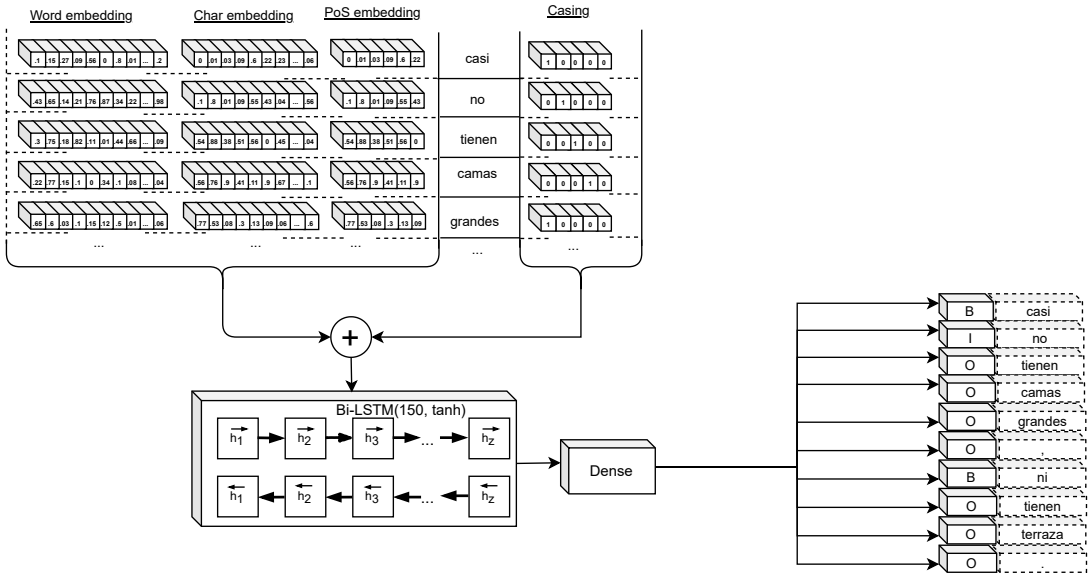


Figure 6-1: Deep learning model for negation trigger detection. Casing information is encoded using a One-hot vector based representation. Bi-LSTM inputs are the concatenated features of each word.

Following a similar approach to the one proposed during the entity recognition research, the developed model includes a Bi-LSTM network (150 neurons per direction) to sequentially process all terms in a sentence and a recurrent network to label each term according to its role as a negation trigger or not[1]. For training this model, we analyzed the following features: words, Part-of-Speech (PoS) labels, characters and casing information. For representing words we used pre-trained Word Embeddings [19] and for representing chars and PoS tags, we used 50-dimensional embeddings calculated during training. The model incorporated a deep learning sub-architecture for character-level term processing [168] and a One-hot vector to represent term casing information. Finally, in order to correct some frequent errors made by the deep learning system, we considered the application of a post-processing phase based on the use of hand-made set of rules. Although the vast majority of errors were related to incorrect IOB constructions, these rules also covered more specific errors detected during the development phase. Fabregat et al. [48] include more details about this model.

6.1.1 Evaluation and analysis

We evaluated the performance of the developed model in the context of the second edition of the NegEs Workshop (NegEs 2019) [163], a meeting proposed during the Iberian Languages Evaluation Forum 2019 [144]. The study of systems for the detection of negation triggers in Spanish (Task A) was one of the main objectives of this workshop. For this purpose the organizers facilitated the SFU Review SP-NEG corpus [162]. This dataset consists of 400 reviews related to 8 different domains (cars, hotels, washing machines, books, cell phones, music, computers and movies). According to the information provided by the organizers, the corpus was randomly divided into training, development and test; ensuring 33 reviews per domain in training, 7 per domain in development and 10 per domain in test.

[1]We used the IOB labeling scheme [131] to label the targets.

Results

Using the training set of the NegEs Workshop, we trained our model during a maximum of 50 epochs using batches of 32 and Adam optimizer [84]. To contextualize the obtained results, Table 6.1 shows a comparison including a set of systems which took part in the 2018 or 2019 editions of NegEs Workshop: IBI [11], UPC [101] and CLiC [11].

	UPC			**CLiC**			**IBI**			**UNED**		
Domain	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Books	**84.19**	**84.52**	**84.35**	80.59	75.79	78.12	80.97	72.62	76.57	84.02	81.35	82.66
Cars	**95.08**	**85.29**	**89.92**	94.92	82.35	88.19	92.73	75.00	82.93	94.83	80.88	87.30
Cell_phones	89.8	**77.19**	**83.02**	87.76	75.44	81.13	**90.48**	66.67	76.77	88.37	66.67	76.00
Computers	91.36	91.36	91.36	90.48	**93.83**	**92.12**	89.06	70.37	78.62	**94.12**	79.01	85.91
Hotels	94	**79.66**	**86.24**	87.5	71.19	78.51	**97.67**	71.19	82.35	93.62	74.58	83.02
Movies	89.68	**85.28**	**87.42**	88.67	81.60	84.99	**90.30**	74.23	81.48	89.86	81.60	85.53
Music	92.96	75.86	83.54	94.44	**78.16**	**85.53**	94.20	74.71	83.33	**95.38**	71.26	81.57
WM	**94.74**	**78.26**	**85.72**	92.98	76.81	84.13	94.34	72.46	81.96	94.34	72.46	81.96
Overall	91.48	**82.18**	**86.45**	89.67	79.40	84.09	91.22	72.16	80.51	**91.82**	75.98	82.99

Table 6.1: NegEs Workshop: Official results by domain for the identification of negation triggers in online product reviews (WM: Washing Machine; P: *precision*; R: *recall*; F: *f-measure*). We included the best proposals presented during the 2018 and 2019 editions. The proposal of Loharja et al. [101] is shown as UPC, Beltrán and González [11] is shown as CLiC, Domínguez-Mas et al. [41] is shown as IBI and Fabregat et al. [48] is shown as UNED.

As it can be seen, the UPC team [101] obtained the best results, highlighting the results of *recall* achieved by them and by the CLiC team [11]. Both teams used architectures based on CRF and high specificity data models, using lists to model information related to terms frequently associated to negation. The use of lists could positively influence the achieved results of *recall*, especially evaluating expressions of relative complexity with low or no frequency of occurrence in the training set. Finally, all participating systems experienced similar difficulties processing domains such as music and mobiles. This performance deterioration may be linked to the presence of domain-specific terms or expressions from other languages.

6.1.2 Discussion

In order to understand the impact of negation on the discourse, we began by exploring the detection of negation triggers in the context of the NegEs Workshop. Although we did not surpass the results of other state of the art systems, this study helped us to identify different elements with a significant impact on the performance. While the use of rules had effects on the precision, some works presented approaches using lists of terms to increase the recall. The use of these lists had very positive effects due to the lack of ambiguity in certain terms related to negation. In addition, since the SFU Review SP-NEG corpus contains information extracted from web pages and written by users without the strict supervision of a proofreader or similar tools, the use of universal representations such as character embeddings or casing information had also interesting effects. These features satisfied the need to represent information that could not be correctly modeled using experience-based data models.

6.2 Exploring negation trigger and scope recognition

Following the exploration of systems for trigger recognition, we extended our study by jointly covering the recognition of negation triggers and their scope. After our participation in the NegEs Workshop, we found that many models established strong dependencies with the studied language. Trying to explore the potential of our model in different languages, we extended the model presented in the previous section to other languages.

In this study we worked on the SFU Review SP-NEG corpus (Spanish) and the BioScope corpus [153] (English). We detailed the characteristics of the SFU corpus in the previous section. The BioScope corpus consists of three parts, electronic health records (EHRs) presented in free text format, full biological articles and abstracts of both scientific and biological studies. The domains included in the BioScope corpus present a complex structure, being the most different the domain of EHRs for the use of a free writing style. The subset of abstracts stands out because it contains more negations than the rest and it is the largest subset. Both corpora were

annotated at token level with labels related to negation triggers and their linguistic scope. In addition, both collections used a similar annotation style and incorporated documents from different scenarios or domains. In short, these similarities encouraged us to propose a study to evaluate a system for the detection of different elements of negation, covering different scenarios and domains, in Spanish and English documents. Fabregat et al. [46] summarize the results of this study.

Deep learning model for the recognition of negation triggers and scopes

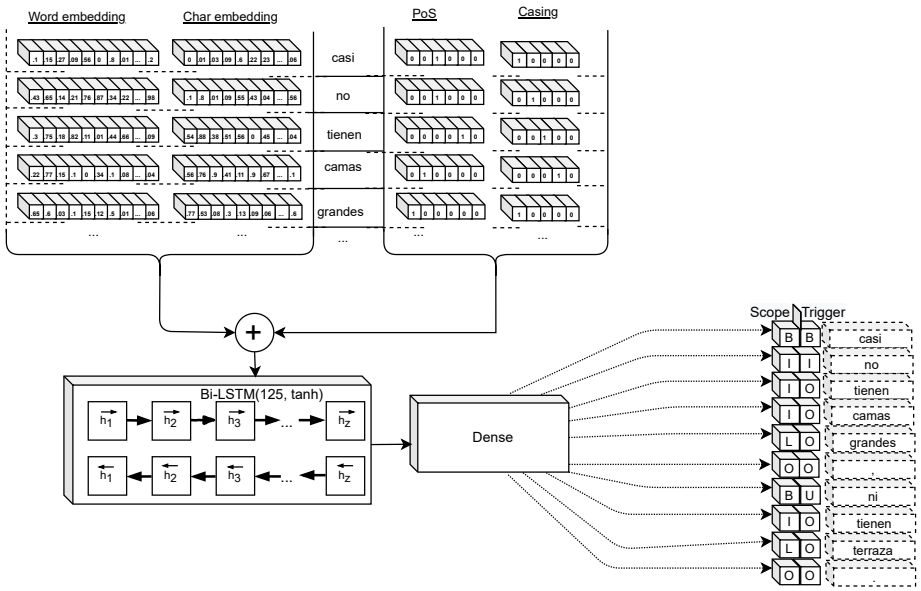


Figure 6-2: Deep learning model for negation trigger and scope detection. Casing information and PoS-tagging are encoded using One-hot vectors. Bi-LSTM inputs are the concatenated features of each word. The model output covers the labeling of triggers and scopes simultaneously.

Among the state of the art on negation scopes recognition, the performance of systems based on deep learning and CRF stands out [52, 98]. Considering this and the good performance achieved by our approach for negation triggers recognition (Section 6.1),

we decided to extend this deep learning system to jointly cover trigger and scope recognition. The proposed model is shown in Figure 6-2.

In order to extend the coverage of the proposed model to handle documents from different languages and domains, and to cover the joint detection of different elements of negation, we needed to include the following modifications:

Features We reduced the number of trainable layers by modeling the casing and PoS information using One-Hot vectors. It was decided to do so because collections such as the BioScope corpus include data sources of modest size and non-free of grammatical errors (e.g. clinical reports), which are an obstacle to consider of interest the generation of high-dimensional representations for the study of PoS. Since we studied documents in English and Spanish, we used two different models of Word Embeddings. For Spanish we used the one presented by Cardellino [19] and for English we used the model developed by Pyysalo et al. [128]. Both models were trained using word2vec and using data from different sources.

Pre-processing We transformed the different datasets into the BILOU labeling scheme [150] (I:In - For tokens part of the annotation; O: Out - For tokens outside the annotation; B:Begin - For the first token of each annotation; L:Last - For the last token of each annotation; U:Unique - Those annotations that have a single token). This annotation scheme allows the representation of partial overlapping and nested entities. We used this labeling scheme to represent both the scope and the negation triggers separately. These two codifications were combined into a single one by means of concatenating the labels. For example, if an expression is both the beginning of a negation and the beginning of a negation triggers, this will be re-labeled with the label "BB". Considering that a negation must has associated a scope, i.e. there are combinations of labels that cannot occur. In total, we generated a total of 17 different labels. Example 6.2 shows an instance annotated following the BILOU format where the first column contains the word and the second column contains the label after joining the scope label and the trigger associated label.

Word	Label	Word	Label
no	BU	no	BU
tendré	IO	me	IO
jamás	IU	gusta	LO
que	IO	por	IO
aceptar	IO	el	IO
un	IO	dinero	LO
trabajo	IO	.	OO
que	IO		

Example 6.2: SFU Review SP-NEG fragment with tag assignment.

This example shows the annotation of two negations. The first one spans from the first term "no" up to the term "dinero" and, the other one is nested and spans from the second term "no" up to the term "gusta".

Post-processing Since in this model we worked with different languages and domains, although we incorporated the use of rules in a post-processing phase, we omitted rules closely linked to specific features of each language. In addition to the safeguards of the BILOU format, we included a set of rules to satisfy the following requirements: Each scope must have at least one associated negation trigger and each annotation must have both a start and an end label, except in the case of a single token annotation. The following are examples of the developed rules:

- If a scope does not have at least one negation trigger associated to it, it is not a scope.

  - Sentence Don't you think it's late?

  - Proposed labels BO BO BO BO BO

  - Corrected labels OO OO OO OO OO

- If a negation trigger does not have one negation scope associated to it, it is not a negation trigger.

- **Sentence** Don't you think it's late?
- **Proposed labels** IS OO OO OO OO
- **Corrected labels** OO OO OO OO OO

- If an annotation starts but does not close, then it finishes in the last term considered by the system as part of the annotation.

  - **Sentence** don't buy it.
  - **Proposed labels** BS IO IO
  - **Corrected labels** BS IO LO

- If an annotation closes a scope but it is not open, it starts with the first trigger of the phrase detected by the system.

  - **Sentence** don't buy it.
  - **Proposed labels** IS IO LO
  - **Corrected labels** BS IO LO

6.2.1 Results and analysis

We divided the evaluation of our system into two parts, training and evaluating using the BioScope corpus. We checked the performance of our system on English documents. Table 6.2 summarizes the performance of our system versus the performance of the systems proposed by Fancellu et al. [52], Fancellu et al. [51] and Li and Lu [98]. Among the conclusions of these works were the good performance of approaches based on Bi-LSTM [51, 52] or CRF [98] and the effectiveness of these approaches to cover different languages and domains [52, 98]. Following an evaluation methodology similar to the one proposed by works from the state of the art, we performed an evaluation using 10-fold cross-validation and considering the evaluation script proposed by Morante and Blanco [115]. Table 6.2 shows the percentage of correctly identified scopes (PCS) and the *f-measure* (F1) at scope level.

	Abstracts		Clinical records		Full papers	
System	PCS	F1	PCS	F1	PCS	F1
Proposed model	80.52	88.54	90.03	94.63	58.99	70.67
Li and Lu [98]	**84.1**	91.3	94.4	95.59	**60.1**	69.23
Fancellu et al. [52]	81.38	**92.11**	94.21	**97.94**	54.54	77.73
Fancellu et al. [51]	73.72	91.35	**95.78**	97.66	51.24	**77.85**

Table 6.2: BioScope corpus (English) - Evaluation of negation scope recognition: Comparison with other state of the art approaches

Table 6.3 shows a performance comparison where we study the performance of our system dealing with the SFU Review SP-NEG corpus. In addition to the results published in the NegEs workshop, we included the results reported by Zafra et al. [164] for trigger and scope recognition. Using two CRF-based systems, Zafra et al. [164] explored the detection of scope and triggers using a wide repertoire of linguistic attributes. In order to contextualize the obtained results with those reported in the state of the art, we used the training and test sets provided by the organizers of NegEs Workshop[2].

	Triggers			Scope		
System	Precision	Recall	F1	Precision	Recall	F1
Proposed Model	91.27	**90.06**	**90.66**	78.71	**76.31**	77.5
Zafra et al. [164]	**91.99**	83.35	87.32	**100 (89.59)**	67.71 (61.91)	**80.68** (73.35)
Fabregat et al. [48]	91.82	75.98	82.99	–	–	–
Loharja et al. [101]	91.48	82.18	86.45	–	–	–

Table 6.3: SFU Review SP-NEG corpus - Evaluation of recognition of both negation scope and negation triggers: Comparison of obtained results by the proposed model with results from other state-of-the-art approaches. Regarding the recognition of negation scope, we present the results obtained by Zafra et al. [164] using the negation triggers extracted from the gold standard and in brackets, those obtained using the predicted triggers.

Analyzing the results obtained by our system, we found similar conclusions

[2]Fabregat et al. [46] summarize the results of our system using 10-fold cross validation.

to those evidenced by works such as Fancellu et al. [52], highlighting the great portability between languages of neural architectures for the detection of negation scopes. Comparing our results with the state of the art, we obtained interesting results processing documents in Spanish, especially detecting Spanish negation triggers. However, although the joint exploration of triggers and negation scope leads to improvements in the detection of triggers, this approach implies the unavailability of triggers for modeling the scope. This fact is relevant when we compare the results of our system in the detection of scope in both Spanish and English documents with the state of the art. Reaffirming the importance of including information about the triggers in the scope detection, the results obtained by Zafra et al. [164] using information not extracted from the gold standard about negation triggers show a clear deterioration in comparison with the results obtained using gold information.

6.2.2 Discussion

Extending our analysis of negation we covered the detection of negation scopes using a joint approach for the detection of triggers and scopes simultaneously. We analyzed the performance of our proposal by processing documents from different domains and languages. Comparing the performance of our system with other systems from the literature, our joint approach led to improvements in the detection of triggers by considering constraints in the labeling model related to the scope, i.e. a trigger has to be under the context of a scope. However, this joint approach required not using information about negation triggers for scope recognition, which is a very relevant limitation if we take into account that many state-of-the-art systems use features extracted from negation triggers to predict the scope. The approach proposed by Zafra et al. [164] is an example where we find a deterioration in the performance of scope recognition by not using negation triggers extracted from the gold standard.

6.3 Negation knowledge on relation extraction

In Section 5.1.3 we analyzed the impact of multi-task models on relation extraction and named entity identification processes. After evidencing positive effects addressing both

tasks jointly, we decided to study other aspects that could affect relation extraction processes. Given the evident impact of negation on discourse understanding, we explored the extraction of relationships by considering aspects related to this linguistic phenomena.

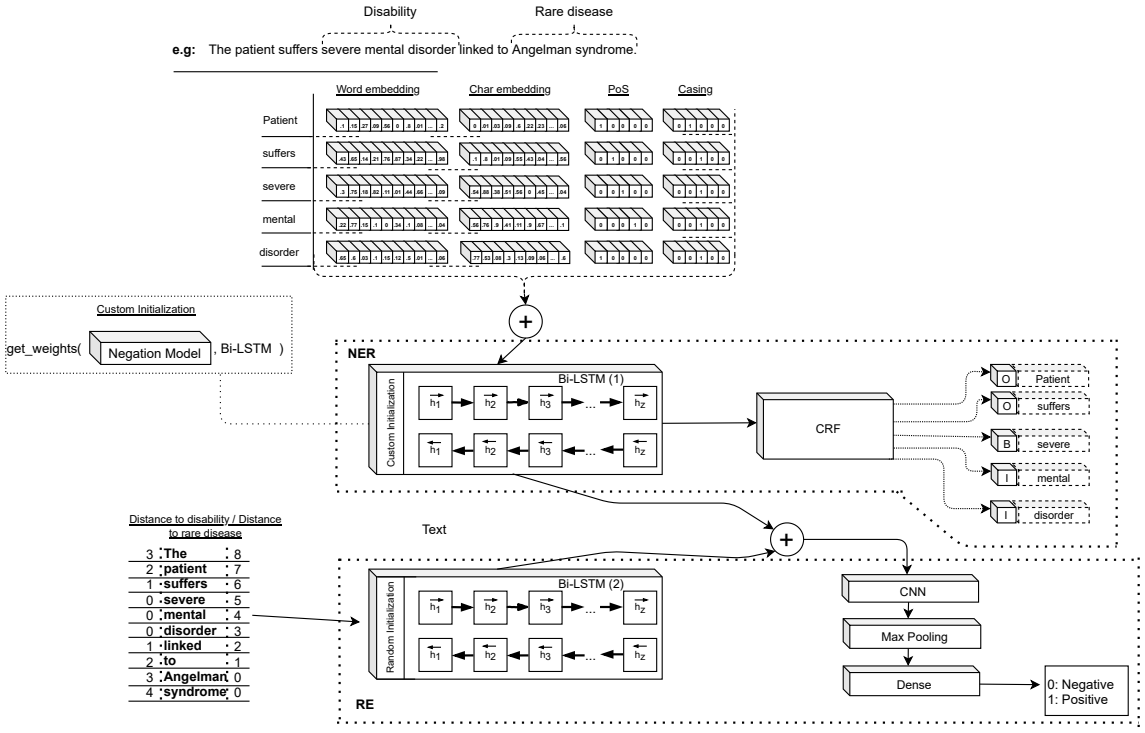


Figure 6-3: RDD Corpus: Deep learning joint model for named entity recognition (NER submodel) and relationship extraction (RE submodel) using a custom initialization based on transfer learning. NER submodel: The model uses a Bi-LSTM+CRF based stack to process the concatenation of the inputs (words and characters embeddings, and one-hot vectors representing Part-of-speech and casing information). The weights of the Bi-LSTM are initialized with the weights of an auxiliary model. RE submodel: The model uses two neural stacks. (1) two Bi-LSTMs (one of them shared with NER submodel) and a convolutional network to process the inputs of more than two dimensions. (2) two densely connected networks (50 and 1 neuron/s) to process the concatenation of the pooling of (1) with the two-dimensional inputs (summarized context embeddings).

Exploring the RDD corpus, we focused this study on relation extraction models between disabilities and rare diseases. This corpus has a set of limitations that make it less feasible to address negation by studying certain types of algorithms i.e. small size and few annotated negations. These facts would affect the application of classical supervised system. For these reasons, we explored transfer learning approaches between models trained with different corpora and focused on different objectives (negation triggers and scopes detection, and relation extraction and named entity recognition).

During this study, we explored an architecture based on the models proposed in Sections 6.2 and 5.1.3. These models shared many aspects, e.g. both were focused on sequence tagging of English documents and both models used a Bi-LSTM network as a central processing component. As shown in Figure 6-3, we exploited these similarities to perform a custom weights initialization of the multi-task model (Section 5.1.3), using the training of the negation processing model (Section 6.2). In order to perform this study and to achieve coherence between the vectorial spaces generated by both models, we trained both systems using the same set of Word Embeddings (Glove of 100 dimensions). To train this model, we proposed a two-stage training scheme. In the first phase, we trained the negation processing model using the criteria outlined in Section 6.2. Next, we used the weights of the trained Bi-LSTM to initialize the Bi-LSTM (1) of the multi-task model. Finally, we trained the multi-task model following the criteria explored in Section 5.1.3.

6.3.1 Results and analysis

We explored the impact of negation on the training of a multi-task model by evaluating the generated system using 10 fold cross-validation. Tables 6.4 and 6.5, show the results obtained for named entity detection and relationship extraction. Both tables present a comparison of the performance achieved by the multi-task model in each training step. We also include in this comparison the results obtained without the knowledge transferred from the negation model (Glove100).

		Step 1: NER Training			Step 2: Joint Training		
		Precision	Recall	F-measure	Precision	Recall	F-measure
Glove100	NER	0.781	0.746	0.763	0.787	0.813	0.8
	RE				0.788	**0.779**	0.783
Glove100 + NEG	NER	**0.818**	**0.806**	**0.812**	**0.823**	**0.827**	**0.825**
	RE				**0.806**	0.775	**0.79**

Table 6.4: RDD Corpus: Results achieved by using a transfer learning approach for relationship extraction and entity recognition (studying the impact of knowledge extracted from a model for negation processing). Glove100 refers to the results obtained prior to this research (Table 5.5). The table shows the results obtained in both training phases: NER sub-model training (Step 1) and full system training (Step 2). Metrics: *precision*, *recall* and *f-measure*. The best results are shown in bold.

As shown in Table 6.4, the use of the studied transfer learning approach improves the results on early training stages of the multi-task model (Step 1: NER Training). Focusing on entity recognition and without considering the application of the knowledge extracted from the negation model, these improvements are comparable to those obtained in the Step 2 of the training (Step 2: Joint Training). In Section 5.5 we commented that the improvements achieved during Step 1 and 2 of the training might be due to the oversampling required by the data format used to represent the relations. Since we limited the study to relations between entity tuples, we replicated the NER annotations of each sentence as many times as positive or negative relations it contains, implying an oversampling process over the NER labels. During Step 2, the positive effect obtained by oversampling is also reproduced in the results obtained using the proposed transfer learning approach.

Table 6.5 shows the results obtained by the proposed model evaluating possible relationships between entities extracted from the gold standard or between entities recognized by this model. In addition we show the results obtained for the NER task, divided by type of entity. According to the results shown in the table, using information extracted from the gold standard as well as evaluating relationships between entities recognized by the system itself, we obtained considerable improvements. The achieved results of *precision* stand out, representing a significant reduction of the number of false positives to evaluate when our system works with entities not extracted from

the gold standard. Finally, analyzing the results of *recall*, although we obtained some improvements, these were affected by the low frequency of some terms, especially those related to rare diseases.

Glove100	**Precision**	**Recall**	**F-measure**
NER (Step 2): Global	0.787	0.813	0.80
NER (Step 2): Disability	0.834	0.856	0.845
NER (Step 2): Rare Disease	0.730	0.757	0.743
RE (Step 2): Predicted entities	0.583	0.558	0.57
RE (Step 2): Gold entities	0.788	0.779	0.783

Glove100+NEG	**Precision**	**Recall**	**F-measure**
NER (Step 2): Global	0.823	0.827	0.825
NER (Step 2): Disability	0.872	0.87	0.871
NER (Step 2): Rare Disease	0.761	0.772	0.766
RE (Step 2): Predicted entities	0.628	0.578	0.597
RE (Step 2): Gold entities	0.806	0.775	0.79

Table 6.5: Transfer Learning. Pipeline NER and RE - RDD Corpus: Results achieved by using a transfer learning approach for the task of relationship extraction using the joint model and information about entities not extracted from the gold standard (studying the impact of knowledge extracted from a model for negation processing). Glove100 refers to the results obtained prior to this research. Metrics: *precision*, *recall* and *f-measure*.

6.3.2 Discussion

In order to study the impact of negation on relation extraction processes, in this section we addressed a transfer learning study between the model presented in Section 6.2 for negation processing, and the multi-task model presented in Section 5.1.3. The results obtained show a positive effect especially in the detection of named entities, which has a positive impact on the exploration of possible relationships between

predicted entities. The transfer of information between both tasks had a comparable impact to that obtained after the second training of the multi-task model (Section 5.1.3 - Table 5.5).

6.4 Conclusions

Although this thesis focuses on the study of techniques for information extraction from medical documents, these methods are influenced by certain linguistic phenomenons. We extended the exploration of techniques for relation extraction and entity recognition by exploring methods for negation processing. Focusing our efforts on the study of the detection of negation scopes and triggers, we explored the negation in different phases.

First, we developed a deep learning-based model to discover negation triggers in Spanish documents. Under the umbrella of the NegEs workshop, this model achieved a remarkable performance without requiring any linguistic or technical features strongly linked to the target language.

After the exploration of negation triggers, we addressed the automatic annotation of negation scopes by developing a deep learning system for the joint annotation of negation triggers and scopes. We extended the analysis covering documents from different languages and domains. For the processing of Spanish documents, the results obtained in the detection of negation triggers stand out, having achieved results above the state of the art using a reduced set of features. These improvements were associated to the extension of the BILOU tagging that we used to jointly consider the scope and triggers tagging. This labeling forced the identification of a negation trigger to the previous identification of an associated scope. However, this labeling approach limited the exploration of negation triggers as input for scopes recognition. Although our model did not achieve improvements in identifying scopes in Spanish documents, we achieved better performance than other systems using triggers not extracted from the gold standard. Regarding the processing of English documents, we evaluated our model only for scopes recognition with findings analogous to those seen when processing Spanish documents.

Finally, we explored transfer learning techniques between the developed negation

model and the multi-task model described in Section 5.1.3. Both models shared the representation space generated during the training of the negation model. We analyzed this scheme of transfer learning to study the impact of negation on entity detection and relation extraction processes. Using this approach, we obtained significant positive effects. These improvements were mostly related to named entity recognition results. The consideration of negation also had positive effects on relation extraction, especially focused on *precision*. Linking the results obtained in both tasks, the reduction of false positives during named entity recognition meant a reduction of candidates to be evaluated by the relation extraction process.

Chapter 7

Conclusions and Future Work

Contents

In previous chapters, we described and contextualized the work carried out. The current chapter represents the conclusion of this thesis. In Section 7.1 we discuss the main contributions generated during the development of the thesis. Later, in Section 7.2, we present the main conclusions of the work carried out as answers to the Research Questions proposed in Chapter 1. We intend to determine through these answers and conclusions whether the main Research Objective has been fulfilled or not. After discussing the different points developed during this thesis, we discuss some lines of research to be explored in future works (Section 7.3). Finally, we present the publications derived from this thesis (Section 7.4).

7.1 Main Contributions

In this first section, we summarize the main contributions developed through the research process described in this thesis:

Corpora about rare diseases and disabilities

Given the lack of annotated resources oriented to the study of disabilities and rare diseases, we have addressed the generation of corpora for the study of these entities and their possible relationships. In Chapter 3 we exhaustively described the annotation guidelines developed for the annotation of disabilities, both in English and Spanish. In addition, we detailed the particularities about the annotation of relationships between disabilities and rare diseases. Using these annotation guidelines, we created the RDD corpus with annotations on disabilities, rare diseases, and the relationships between both types of entities. Additionally, in order to enable the study of languages beyond English, we developed the DIANN corpus, which includes abstracts in Spanish and English with annotations about the mentioned disabilities. Both corpora contain annotations on negation when it affects one or more disabilities.

Recognition of mentions to disabilities and rare diseases

In order to study the generation of systems for the recognition of disabilities and rare diseases, in Chapter 4 we analyzed this task using different approaches. First, we studied the strengths and weaknesses of basic supervised systems, based on

machine learning and deep learning techniques. After that, we extended the analysis of disability detection by examining the performance of the systems presented during the DIANN Shared Task. We also took part in this task analyzing an unsupervised approach based on variant generation.

Following the analysis of the proposals presented in DIANN Shared Task, and in an effort to explore in additional contexts some of the drawn conclusions, we developed and tested a generic deep learning architecture for the de-identification of clinical reports and the extraction of keyphrases. Finally, and considering the different lessons learned, we developed a deep learning system for disability detection improving the results achieved until now.

Identification of relationships between disabilities and rare diseases

Focusing on the study of relationships between disabilities and rare diseases, in Chapter 5 we discussed approaches based on both machine learning and deep learning. Having obtained a good performance using deep learning based systems, we extended this framework by analyzing different types of features and deep learning architectures obtaining significant improvements. Finally, in order to analyze possible refinements in the processing of information related to rare diseases and disabilities, in Chapter 5, we described the development of a joint model for entity detection and relationship extraction. Using this joint approach, we obtained different improvements, especially in the analysis of relationships between entities not extracted from the gold standard.

Analysis of negation triggers and scopes, and their impact on named entity recognition and relation extraction

Exploring linguistic aspects with influence on information extraction processes, in Chapter 6, we discussed our study about negation. First, we successfully explored an approach based on deep learning for the detection of triggers in Spanish documents. After that and in order to study possible improvements through the joint learning of different tasks, we developed a model for the joint detection of negation triggers and scopes. The improvements obtained in the identification of negation triggers, especially in Spanish, were outstanding. Finally, we explored the impact of negation on relationship extraction and entity detection processes. Using an approach based

on weight sharing, we obtained relevant improvements on entity detection that had a positive impact on the identification of relationships between entities not extracted from the gold standard.

7.2 Answers to Research Questions

In Chapter 1, we described the different objectives of this thesis organized under a main research objective:

> **Research Objective**
>
> To study named entity recognition and relationship extraction for the processing of documents related to disabilities and rare diseases. To analyze the performance of automatic systems dealing with documents of different languages and to study the effect of jointly exploring both entity detection and relation extraction. To explore the impact of linguistic aspects such as negation on the extraction of relationships.

Having addressed the study and development of these objectives, in this section, we try to answer the different research questions by way of summarizing the main conclusions of this thesis.

Research Question 1. *Considering the difficulties inherent to the biomedical domain, do disabilities and rare diseases present additional difficulties for information extraction?*

The detection of disabilities involves the consideration of some aspects that make the processing of related documents particularly difficult. Some of these aspects were highlighted during the generation of the corpora described in Chapter 3. First, considering the definition of disability established by the WHO, we found some ambiguity in determining whether or not a disease is a disability. Under WHO

criteria, a disease is considered a disability if it leads to a deterioration of certain cognitive or corporal functions. At the same time, the WHO established a set of thresholds for the duration and intensity of these impairments and their consequences on the day-to-day life of the affected people. This casuistry complicates the generation of specific resources oriented to the study of this type of entities. We confirmed this fact during the analysis of the unsupervised systems presented at the workshop DIANN for the study of disabilities. They had notable problems in differentiating between disease and disability, especially those systems that relied on the use of thesauruses. On the other hand, and given its definition, a rare disease is only distinguished from a common disease by its degree of incidence, which means that the development of systems for the automatic identification of mentions of rare diseases also involves dealing with some hidden semantics. Comparing the ambiguity of these entities with the ambiguity found in disabilities, rare diseases are less complex to address through the development of dictionaries, a fact that is evidenced in some Orphanet contributions for the generation of resources oriented to the study of these entities.

Analyzing the way in which disabilities and rare diseases are presented, both types of entities imply the consideration of a wide range of variants to refer to the same concept. This fact is especially prevalent in the study of disabilities, where highly complex cases may arise from the use of a free narrative style and the need to specify aspects of this condition related to duration and severity. In Chapter 4, we studied the performance of different systems under different matching criteria and we found a clear difficulty in the correct identification of all the terms related to disability.

Finally, although there are many efforts in the development of resources oriented to the biomedical domain, the scarcity of resources focused on disabilities and rare diseases hinders its study. In Chapter 5 and 6, we obtained improvements in disability detection due to the oversampling performed to jointly model disability detection and relation extraction. Even considering the improvements obtained by artificially increasing the training data, there is a gap for improvement that could be addressed by generating new annotated collections.

Research Question 2. *Which is the impact of changing the language on the performance of systems for automatic detection of biomedical entities?*

During the DIANN Workshop, some multi-language models were presented and although they obtained a competitive performance, they raised some doubts about the viability of this type of approaches versus others focused on a single language. Analyzing supervised systems for the detection of disabilities (Section 4.4), we explored different systems for the processing of documents in English and Spanish. During this analysis we found that the use of certain features, such as Part-of-Speech, resulted in improvements during the handling of English documents and failed during the processing of Spanish documents. A similar fact occurred when studying the impact of features based on embeddings. We studied an attention model based on the modeling of reference terms using embeddings i.e., "disability" for English and "discapacidad" for Spanish. Analyzing the results obtained using this model, we only found improvements in the processing of English documents. Although the terms used represent the same concept, the embedding model used for each language provides similarity spaces that show the differences between the number of contexts where other tuples of terms, such as ("impairment" [English], "deterioro" [Spanish]), occur. In summary, although we approached the problem using the same architecture for both languages, we obtained the best results exploring different sets of features for each language.

Research Question 3. *Analyzing small datasets, how useful is the use of multitasking systems for entity recognition and relationship extraction?*

We explored joint approaches for relationship extraction and entity recognition. During these experiments we focused on the study of disabilities and rare diseases. In order to jointly cover both tasks, we applied oversampling on the NER training set, which reduced the impact of certain aspects defined during the training phase e.g. batch size. At the same time, the joint learning of both tasks allowed a better assimilation of the dependency relationships between the different terms due to the need to explicitly indicate, that an entity to be recognized is also part of a relationship with another entity mentioned in the same sentence. Although the improvements we obtained were mainly focused on named entity recognition, these improvements also led to some advances in the extraction of relations.

Research Question 4. *Analyzing transfer learning between different tasks, what is the effect of negation processing approaches on systems for relation extraction?*

During the analysis of negation discussed in Chapter 6, we explored different approaches similar to those explored for named entity recognition (Chapter 4). We obtained interesting improvements in the joint exploration of entities and relations by incorporating a weight initialization based on the weights extracted from a system for negation processing. This transfer learning scheme led to improvements in entity recognition that positively affected the extraction of relations. We obtained a better representation of the entities by learning the boundaries of negation scopes i.e. if an entity is included within a negation, the length of this entity is limited by the length of the negation.

7.3 Future Lines of Work

In previous chapters we analyzed, by exploring different tasks, the processing of documents related to disabilities and rare diseases. Although we have answered the different research questions proposed at the beginning of this document, some possible lines of future work have emerged during the development of this thesis.

Exploration of more specific entities

During this thesis we developed different systems for the identification of disabilities in scientific documents. After analyzing the results of the proposed systems, we have identified several errors related to the identification of diseases as disabilities. Since a disability can be defined as the impairment of a physical and/or cognitive function, extending the identification of disabilities to more specific entities, including the recognition of the functions affected and the type of impairment suffered, could lead to different improvements. Considering that the DIANN and RDD corpus contains this type of information, these datasets may be useful for the study of these additional entities. The following example shows an annotated sentence extracted from the RDD corpus:

"After the accident the youth presented <dis id=0> serious **<imp>** difficulties **</imp>** in **<fun>** walking **</fun>** and **<fun>** talking **</fun>** </dis>.".

Besides the label "dis", we used specific labels to annotate terms related to the affected functions (walking/speaking) and the type of impairment (difficulties). In summary, the presence of functions and terms denoting a deterioration of a condition is a similar aspect to the one identified during the joint recognition of negation scopes and triggers. The improvements found during negation processing could be obtained with this new proposal.

Learning strategies

From multi-class approaches to multi-task learning, during Chapter 5, we analyzed the performance of a joint system for entity recognition and relation extraction. Effects derived from this joint approach, such as the oversampling performed on the NER labels, led to significant improvements in both entity recognition and relationship extraction. Considering this ensemble, the incorporation of tasks such as PoS labeling or semantic role labeling could lead to further improvements and a better understanding of the performance obtained by this approach.

On the other hand, in Section 6.3, we analyzed a transfer learning approach based on weight sharing between different systems. Using this approach, we explored the effect of negation on named entity recognition and relation extraction. Although we obtained interesting improvements, a more exhaustive exploration, including tasks such as the detection of speculative statements, could lead to similar improvements.

Exploration of relationships multi-line

During this thesis we analyzed the identification of relationships between entities mentioned in the same sentence. It would be interesting to extend this experimentation to contexts evaluating relationships between entities mentioned in different sentences.

Modern language models

During the development of this thesis, in different NLP tasks, the use of contextualized embeddings, such as BERT, made significant improvements. The use of these resources

could lead to improvements in different aspects of the analyzed tasks.

Multilingual approaches

With the development of the DIANN corpus and the DIANN workshop, we studied the detection of disabilities in a multilingual context, however, we did not study in depth the feasibility of multilingual supervised systems exploiting the duality of this dataset. Although we have carried out some experiments using multilingual word embeddings, we have not documented them due to the low obtained performance. A detailed analysis is needed to verify the potential of this type of datasets to train supervised systems for the identification of disabilities in documents from different languages.

Negation triggers

On the study carried out about negation processing, we covered the development of a system for the joint annotation of negation triggers and scopes. During this work we evidenced that an aspect of great importance in the study of scopes is the presence of triggers. The exploration of features derived from the study of negation triggers could lead to improvements in the correct recognition of scopes.

Knowledge transfer

Finally, we want to generate software packages with the developed systems to simplify the use of the contributions derived from this thesis and make them available to the scientific community.

7.4 Publications

During the development of this thesis, the results and conclusions reached by us have been published in different specialized sources. Whereas the publishing in specialized journals has provided us important feedback on the reached conclusions, the participation in workshops has allowed us to evaluate the developed proposals under the same evaluation framework.

- Hermenegildo Fabregat, Lourdes Araujo, Juan Martínez-Romo: **Deep neural models for extracting entities and relationships in the new RDD corpus relating disabilities and rare diseases.** Comput. Methods Programs Biomed. 164: 121-129 (2018)

- Hermenegildo Fabregat, Juan Martínez-Romo, Lourdes Araujo: **Overview of the DIANN Task: Disability Annotation Task.** IberEval@SEPLN 2018: 1-14

- Hermenegildo Fabregat, Juan Martínez-Romo, Lourdes Araujo: **UNED at DIANN 2018: Unsupervised System for Automatic Disabilities Labelling in Medical Scientific Documents.** IberEval@SEPLN 2018: 53-60

- Hermenegildo Fabregat, Andres Duque Fernandez, Juan Martínez-Romo, Lourdes Araujo: **NLP_UNED at eHealth-KD Challenge 2019: Deep Learning for Named Entity Recognition and Attentive Relation Extraction.** IberLEF@SEPLN 2019: 67-77

- Hermenegildo Fabregat, Andrés Duque, Juan Martínez-Romo, Lourdes Araujo: **Extending a Deep Learning Approach for Negation Cues Detection in Spanish.** IberLEF@SEPLN 2019: 369-377

- Hermenegildo Fabregat, Andrés Duque, Juan Martínez-Romo, Lourdes Araujo: **De-Identification through Named Entity Recognition for Medical Document Anonymization.** IberLEF@SEPLN 2019: 663-670

- Hermenegildo Fabregat, Lourdes Araujo, Juan Martínez-Romo: **Deep learning approach for negation trigger and scope recognition.** Proces. del Leng. Natural 62: 37-44 (2019)

- Hermenegildo Fabregat, Juan Martínez-Romo, Lourdes Araujo: **Understanding and Improving Disability Identification in Medical Documents.** IEEE Access 8: 155399-155408 (2020)

- Razan Masood, Mengjiao Hu, Hermenegildo Fabregat, Ahmet Aker, Norbert Fuhr: **Anorexia Topical Trends in Self-declared Reddit Users.** CIRCLE 2020

Appendix A

Tables: Annotation process

CORPUS FUNCTIONALITY	FUNCTIONALITY LIST ORPHANET
ability to recognise face	Interpersonal relations
academic	Learning
attention	Focusing attention
audiological	Participating in a conversation
auditory	Listening
auricular	Listening
autonomic	Life management
behavior	Life management
behaviour	Life management
brain	Understanding
chew	Eating
cognitive	Solving problems, Communicating with others ...
communication	Communicating with others
cortical	Understanding
development	b560 Growth maintenance functions / ICF-CY
emotional	Interpersonal relations
executive function	Motor skills
face perception	Interacting with other people
face recognition	Interacting with other people
feeding	Eating
functional	Life management
gait	Motor skills
growth	b560 Growth maintenance functions / ICF-CY
hearing	Listening
intellectual	Understanding
jaw opening	Eating
language	Acquiring language
learning	Learning
memory	Understanding
mental	Understanding
motor	Motor skills
movement	Motor skills
neurological	Understanding

Table A.1: [Part 1] Lists of functions having been identified during the corpus annotation as part of a disability. Additionally, we show the correspondence of these functions, with those being listed in the Orphanet Functioning Thesaurus.

CORPUS FUNCTIONALITY	FUNCTIONALITY LIST ORPHANET
object recognition	Using objects
ocular	Watching
ophthalmologic	Watching
optic	Watching
oral	Speaking
orofacial	Communicating with others
personality	Controlling one's own behaviour
phonological	Communicating with others
physical	Motor skills
processing speed	Education
psychiatric	Handling stress and other psychological demands
psychological	Handling stress and other psychological demands
receptive vocabulary	Receiving messages
retina	Watching
retino-cerebellar	Watching
running	Motor skills
sensitivity to sound	Listening
sensorial	Listening , Watching,...
sensory	Listening , Watching, ...
skeletal	Motor skills
sleep	Daily activities
social	Social life
speaking	Communicating with others
speech	Communicating with others
stand	Maintaining a standing position
swallow	Eating, Drinking
verbal	Communicating with others
vision	Watching
visual	Watching
visuo-spatial	Watching
visuospatial	Watching
walk	Walking

Table A.2: [Part 2] List of functions identified, during corpus annotation, as part of a disability. Additionally, we show the correspondence of these functions, with those listed in the Orphanet Functioning Thesaurus.

aberration	delay	limitation
abnormal	detachment	limited
absence	deterioration	loss
absent	difficulty	lower
alteration	diminished	maladaptive repetitive
anomaly	disability	no
athetoid	disabled	poor
atrophy	disease	problem
compromise	disorder	reduced
decline	disturbance	regression
decreased	dysfunction	restriction
defect	dystrophy	retard
deficiency	failure	self-injuring
deficient	impaired	self-injurious
deficit	impairment	stagnant
degeneration	inability	unable
degradation	issue	weak

Table A.3: Annotation process: List of impairment words obtained after the annotation process.

Appendix B

Tables: DIANN Shared Task - Results

Spanish	P	R	F
IxaMed R1	0.757	0.817	0.786
UC3M_1 R2	0.818	0.646	0.722
UC3M_1 R1	0.801	0.651	0.718
UPC_3 R2	0.807	0.603	0.69
UPC_3 R1	0.814	0.594	0.687
UC3M_1 R3	0.801	0.563	0.662
IXA_2 R1	0.65	0.642	0.646
IXA_2 R3	0.636	0.655	0.645
UPC_3 R3	0.67	0.603	0.634
IXA_2 R2	0.641	0.616	0.628
UPC_2 R1	0.732	0.502	0.596
SINAI_1 R3	0.459	0.345	0.394
LSI_UNED R3	0.41	0.249	0.31
LSI_UNED R2	0.396	0.249	0.306
LSI_UNED R1	0.393	0.249	0.305
GPLSIUA_1 R1	0.813	0.17	0.282
GPLSIUA_1 R2	0.796	0.17	0.281
SINAI_1 R2	0.181	0.415	0.252
SINAI_1 R1	0.022	0.485	0.042

Table B.1: Disability recognition: Spanish - Exact matching. Precision (P), Recall (R) and F-measure (F).

English	P	R	F
IxaMed R1	0.786	0.86	0.821
UC3M_1 R1	0.778	0.72	0.748
UC3M_1 R2	0.759	0.663	0.708
UC3M_1 R3	0.775	0.65	0.707
UPC_3 R1	0.799	0.605	0.689
UPC_3 R2	0.795	0.605	0.687
UPC_2 R1	0.756	0.56	0.643
UPC_3 R3	0.655	0.617	0.636
LSI_UNED R3	0.671	0.597	0.632
LSI_UNED R2	0.639	0.597	0.617
LSI_UNED R1	0.633	0.597	0.614
IXA_2 R1	0.701	0.531	0.604
IXA_2 R2	0.706	0.494	0.581
SINAI_1 R3	0.625	0.37	0.465
GPLSIUA_1 R2	0.884	0.251	0.391
GPLSIUA_1 R1	0.881	0.243	0.381
SINAI_1 R2	0.222	0.428	0.293
SINAI_1 R1	0.016	0.593	0.032

Table B.2: Disability recognition: English - Exact matching. Precision (P), Recall (R) and F-measure (F).

Spanish	P	R	F
IxaMed R1	0.822	0.886	0.853
UC3M_1 R1	0.882	0.716	0.79
UC3M_1 R2	0.878	0.694	0.776
UPC_3 R2	0.889	0.664	0.76
UPC_3 R1	0.898	0.655	0.758
IXA_2 R3	0.712	0.734	0.723
UC3M_1 R3	0.876	0.616	0.723
IXA_2 R1	0.721	0.712	0.716
UPC_3 R3	0.743	0.668	0.703
IXA_2 R2	0.705	0.677	0.69
UPC_2 R1	0.828	0.568	0.674
LSI_UNED R2	0.847	0.533	0.654
LSI_UNED R1	0.841	0.533	0.652
LSI_UNED R3	0.842	0.511	0.636
SINAI_1 R3	0.512	0.384	0.439
GPLSIUA_1 R2	0.959	0.205	0.338
GPLSIUA_1 R1	0.958	0.201	0.332
SINAI_1 R2	0.204	0.467	0.284
SINAI_1 R1	0.026	0.568	0.05

Table B.3: Disability recognition: Spanish - Partial matching. Precision (P), Recall (R) and F-measure (F).

English	P	R	F
IxaMed R1	0.842	0.922	0.88
LSI_UNED R3	0.856	0.761	0.806
UC3M_1 R1	0.822	0.761	0.791
LSI_UNED R2	0.815	0.761	0.787
LSI_UNED R1	0.808	0.761	0.784
UC3M_1 R2	0.835	0.728	0.778
UC3M_1 R3	0.828	0.695	0.756
UPC_3 R1	0.875	0.663	0.754
UPC_3 R2	0.865	0.658	0.748
UPC_3 R3	0.742	0.7	0.72
UPC_2 R1	0.822	0.609	0.7
IXA_2 R1	0.761	0.576	0.656
IXA_2 R2	0.788	0.551	0.649
SINAI_1 R3	0.688	0.407	0.512
GPLSIUA_1 R1	0.94	0.259	0.406
GPLSIUA_1 R2	0.913	0.259	0.404
SINAI_1 R2	0.252	0.486	0.332
SINAI_1 R1	0.019	0.704	0.038

Table B.4: Disability recognition: English - Partial matching. Precision (P), Recall (R) and F-measure (F).

Spanish	P	R	F
IxaMed R1	0.889	0.727	0.8
IXA_2 R1	1	0.545	0.706
IXA_2 R2	0.929	0.591	0.722
IXA_2 R3	0.923	0.545	0.686
UPC_2 R1	0.737	0.636	0.683
UPC_3 R3	0.688	0.5	0.579
UPC_3 R1	0.647	0.5	0.564
UPC_3 R2	0.647	0.5	0.564
SINAI_1 R3	0.667	0.091	0.16
SINAI_1 R2	0.333	0.045	0.08
GPLSIUA_1 R1	0	0	0
GPLSIUA_1 R2	0	0	0
LSI_UNED R1	0	0	0
LSI_UNED R2	0	0	0
LSI_UNED R3	0	0	0
SINAI_1 R1	0	0	0
UC3M_1 R1	0	0	0
UC3M_1 R2	0	0	0
UC3M_1 R3	0	0	0

Table B.5: Negated disability recognition: Spanish - Exact matching. Precision (P), Recall (R) and F-measure (F).

English	P	R	F
UPC_3 R1	0.773	0.739	0.756
UPC_3 R2	0.773	0.739	0.756
UPC_3 R3	0.696	0.696	0.696
GPLSIUA_1 R1	0.647	0.478	0.55
UPC_2 R1	0.647	0.478	0.55
GPLSIUA_1 R2	0.611	0.478	0.537
IXA_2 R1	0.667	0.435	0.526
IXA_2 R2	0.75	0.391	0.514
SINAI_1 R3	0.526	0.435	0.476
IxaMed R1	0.476	0.435	0.455
SINAI_1 R2	0.306	0.478	0.373
SINAI_1 R1	0.25	0.391	0.305
LSI_UNED R2	0.188	0.13	0.154
LSI_UNED R3	0.188	0.13	0.154
LSI_UNED R1	0.176	0.13	0.15
UC3M_1 R1	0	0	0
UC3M_1 R2	0	0	0
UC3M_1 R3	0	0	0

Table B.6: Negated disability recognition: English - Exact matching. Precision (P), Recall (R) and F-measure (F).

Spanish	P	R	F
IxaMed R1	1	0.818	0.9
UPC_3 R3	1	0.727	0.842
UPC_2 R1	0.895	0.773	0.829
UPC_3 R1	0.941	0.727	0.821
UPC_3 R2	0.941	0.727	0.821
UC3M_1 R3	1	0.682	0.811
IXA_2 R3	1	0.591	0.743
IXA_2 R2	0.929	0.591	0.722
IXA_2 R1	1	0.545	0.706
UC3M_1 R2	0.909	0.455	0.606
SINAI_1 R3	1	0.136	0.24
UC3M_1 R1	1	0.136	0.24
LSI_UNED R1	0.75	0.136	0.231
LSI_UNED R2	0.75	0.136	0.231
LSI_UNED R3	0.75	0.136	0.231
SINAI_1 R2	0.667	0.091	0.16
GPLSIUA_1 R1	0.5	0.091	0.154
GPLSIUA_1 R2	0.4	0.091	0.148
SINAI_1 R1	0.125	0.045	0.067

Table B.7: Negated disability recognition: Spanish - Partial matching. Precision (P), Recall (R) and F-measure (F).

English	P	R	F
IxaMed R1	1	0.913	0.955
UPC_3 R1	0.955	0.913	0.933
UPC_3 R2	0.955	0.913	0.933
UPC_3 R3	0.913	0.913	0.913
SINAI_1 R3	1	0.826	0.905
GPLSIUA_1 R1	0.941	0.696	0.8
UPC_2 R1	0.941	0.696	0.8
IXA_2 R1	1	0.652	0.789
GPLSIUA_1 R2	0.889	0.696	0.78
UC3M_1 R3	1	0.609	0.757
LSI_UNED R2	0.875	0.609	0.718
LSI_UNED R3	0.875	0.609	0.718
LSI_UNED R1	0.824	0.609	0.7
IXA_2 R2	1	0.522	0.686
SINAI_1 R1	0.556	0.87	0.678
SINAI_1 R2	0.556	0.87	0.678
UC3M_1 R2	0.875	0.304	0.452
UC3M_1 R1	1	0.043	0.083

Table B.8: Negated disability recognition: English - Partial matching. Precision (P), Recall (R) and F-measure (F).

Spanish	P	R	F
IxaMed R1	0.746	0.795	0.77
UC3M_1 R1	0.769	0.568	0.653
UPC_3 R2	0.772	0.563	0.652
UPC_3 R1	0.779	0.555	0.648
UC3M_1 R2	0.749	0.559	0.64
IXA_2 R1	0.644	0.616	0.629
IXA_2 R3	0.626	0.629	0.627
UC3M_1 R3	0.731	0.546	0.625
IXA_2 R2	0.633	0.594	0.613
UPC_3 R3	0.64	0.559	0.597
UPC_2 R1	0.71	0.48	0.573
SINAI_1 R3	0.411	0.284	0.336
LSI_UNED R3	0.424	0.245	0.31
LSI_UNED R2	0.409	0.245	0.306
LSI_UNED R1	0.406	0.245	0.305
SINAI_1 R2	0.157	0.349	0.217
GPLSIUA_1 R1	0.692	0.118	0.201
GPLSIUA_1 R2	0.659	0.118	0.2
SINAI_1 R1	0.018	0.402	0.035

Table B.9: Negated and non negated disability recognition: Spanish - Exact matching. Precision (P), Recall (R) and F-measure (F).

English	P	R	F
IxaMed R1	0.746	0.811	0.777
UC3M_1 R1	0.749	0.626	0.682
UPC_3 R1	0.772	0.584	0.665
UPC_3 R2	0.768	0.584	0.664
UC3M_1 R3	0.712	0.609	0.656
UC3M_1 R2	0.706	0.572	0.632
LSI_UNED R3	0.657	0.568	0.609
UPC_3 R3	0.626	0.593	0.609
UPC_2 R1	0.724	0.519	0.604
LSI_UNED R2	0.624	0.568	0.595
LSI_UNED R1	0.616	0.568	0.591
IXA_2 R1	0.672	0.49	0.567
IXA_2 R2	0.685	0.457	0.548
SINAI_1 R3	0.573	0.337	0.425
GPLSIUA_1 R2	0.806	0.239	0.368
GPLSIUA_1 R1	0.812	0.23	0.359
SINAI_1 R2	0.199	0.395	0.264
SINAI_1 R1	0.015	0.543	0.029

Table B.10: Negated and non negated disability recognition: English - Exact matching. Precision (P), Recall (R) and F-measure (F).

Spanish	P	R	F
IxaMed R1	0.82	0.873	0.846
UC3M_1 R3	0.889	0.664	0.76
UPC_3 R2	0.88	0.642	0.742
UC3M_1 R2	0.865	0.646	0.74
UPC_3 R1	0.89	0.633	0.74
UC3M_1 R1	0.864	0.638	0.734
IXA_2 R3	0.7	0.703	0.702
IXA_2 R1	0.708	0.677	0.692
UPC_3 R3	0.735	0.642	0.685
IXA_2 R2	0.693	0.651	0.671
UPC_2 R1	0.819	0.555	0.661
LSI_UNED R2	0.803	0.48	0.601
LSI_UNED R1	0.797	0.48	0.599
LSI_UNED R3	0.803	0.463	0.587
SINAI_1 R3	0.468	0.323	0.382
GPLSIUA_1 R2	0.878	0.157	0.267
GPLSIUA_1 R1	0.897	0.153	0.261
SINAI_1 R2	0.18	0.402	0.249
SINAI_1 R1	0.022	0.48	0.042

Table B.11: Negated and non negated disability recognition: Spanish - Partial matching. Precision (P), Recall (R) and F-measure (F).

English	P	R	F
IxaMed R1	0.841	0.914	0.876
LSI_UNED R3	0.843	0.728	0.781
UC3M_1 R3	0.832	0.712	0.767
LSI_UNED R2	0.801	0.728	0.763
LSI_UNED R1	0.79	0.728	0.758
UPC_3 R1	0.87	0.658	0.749
UPC_3 R2	0.859	0.654	0.743
UC3M_1 R2	0.817	0.663	0.732
UC3M_1 R1	0.803	0.671	0.731
UPC_3 R3	0.735	0.695	0.715
UPC_2 R1	0.822	0.588	0.686
IXA_2 R1	0.757	0.551	0.638
IXA_2 R2	0.784	0.523	0.627
SINAI_1 R3	0.685	0.403	0.508
GPLSIUA_1 R1	0.942	0.267	0.417
GPLSIUA_1 R2	0.903	0.267	0.413
SINAI_1 R2	0.242	0.481	0.322
SINAI_1 R1	0.019	0.691	0.037

Table B.12: Negated and non negated disability recognition: English - Partial matching. Precision (P), Recall (R) and F-measure (F).

Bibliography

[1] Rodrigo Agerri and German Rigau. Simple language independent sequence labelling for the annotation of disabilities in medical texts. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages co-located with 34th Conference of the Spanish Society for Natural Language Processing, Sevilla - Spain*, volume 2150 of *CEUR Workshop Proceedings*, pages 25–30. CEUR-WS.org, 2018.

[2] Rodrigo Agerri, Josu Bermudez, and German Rigau. IXA pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, Reykjavik - Iceland*, pages 3823–3828. ELRA, 2014.

[3] Ahmet Aker, Alfred Sliwa, Fahim Dalvi, and Kalina Bontcheva. Rumour verification through recurring information and an inner-attention mechanism. *Online Social Networks Media*, 13, 2019.

[4] Christoph Alt, Marc Hübner, and Leonhard Hennig. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1134. URL `https://www.aclweb.org/anthology/P19-1134`.

[5] Jorge Mederos Alvarado, Ernesto Quevedo Caballero, Alejandro Rodríguez Pérez, and Rocío Cruz Linares. UH-MAJA-KD at ehealth-kd challenge 2019: Deep learning models for knowledge discovery in spanish ehealth documents. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 85–94. CEUR-WS.org, 2019.

[6] Peggy M. Andersen, Philip J. Hayes, Steven P. Weinstein, Alison K. Huettner, Linda M. Schmandt, and Irene B. Nirenburg. Automatic Extraction of Facts from Press Releases to Generate News Stories. In *Third Conference on Applied Natural Language Processing*, pages 170–177, Trento, Italy, 1992. ACL.

[7] Gavin Andrews, Tim Slade, and Cathy Issakidis. Deconstructing current comorbidity: data from the australian national survey of mental health and well-being. *British Journal of Psychiatry*, 181(4):306–314, 2002.

[8] Chinatsu Aone, Lauren Halverson, Tom Hampton, and Mila Ramos-Santacruz. SRA: de-

scription of the IE2 system used for MUC-7. In *Seventh Message Understanding Conference: Proceedings of a Conference Held in Fairfax, Virginia - USA*. ACL, 1998.

[9] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings. AMIA Symposium*, pages 17–21, 2001.

[10] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montréal - Canada*, pages 86–90. Morgan Kaufmann Publishers / ACL, 1998.

[11] Javier Beltrán and Mónica González. Detection of negation cues in spanish: The clic-neg system. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 352–360. CEUR-WS.org, 2019.

[12] Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Padó. Germeval 2014 Named Entity Recognition Shared Task: Companion Paper. In *KONVENS 2014 / Workshop Proceedings of the 12th KONVENS*, 2014.

[13] Akash Bharadwaj, David R. Mortensen, Chris Dyer, and Jaime G. Carbonell. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472. ACL, 2016.

[14] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, 2017.

[15] Robert Bossy, Louise Deléger, Estelle Chaix, Mouhamadou Ba, and Claire Nédellec. Bacteria biotope at bionlp open shared tasks 2019. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, BioNLP-OST@EMNLP-IJNCLP 2019, Hong Kong, - China*, pages 121–131. ACL, 2019.

[16] Àlex Bravo, Pablo Accuosto, and Horacio Saggion. Lastus-taln at iberlef 2019 ehealth-kd challenge: Deep learning approaches to information extraction in biomedical texts. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 51–59. CEUR-WS.org, 2019.

[17] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

[18] Markus Bundschus, Mathäus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinform.*, 9, 2008.

[19] Cristian Cardellino. Spanish Billion Words Corpus and Embeddings, August 2019. URL `https://crscardellino.github.io/SBWCE/`.

[20] Arantza Casillas, Nerea Ezeiza, Iakes Goenaga, Alicia Pérez, and Xabier Soto. Measuring the effect of different types of unsupervised word representations on medical named entity recognition. *International Journal of Medical Informatics*, 129:100–106, 2019. ISSN 1386-5056.

doi: https://doi.org/10.1016/j.ijmedinf.2019.05.022. URL `https://www.sciencedirect.com/science/article/pii/S1386505618310311`.

[21] Neus Català and Mario Martín. Coin_flipper at ehealth-kd challenge 2019: Voting lstms for key phrases and semantic relation identification applied to spanish ehealth texts. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 17–25. CEUR-WS.org, 2019.

[22] Wendy W Chapman, Dieter Hilert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E Chapman, Michael Conway, Melissa Tharp, Danielle L Mowery, and Louise Deleger. Extending the negex lexicon for multiple languages. *Studies in health technology and informatics*, 192:677, 2013.

[23] Wendy Webber Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, 2001.

[24] Hai Leong Chieu and Hwee Tou Ng. Teaching a weaker classifier: Named entity recognition on upper case text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia - USA*, pages 481–488. ACL, 2002.

[25] Hyejin Cho and Hyunju Lee. Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics*, 20(1):735, 2019.

[26] Md. Faisal Mahbub Chowdhury and Alberto Lavelli. Fbk-irst : A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information. In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta - USA*, pages 351–355. ACL, 2013.

[27] Alexander Clark. Combining distributional and morphological information for part of speech induction. In *EACL 2003, 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest - Hungary*, pages 59–66. ACL, 2003.

[28] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning, Helsinki - Finland*, ICML '08, page 160–167. ACM, 2008.

[29] Sandra Collovini, Joaquim Francisco Santos Neto, Bernardo Scapini Consoli, Juliano Terra, Renata Vieira, Paulo Quaresma, Marlo Souza, Daniela Barreiro Claro, and Rafael Glauber. Iberlef 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 390–410. CEUR-WS.org, 2019.

[30] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995.

[31] Viviana Cotik, Vanesa Stricker, Jorge Vivaldi, and Horacio Rodríguez Hontoria. Syntactic methods for negation detection in radiology reports in spanish. In *Proceedings of the 15th Workshop on BioNLP 2016: Berlin - Germany*, pages 156–165. ACL, 2016.

[32] Isaac G. Councill, Ryan T. McDonald, and Leonid Velikovich. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of*

the Workshop on Negation and Speculation in Natural Language Processing, NeSp-NLP@ACL 2010, Uppsala - Sweden, pages 51–59. University of Antwerp, 2010.

[33] Alessandro Cucchiarelli, Danilo Luzi, and Paola Velardi. Automatic semantic tagging of unknown proper names. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics: Proceedings of the Conference, Quebec - Canada*, pages 286–292. Morgan Kaufmann Publishers / ACL, 1998.

[34] Miguel Ángel García Cumbreras, Julio Gonzalo, Eugenio Martínez Cámara, Raquel Martínez-Unanue, Paolo Rosso, Jorge Carrillo-de-Albornoz, Soto Montalvo, Luis Chiruzzo, Sandra Collovini, Yoan Gutiérrez, Salud M. Jiménez Zafra, Martin Krallinger, Manuel Montes-y-Gómez, Reynier Ortega-Bueno, and Aiala Rosá, editors. *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, volume 2421 of *CEUR Workshop Proceedings*, 2019. CEUR-WS.org.

[35] Arjun Das, Debasis Ganguly, and Utpal Garain. Named entity recognition with word embeddings and wikipedia categories for a low-resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 16(3):18:1–18:19, 2017.

[36] *Proceedings of the 6th Conference on Message Understanding, Columbia - USA*, 1995. Defense Advanced Research Projects Agency and Software and Intelligent Systems Technology Office, ACL.

[37] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

[38] I. N. Dewi, S. Dong, and J. Hu. Drug-drug interaction relation extraction with deep convolutional neural networks. In *2017 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1795–1802, 2017.

[39] Daniela Oliveira Ferreira do Amaral, Maiki Buffet, and Renata Vieira. Comparative analysis between notations to classify named entities using conditional random fields. In *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology, STIL 2015, Natal - Brazil*, pages 27–31. Sociedade Brasileira de Computação, 2015.

[40] Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, 2014.

[41] Lluís Domínguez-Mas, Francesco Ronzano, and Laura I. Furlong. Supervised learning approaches to detect negation cues in spanish reviews. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 361–368. CEUR-WS.org, 2019.

[42] John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online

learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[43] Martine Enger, Erik Velldal, and Lilja Øvrelid. An open-source tool for negation detection: a maximum-margin approach. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 64–69, Valencia, Spain, 2017. ACL.

[44] Hermenegildo Fabregat, Lourdes Araujo, and Juan Martínez-Romo. Deep neural models for extracting entities and relationships in the new RDD corpus relating disabilities and rare diseases. *Computer Methods and Programs in Biomedicine*, 164:121–129, 2018.

[45] Hermenegildo Fabregat, Juan Martínez-Romo, and Lourdes Araujo. Overview of the DIANN task: Disability annotation task. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages co-located with 34th Conference of the Spanish Society for Natural Language Processing, Sevilla - Spain*, volume 2150 of *CEUR Workshop Proceedings*, pages 1–14. CEUR-WS.org, 2018.

[46] Hermenegildo Fabregat, Lourdes Araujo, and Juan Martínez-Romo. Deep learning approach for negation trigger and scope recognition. *Procesamiento del Lenguaje Natural*, 62:37–44, 2019.

[47] Hermenegildo Fabregat, Andrés Duque, Juan Martínez-Romo, and Lourdes Araujo. De-identification through named entity recognition for medical document anonymization. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 663–670. CEUR-WS.org, 2019.

[48] Hermenegildo Fabregat, Andrés Duque, Juan Martínez-Romo, and Lourdes Araujo. Extending a deep learning approach for negation cues detection in spanish. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 369–377. CEUR-WS.org, 2019.

[49] Hermenegildo Fabregat, Andres Duque Fernandez, Juan Martínez-Romo, and Lourdes Araujo. Nlp_uned at ehealth-kd challenge 2019: Deep learning for named entity recognition and attentive relation extraction. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 67–77. CEUR-WS.org, 2019.

[50] Hermenegildo Fabregat, Juan Martínez-Romo, and Lourdes Araujo. Understanding and improving disability identification in medical documents. *IEEE Access*, 8:155399–155408, 2020.

[51] Federico Fancellu, Adam Lopez, and Bonnie L. Webber. Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin - Germany*, volume 1. ACL, 2016.

[52] Federico Fancellu, Adam Lopez, Bonnie L. Webber, and Hangfeng He. Detecting negation scope is easy, except when it isn't. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia - Spain*, volume 2, pages 58–63. ACL, 2017.

[53] K Fukuda, A Tamura, T Tsunoda, and T Takagi. Toward information extraction: identifying

protein names from biological papers. *Pacific Symposium on Biocomputing*, page 707—718, 1998.

[54] Ken-ichiro Fukuda, Tatsuhiko Tsunoda, Ayuchi Tamura, Toshihisa Takagi, et al. Toward information extraction: identifying protein names from biological papers. In *Pac symp biocomput*, volume 707, pages 707–718, 1998.

[55] Jun-ichi Fukumoto, Fumito Masui, M. Shimcheta, and M. Saski. Description of the oki system as used for MUC-7. In *Seventh Message Understanding Conference: Proceedings of a Conference Held in Fairfax, Virginia - USA*. ACL, 1998.

[56] Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *11th Conference of the European Chapter of the Association for Computational Linguistics, Trento - Italy*. ACL, 2006.

[57] Talmy Givón. *English grammar: A function-based introduction*, volume 2. John Benjamins Publishing, 1993.

[58] Iakes Goenaga, Aitziber Atutxa, Koldo Gojenola, Arantza Casillas, Arantza Díaz de Ilarraza, Nerea Ezeiza, Maite Oronoz, Alicia Pérez, and Olatz Perez-de-Viñaspre. A hybrid approach for automatic disability annotation. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages co-located with 34th Conference of the Spanish Society for Natural Language Processing, Sevilla - Spain*, volume 2150 of *CEUR Workshop Proceedings*, pages 31–36. CEUR-WS.org, 2018.

[59] Iakes Goenaga, Sergio Santana, Sara Santiso Gonzáles, Koldo Gojenola, Alicia Pérez, and Arantza Casillas. Ixamed at ehealth-kd challenge 2019: Using different paradigms to solve clinical relation extraction. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 43–50. CEUR-WS.org, 2019.

[60] I Goldin and Wendy W Chapman. Learning to detect negation with 'not'in medical texts. In *Proceedings of the Workshop on Text Analysis and Search for Bioinformatics, ACM SIGIR*. ACM SIGIR, 2003.

[61] Aitor Gonzalez Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10. ACL, 2019.

[62] Sergey Goryachev, Margarita Sordo, Qing T Zeng, and Long Ngo. Implementation and evaluation of four different methods of negation detection. *Boston, MA: DSG*, 2006.

[63] Ralph Grishman and Beth Sundheim. Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 1 of *COLING '96*, page 466–471, USA, 1996. ACL.

[64] Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. Table filling multi-task recurrent neural network for joint entity and relation extraction. In *Proceedings of COLING 2016, 26th International Conference on Computational Linguistics, Osaka - Japan*, pages 2537–2547. ACL, 2016.

[65] Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a benchmark corpus to support the automatic

extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892, 2012.

[66] Kai Hakala and Sampo Pyysalo. Biomedical named entity recognition with multilingual BERT. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5709. URL `https://www.aclweb.org/anthology/D19-5709`.

[67] Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851, 2009.

[68] Fadi Hassan, Mohammed Jabreel, Najlaa Maaroof, David Sánchez, Josep Domingo-Ferrer, and Antonio Moreno. Recrf: Spanish medical document anonymization using automatically-crafted rules and CRF. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 727–734. CEUR-WS.org, 2019.

[69] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38. ACL, 2010.

[70] María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics*, 46(5):914–920, 2013. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2013.07.011. URL `https://www.sciencedirect.com/science/article/pii/S1532046413001123`.

[71] R. Herwando, M. A. Jiwanggi, and M. Adriani. Medical entity recognition using conditional random field (crf). In *2017 International Workshop on Big Data and Information Security*, pages 57–62, 2017.

[72] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8):1735–1780, 1997.

[73] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Learning and Intelligent Optimization - 5th International Conference, Rome - Italy*, volume 6683 of *Lecture Notes in Computer Science*, pages 507–523. Springer, 2011.

[74] Hideki Isozaki and Hideto Kazawa. Efficient Support Vector Classifiers for Named Entity Recognition. In *Proceedings of the 19th International Conference on Computational Linguistics*, volume 1 of *COLING '02*, page 1–7, USA, 2002. ACL.

[75] Mohammed Jabreel, Fadi Hassan, Najlaa Maaroof, David Sánchez, Josep Domingo-Ferrer, and Antonio Moreno. E2EJ: anonymization of spanish medical records using end-to-end joint neural networks. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 712–719. CEUR-WS.org, 2019.

[76] Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu. Overview of the First Natural

Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0). *Drug Safety*, 42(1):99–111, 2019.

[77] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[78] Z. Ju, J. Wang, and F. Zhu. Named entity recognition from biomedical text using svm. In *2011 5th International Conference on Bioinformatics and Biomedical Engineering*, pages 1–4, 2011.

[79] U. Kanimozhi and D. Manjula. A crf based machine learning approach for biomedical named entity recognition. In *2017 Second International Conference on Recent Trends and Challenges in Computational Models*, pages 335–342, 2017.

[80] Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta, and Jun'ichi Tsujii. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain, Philadelphia - USA*, pages 1–8. ACL, 2002.

[81] Martijn G. Kersloot, Francis Lau, Ameen Abu Hanna, Derk L. Arts, and Ronald Cornet. Automated SNOMED CT concept and attribute relationship detection through a web-based implementation of cTAKES. *Journal of Biomedical Semantics*, 10(1):14, 2019.

[82] Ji-Hwan Kim and Philip C Woodland. A rule-based named entity recognition system for speech input. In *Sixth International Conference on Spoken Language Processing*, 2000.

[83] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Conference Track Proceedings of 3rd International Conference on Learning Representations, San Diego - USA*, 2015.

[84] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[85] Paul R. Kingsbury and Martha Palmer. From treebank to propbank. In *Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas - Spain*. ELRA, 2002.

[86] Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. A review corpus annotated for negation, speculation and their scope. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 2012. ELRA.

[87] George R. Krupka and Kevin Hausman. Isoquest inc.: Description of the netowl. In *Seventh Message Understanding Conference: Proceedings of a Conference Held in Fairfax, Virginia - USA*. ACL, 1998.

[88] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown - USA*, pages 282–289. Morgan Kaufmann, 2001.

[89] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego - USA, pages 260–270. ACL, 2016.

[90] Lukas Lange, Heike Adel, and Jannik Strötgen. NLNDE: the neither-language-nor-domain-experts' way of spanish medical document de-identification. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 671–678. CEUR-WS.org, 2019.

[91] Emanuele Lapponi, Erik Velldal, Lilja Øvrelid, and Jonathon Read. Uio 2: Sequence-labeling negation using dependency features. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics, *SEM 2012, Montréal - Canada*, pages 319–327. ACL, 2012.

[92] Lydia Lazib, Yanyan Zhao, Bing Qin, and Ting Liu. Negation scope detection with recurrent neural networks models in review texts. In *Proceedings, Part I: Social Computing - Second International Conference of Young Computer Scientists, Engineers and Educators, Harbin - China*, volume 623 of *Communications in Computer and Information Science*, pages 494–508. Springer, 2016.

[93] Robert Leaman and Zhiyong Lu. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, 32(18):2839–2846, 2016.

[94] Joohong Lee, Sangwoo Seo, and Yong Suk Choi. Semantic relation classification via bidirectional LSTM networks with entity-aware attention using latent entity typing. *Symmetry*, 11 (6):785, 2019.

[95] Fernando Sánchez León. Resource-based anonymization for spanish clinical cases. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 704–711. CEUR-WS.org, 2019.

[96] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore - USA*, volume 2, pages 302–308. ACL, 2014.

[97] Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics*, 18(1):198:1–198:11, 2017.

[98] Hao Li and Wei Lu. Learning with structured representations for negation scope extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne - Australia*, volume 2, pages 533–539. ACL, 2018.

[99] Pengfei Li and Kezhi Mao. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*, 115:512–523, 2019.

[100] D. A. Lindberg, B. L. Humphreys, and A. T. McCray. The Unified Medical Language System. *Methods of information in medicine*, 32(4):281–291, 1993.

[101] Henry Loharja, Lluìs Padró, and Jorge Turmo Borras. Negation cues detection using crf on spanish product review texts. In *NEGES 2018: Workshop on Negation in Spanish: Proceedings book. Seville - Spain.*, pages 49–54, 2018.

[102] Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, María Teresa Martín-Valdivia, and Salud María Jiménez Zafra. SINAI at DIANN - ibereval 2018. annotating disabilities in multi-language systems with UMLS. In *Proceedings of the Third Workshop on Evaluation of Human*

Language Technologies for Iberian Languages co-located with 34th Conference of the Spanish Society for Natural Language Processing, Sevilla - Spain, volume 2150 of *CEUR Workshop Proceedings*, pages 37–43. CEUR-WS.org, 2018.

[103] Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. Automatic de-identification of medical texts in spanish: the MEDDOCAN track, corpus, guidelines, methods and evaluation of results. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 618–638. CEUR-WS.org, 2019.

[104] Alexa T. McCray. The umls semantic network. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 503—507, 1989.

[105] Salvador Medina and Jordi Turmo. TALP-UPC at ehealth-kd challenge 2019: A joint model with contextual embeddings for clinical information extraction. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 78–84. CEUR-WS.org, 2019.

[106] Salvador Medina, Jordi Turmo, Henry Loharja, and Lluís Padró. Semi-supervised learning for disabilities detection on english and spanish biomedical text. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages co-located with 34th Conference of the Spanish Society for Natural Language Processing, Sevilla - Spain*, volume 2150 of *CEUR Workshop Proceedings*, pages 66–73. CEUR-WS.org, 2018.

[107] Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M. Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C. Max Schmidt, Hongfang Liu, and Mathew Palakal. Deepen: A negation detection system for clinical text incorporating dependency relation into negex. *Journal of Biomedical Informatics*, 54:213–219, 2015.

[108] Adam Meyers, Catherine Macleod, Roman Yangarber, Ralph Grishman, Leslie Barrett, and Ruth Reeves. Using NOMLEX to produce nominalization patterns for information extraction. In *The Computational Treatment of Nominals*, 1998.

[109] Andrei Mikheev, Marc Moens, and Claire Grover. Named Entity Recognition without Gazetteers. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, page 1–8, USA, 1999. ACL.

[110] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Workshop Track Proceedings of the 1st International Conference on Learning Representations, Scottsdale - USA*, 2013.

[111] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38 (11):39–41, 1995.

[112] Scott Miller, Michael Crystal, Heidi Fox, Lance A. Ramshaw, Richard M. Schwartz, Rebecca Stone, and Ralph M. Weischedel. BBN: description of the SIFT system as used for MUC-7. In *Seventh Message Understanding Conference: Proceedings of a Conference Held in Fairfax, Virginia - USA*. ACL, 1998.

[113] Kevin J. Mitchell, Michael J. Becich, Jules J. Berman, Wendy W. Chapman, John R. Gilbertson, Dilip Gupta, James Harrison, Elizabeth Legowski, and Rebecca S. Crowley. Implementation

and evaluation of a negation tagger in a pipeline-based system for information extraction from pathology reports. In *MEDINFO 2004 - Proceedings of the 11th World Congress on Medical Informatics, San Francisco -USA*, volume 107 of *Studies in Health Technology and Informatics*, pages 663–667. IOS Press, 2004.

[114] Soto Montalvo, Maite Oronoz, Horacio Rodríguez Hontoria, and Raquel Martínez. Biomedical abbreviation recognition and resolution by prosa-med. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages co-located with 33th Conference of the Spanish Society for Natural Language Processing: Murcia - Spain*, pages 247–254. CEUR-WS. org, 2017.

[115] Roser Morante and Eduardo Blanco. *sem 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics, Montréal - Canada*, pages 265–274. ACL, 2012.

[116] Roser Morante and Walter Daelemans. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the BioNLP Workshop, BioNLP@HLT-NAACL 2009, Boulder - USA*, pages 28–36. ACL, 2009.

[117] Isabel Moreno, María Teresa Romá-Ferri, and María Paloma Moreda Pozo. A domain and language independent named entity classification approach based on profiles and local information. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna - Bulgaria*, pages 510–518. INCOMA Ltd., 2017.

[118] Isabel Moreno, María Teresa Romá-Ferri, and Paloma Moreda. GPLSIUA team at the DIAAN 2018 task. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages co-located with 34th Conference of the Spanish Society for Natural Language Processing, Sevilla - Spain*, volume 2150 of *CEUR Workshop Proceedings*, pages 15–24. CEUR-WS.org, 2018.

[119] Kerim M. Munir. The co-occurrence of mental disorders in children and adolescents with intellectual disability/intellectual developmental disorder. *Current Opinion in Psychiatry*, 29 (2), 2016.

[120] Pradeep G. Mutalik, Aniruddha Deshpande, and Prakash M. Nadkarni. Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents: A Quantitative Study Using the UMLS. *Journal of the American Medical Informatics Association*, 8(6): 598–609, 2001.

[121] Isar Nejadgholi, Kathleen C. Fraser, and Berry de Bruijn. Extensive Error Analysis and a Learning-Based Evaluation of Medical Entity Recognition Systems to Approximate User Experience. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 177–186. ACL, 2020.

[122] World Health Organization. *International classification of functioning, disability, and health : children & youth version : ICF-CY.* WHO, 2007.

[123] World Health Organization et al. *International classification of impairments, disabilities, and handicaps: a manual of classification relating to the consequences of disease, published in accordance with resolution WHA29. 35 of the Twenty-ninth World Health Assembly, May 1976.* WHO, 1980.

[124] World Health Organization et al. *International classification of functioning, disability and health: ICF*. WHO, 2001.

[125] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Doha - Qatar: A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014.

[126] Naiara Pérez, Laura García-Sardiña, Manex Serras, and Arantza del Pozo. Vicomtech at MEDDOCAN: medical document anonymization. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 696–703. CEUR-WS.org, 2019.

[127] Alejandro Piad-Morffis, Yoan Gutiérrez, Juan Pablo Consuegra-Ayala, Suilan Estevez-Velarde, Yudivián Almeida-Cruz, Rafael Muñoz, and Andrés Montoyo. Overview of the ehealth knowledge discovery challenge at iberlef 2019. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 1–16. CEUR-WS.org, 2019.

[128] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*, pages 39–44. LBM, 2013.

[129] Zhong Qian, Peifeng Li, Qiaoming Zhu, Guodong Zhou, Zhunchen Luo, and Wei Luo. Speculation and negation scope detection via convolutional neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 815–825. ACL, 2016.

[130] Changqin Quan, Meng Wang, and Fuji Ren. An unsupervised text mining method for relation extraction from biomedical literature. *PLOS ONE*, 9(7):1–8, 2014.

[131] Lance A. Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora, VLC@ACL 1995, Cambridge - USA*, 1995.

[132] Marek Rei, Gamal K. O. Crichton, and Sampo Pyysalo. Attending to characters in neural sequence labeling models. In *26th International Conference on Computational Linguistics - Proceedings of the Conference: Technical Papers*, pages 309–318. ACL, 2016.

[133] I Rigoutsos and A Floratos. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics*, 14(1):55–67, 1998.

[134] Bryan Rink and Sanda M. Harabagiu. UTD: classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala - Sweden*, pages 256–259. ACL, 2010.

[135] Angus Roberts, Robert J. Gaizauskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay (Subbarao) Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, and Bill Wheeldin. The CLEF corpus: Semantic annotation of clinical text. In *AMIA 2007, American Medical Informatics Association Annual Symposium, Chicago - USA*. AMIA, 2007.

[136] Lior Rokach, Roni Romano, and Oded Maimon. Negation recognition in medical narrative reports. *Inf. Retr.*, 11(6):499–538, 2008.

[137] Paolo Rosso, Julio Gonzalo, Raquel Martínez, Soto Montalvo, and Jorge Carrillo de Albornoz, editors. *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages co-located with 34th Conference of the Spanish Society for Natural Language Processing, Sevilla - Spain*, volume 2150 of *CEUR Workshop Proceedings*, 2018.

[138] Alejandro Ruiz-de-laCuadra, José Luis López Cuadrado, Israel González-Carrasco, and Belén Ruíz-Mezcua. Hulat-taska at ehealth-kd challenge 2019: Sequence key phrases recognition in the spanish clinical narrative. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 26–34. CEUR-WS.org, 2019.

[139] Sunita Sarawagi and William W Cohen. Semi-markov conditional random fields for information extraction. In *Advances in Neural Information Processing Systems 17*, pages 1185–1192. NIPS, 2005.

[140] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[141] Arrigo Schieppati, Jan-Inge Henter, Erica Daina, and Anita Aperia. Why rare diseases are an important medical and social issue. *The Lancet*, 371(9629):2039–2041, 2008.

[142] Kent A. Spackman, Keith E. Campbell, and Roger A. Côté. SNOMED RT: a reference terminology for health care. In *AMIA 1997, American Medical Informatics Association Annual Symposium, Nashville, TN, USA*. AMIA, 1997.

[143] Suzanne Steffenburg, Christopher L. Gillberg, Ulf Steffenburg, and Mårten Kyllerman. Autism in angelman syndrome: a population-based study. *Pediatric Neurology*, 14(2):131 – 136, 1996.

[144] Víctor Suárez-Paniagua. VSP at ehealth-kd challenge 2019: Recurrent neural networks for relation classification in spanish ehealth documents. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 95–104. CEUR-WS.org, 2019.

[145] Cong Sun and Zhihao Yang. Transfer learning in biomedical named entity recognition: An evaluation of BERT in the PharmaCoNER task. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 100–104, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5715. URL `https://www.aclweb.org/anthology/D19-5715`.

[146] Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *COLING-02: The 6th Conference on Natural Language Learning)*, 2002.

[147] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003.

[148] G. Tottie. *Negation in English Speech and Writing: A Study in Variation*. Quantitative analyses of linguistic structure. Academic Press, 1991.

[149] Richard Tzong-Han Tsai, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang,

Ting-Yi Sung, and Wen-Lian Hsu. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7:92, 2006.

[150] Kiyotaka Uchimoto, Qing Ma, Masaki Murata, Hiromi Ozaku, and Hitoshi Isahara. Named entity extraction based on A maximum entropy model and transformation rules. In *38th Annual Meeting of the Association for Computational Linguistics, Hong Kong - China*, pages 326–335. ACL, 2000.

[151] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, Long Beach - USA*, pages 5998–6008, 2017.

[152] Pol Alvarez Vecino and Lluís Padró. Basic CRF approach to DIANN 2018 shared task. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages co-located with 34th Conference of the Spanish Society for Natural Language Processing, Sevilla - Spain*, volume 2150 of *CEUR Workshop Proceedings*, pages 61–65. CEUR-WS.org, 2018.

[153] Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S-11), 2008.

[154] Charlotte von der Lippe, Plata S. Diesen, and Kristin B. Feragen. Living with a rare disorder: a systematic review of the qualitative literature. *Molecular genetics & genomic medicine*, 5 (6):758–773, 2017.

[155] Piek Vossen. Introduction to eurowordnet. *Computers and the Humanities*, 32(2-3):73–89, 1998.

[156] Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin - Germany*, volume 1. ACL, 2016.

[157] Steffanie S Weinreich, R Mangon, JJ Sikkens, ME Teeuw, and MC Cornel. Orphanet: a european database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, 152(9):518–519, 2008.

[158] Yorick Wilks. Information Extraction as a Core Language Technology. In *International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, SCIE '97, page 1–9, Berlin, Heidelberg, 1997. Springer-Verlag.

[159] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data mining: practical machine learning tools and techniques, 3rd Edition*. Morgan Kaufmann, Elsevier, 2011.

[160] Kui Xue, Yangming Zhou, Zhiyuan Ma, Tong Ruan, Huanhuan Zhang, and Ping He. Fine-tuning BERT for joint entity and relation extraction in chinese medical text. In Illhoi Yoo, Jinbo Bi, and Xiaohua Hu, editors, *2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019, San Diego, CA, USA, November 18-21, 2019*, pages 892–897. IEEE, 2019. doi: 10.1109/BIBM47256.2019.8983370. URL `https://doi.org/10.1109/BIBM47256.2019.8983370`.

[161] Alexander Yates, Michele Banko, Matthew Broadhead, Michael J. Cafarella, Oren Etzioni, and Stephen Soderland. Textrunner: Open information extraction on the web. In *Human Language*

Technology Conference of the North American Chapter of the Association of Computational Linguistics, Rochester - USA, pages 25–26. ACL, 2007.

[162] Salud M. Jiménez Zafra, Mariona Taulé, María Teresa Martín-Valdivia, Luis Alfonso Ureña López, and Maria Antònia Martí. SFU reviewsp-neg: a spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. *Language Resources and Evaluation*, 52(2):533–569, 2018.

[163] Salud María Jiménez Zafra, Noa Patricia Cruz Díaz, Roser Morante, and María Teresa Martín Valdivia. NEGES 2019 task: Negation in spanish. In *Proceedings of IberLEF co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao - Spain*, volume 2421 of *CEUR Workshop Proceedings*, pages 329–341. CEUR-WS.org, 2019.

[164] Salud María Jiménez Zafra, Roser Morante, Eduardo Blanco, María Teresa Martín Valdivia, and Luis Alfonso Ureña López. Detecting negation cues and scopes in spanish. In *Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020, Marseille - France*, pages 6902–6911. ELRA, 2020.

[165] Renzo M. Rivera Zavala, Paloma Martínez, and Isabel Segura-Bedmar. A hybrid bi-lstm-crf model to recognition of disabilities from biomedical texts. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages co-located with 34th Conference of the Spanish Society for Natural Language Processing, Sevilla - Spain*, volume 2150 of *CEUR Workshop Proceedings*, pages 44–52. CEUR-WS.org, 2018.

[166] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. E. Hinton. On rectified linear units for speech processing. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3517–3521, 2013.

[167] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Philadelphia - USA*, pages 71–78, 2002.

[168] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *CoRR*, abs/1509.01626, 2015.

[169] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin - Germany*, volume 2. ACL, 2016.

[170] Bowei Zou, Qiaoming Zhu, and Guodong Zhou. Negation and speculation identification in chinese language. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Beijing - China*, volume 1, pages 656–665. ACL, 2015.