UNIVERSIDAD DE JAÉN

Doctoral thesis

Biomedical entities recognition in Spanish combining word embeddings

Jaén, Spain

April, 2021

Presented by:

Pilar López-Úbeda

Supervised by:

PhD. M. Teresa Martín-Valdivia

PhD. L. Alfonso Ureña-López

PhD. Manuel Carlos Díaz-Galiano

Abstract

The advent of the Internet, and more specifically Web 2.0, has contributed to the proliferation of large volumes of unstructured information available digitally. The growth of electronic data is especially important in specific areas such as biomedicine, where the number of published documents (articles, clinical and technical reports, among others) is increasing exponentially. In order to organize and manage this data, several manual curation efforts have been made to identify relevant information in the texts. However, manual review of these documents for clinical knowledge extraction is costly and time-consuming. One of the main objectives of Natural Language Processing (NLP) is to facilitate these tasks by proposing automated methods for optimizing the workflow of healthcare professionals. Specifically, automated systems can help healthcare professionals as decision support systems and by managing patients' medical data in a short time. In fact, the application of NLP in the field of biomedicine has attracted the attention of the research community in recent years due to the development of interesting systems showing the advantages of using NLP techniques in this field.

In biomedical text mining, Named Entity Recognition (NER) is an important task in the field of NLP used to extract significant knowledge from textual

documents. The goal of NER is to identify chunks of text that refer to specific entities of interest, such as protein names, drugs, symptoms, and diseases, reporting to medical experts a large amount of the knowledge embedded in the textual data.

On the other hand, machine learning and deep learning methods have shown significant improvements in several NLP tasks such as machine translation and text generation. In this thesis we aim to take advantage of this technology and apply it to the biomedical NER task in Spanish. To accomplish this goal, we propose a model based on neural networks that is able to process the text included in health documents. The neural network architecture is composed of a Bidirectional Long Short Term Memory (BiLSTM) with a layer of Conditional Random Field (CRF) to predict each word as a proper entity. To represent the text we employ a combination of word embeddings providing knowledge to each word according to the combination selected. Moreover, we generate new domain and language-specific word embeddings to test their effectiveness. This approach is evaluated in three scenarios of different biomedical sub-domains. Finally, we demonstrate that the combination of different word embeddings as input to the neural network improves the state-of-the-art results in the applied scenarios.

Acknowledgments

It is a great pleasure to express my respect to the many people who have supported me throughout my doctoral study at the University of Jaén.

First of all, I would like to thank my supervisors Maite, Alfonso, and Manuel Carlos. They kindly guided me and gave me the best advice for my professional and personal future. Without them, I would not have been able to achieve such fruitful results.

I am grateful to my colleagues of the SINAI group, Fernando, Arturo, Miguel Ángel, Manolo, Flor, Salud, and Loles, who have been a point of reference in my thesis. A special thanks to my friend Flor for all the discussions, the new ideas, the work together, and the valuable time we have spent together.

I could not miss the opportunity to thank the external researchers for their work. Stefan Schulz, Alexandra Pomares, Antonio Luna, Teodoro Martín, Pablo López, Jose Perea, and Michel Oleynik have been excellent collaborators.

Finally, I would like to express my sincere gratitude and appreciation to my parents, my sister and my brother-in-law, who support all my decisions, believing in me with much affection and love. A special thanks to Jose, my life partner, who helped me overcome all the difficulties of this adventure with positivity, love and unceasing encouragement.

Table of Contents

List of Tables

List of Figures

Abbreviations

General notation

c.f	*conf er/conf eratur* (english: compare)
e.g.	*exemplum gratia* (english: for example)
i.e.	*id est* (english: that is)

Biomedical notation

EHR	Electronic Health Records
ADE	Adverse Drug Event
COVID-19	COronaVIrus Disease 2019
CT	Computed Tomography
DDI	Drug-Drug Interactions
UMLS	Unified Medical Language System
GO	Gene Ontology
SNOMED-CT	Systematized Nomenclature of Medicine - Clinical Terms
MeSH	Medical Subject Headings
NLM	National Library of Medicine
ICD	International Classification of Diseases
EMEA	European MEdicines Agency
WHO	World Health Organization

IARC	International Agency for Research on Cancer

Natural language processing

AI	Artificial Intelligence
NLP	Natural Language Processing
IE	Information Extraction
NER	Named Entity Recognition
Tf-Idf	Term frequency – Inverse document frequency
NED	Named Entity Disambiguation
IR	Information Retrieval
CDS	Clinical Decision Support
OOV	Out-Of-Vocabulary
BOW	Bag-Of-Word
SNE	Stochastic Neighbor Embedding
WE	Word Embedding

Machine learning

ML	Machine Learning
CRF	Conditional Random Field
NB	Naive Bayes
LR	Logistic Regression
SVM	Support Vector Machine
LDA	Latent Dirichlet Allocation

Deep learning

RNN	Recurrent Neural Network

FNN	Feedforward Neural Network
SLP	Single-Layer Perceptron
MLP	Multi-Layer Perceptron
LSTM	Long Short-Term Memory
MSE	Mean Squared Error
MAPE	Mean Absolute Percentage Error
MAE	Mean Absolute Error
ReLU	Rectified Linear Unit
SGD	Stochastic Gradient Descent
CNN	Convolutional Neural Network
GRU	Gated Recurrent Units

Transfer learning

MLM	Masked Language Modelling
CLM	Causal Language Modeling
TLM	Translation Language Modeling
SPM	Sentence Piece model

Chapter 1

Introduction

One of the main purposes of clinical text mining is the possibility to process and analyze the large volumes of textual information contained in medical records. Through this treatment of the information, we attempt to answer questions such as, which patients presented a certain condition? What kind of conditions were used to detect the disease? What were the results of the tests performed? What was the treatment given? These questions could seem quite simple for some medical professionals, but they become extremely complex when managed automatically by computational systems.

In the biomedical domain, we can find large collections of free textual information (medical reports, Electronic Health Records - EHR, scientific papers, among others) that contain very relevant data that need to be studied in depth. However, current health information systems are not prepared to analyze and extract this knowledge due to the time and cost involved in processing it manually. The field of artificial intelligence known as Natural Language Processing (NLP) is being applied to medical documents to build applications that can understand and analyze this huge amount of textual

information automatically [1]. Following the definition by Chowdhury [2]:

> *"NLP is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things."*

Many researchers in the NLP field focus on the area of Information Extraction (IE) in the biomedical domain to address these challenges. IE systems take natural language text as input and produce structured information specified by certain criteria and that is relevant to a particular application. Depending on the different inputs of IE systems and expected outputs, many sub-tasks can be defined such as Named Entity Recognition (NER).

The term Named Entity was established in 1996, at the 6th Message Understanding Conference (MUC-6), to refer to "unique identifiers of entities" [3]. In broad terms, the NER task consists of locating and classifying parts of the text into pre-defined categories such as places, people, organizations, expressions of time, and quantities. However, in the biomedical domain, the important entities included in documents are not limited to those mentioned above. In this particular case, it is necessary to recognize some special types of named entities, such as diseases, procedures, treatments, and drugs, among others.

NER not only serves as a task included in IE but also plays an essential role in a variety of NLP applications such as information retrieval, automatic text summarizing, or question answering [4]. Additionally, the recognition of biomedical entities in a text can be a starting point for the subsequent extraction of relationships between entities, allowing these concepts to be represented in some coherent and standardized form.

In this thesis, we focus on information extraction from Spanish biomedical texts, more specifically, on the NER task. Spanish has more than 480 million native speakers[1] and nowadays there is a worldwide interest in processing medical texts in this language. With this study, we aim to advance the task of biomedical NER in this relevant language and thus answer the above-mentioned questions.

In order to accomplish this study, we propose a methodology based on deep learning. Furthermore, different word embeddings are used in combination to obtain a better representation of each word. With this approach, we aim to achieve the desired final goal: to recognize biomedical entities accurately in different scenarios.

In the following sections, we show the motivations that have led us to address the NER task in the biomedical field. We also describe the objectives and goals that we intend to achieve with this thesis. Subsequently, we formulate the hypotheses which will serve as a basis for initiating the research. And finally, we explain the methodology followed during the research process.

1.1 Motivation

Over the years, the recognition of biomedical entities has motivated the scientific community to continue developing automatic systems to facilitate the extraction of medical knowledge. NER is a difficult task to solve that can help in many other medical-related systems such as those presented below:

- **Clinical decision support**. Clinical decision support emphasizes the

[1]Spanish language: `https://en.wikipedia.org/wiki/Spanish_language`

ability to produce evidence-based reports on daily health services to assist experts in their decisions and actions [5]. This information can be used by NLP methods that develop evidence-based applications detecting early warning for the monitoring of disorders and the development of personalized patient care [6, 7]. Automated NER systems can provide real-time results, which means that entities such as diseases can be detected immediately. This evidence can be used to help professionals identify emerging health problems, for instance, to alert them to the presence of certain unexpected findings [8].

- **Entity representation**. In the NER task, different words can have similar meanings. This problem is caused by the multiple ways in which a particular entity can be represented and written. For instance, "*adriamicina*" (adriamycin), "*doxorrubicina*" (doxorubicin) and "*hidroxildaunorrubicina*" (hydroxyldaunorubicin) refer to the same drug widely used in cancer chemotherapy. Another consideration is that entities appear as acronyms or their descriptions, e.g. "*enfermedad pulmonar obstructiva crónica*" (chronic obstructive pulmonary disease) and "*EPOC*" (COPD) are referred to as the same disease (chronic inflammatory lung disease causing airflow obstruction of the lungs).

  On the other hand, an acronym does not always have a unique description, it can be interpreted as two different entities depending on the context. For instance, in Spanish, PCR can be referred to "*parada cardiorrespiratoria*" (cardiorespiratory arrest), "*Reacción en Cadena de la Polimerasa*" (Polymerase Chain Reaction) or "*Proteína C-Reactiva*" (Protein

C-Reactive). Finally, as we see in the examples, biological entities may also have multi-word names, so the problem is additionally complicated by the need to determine name boundaries and resolve overlap of candidate names.

- **Sub-domain of application**. The biomedical NER community has the opportunity to further advance the recognition of any type of named entity in different sub-domains such as the oncological, radiological, and pharmaceutical domains. Existing evaluation forums focus on specific applications that allow researchers to address specific tasks and progressively improve NER techniques. In particular, there are challenges related to drug and gene extraction such as PharmaCoNER [9] and CHEMDNER [10], and the identification of cancer problems [11]. In addition, in the pharmacological domain, there are workshops concerning the extraction of Adverse Drug Events (ADEs) [12, 13]. Relationship extraction was also born as a consequence of the NER task, since once the entity is recognized, it could be related to other entities [14, 15]. As we can see, there is a great concern to generate models that can identify entities according to a specific biomedical domain, which has been an extra motivation in the achievement of this thesis.

- **Basis for other NLP tasks**. Biomedical entity recognition serves as the basis for many other crucial areas of information management, such as classification tasks, question answering, information retrieval, and text summarization [16, 17, 18]. For instance, the use of NER becomes important for analyzing the clinical text and obtaining the most relevant tags

in each report, allowing the classification of documents. Regarding the question answering task, it is common practice to use NER systems to expand the query using synonyms of entities, descriptions, and acronyms in order to obtain better results. Another important challenge offered by the NER is to assign each entity a unique identifier in a database or controlled vocabulary. This process is known as entity normalization in which, once a biomedical entity has been identified, it can be shared in a standardized way with other systems.

- **Extracting structured information**. Biomedical NER is a task that facilitates medical professionals in structuring reports contributing to solutions such as providing a summary of patient conditions or serving as a tool to organize the documentation of the physician's decision-making process, plan development, and patient outcomes.

As we can see, there are many problems and difficulties that can be solved indirectly by advancing in solving the NER task in the health domain. Therefore, the main motivation of this thesis is to extract relevant knowledge from biomedical texts that can be useful and helpful for open problems in the medical area.

1.2 Objective

The main objective of this thesis focuses on the study, analysis, and development of NLP techniques and tools for the NER task in the biomedical domain in Spanish. Specifically, it focuses on the study and applicability of different

combinations of word embeddings as word representation.

This general objective has been defined through the following specific objectives:

- Collect resources available in Spanish annotated with biomedical entities used in different challenges.
- Study and select the existing word embeddings in Spanish serving as input to the network.
- Propose a deep learning-based method for NER in the biomedical domain that can take a combination of different word embeddings as input.
- Generate a new word embedding for Spanish focused on the biomedical domain to see how effective it is in comparison to existing ones.
- Evaluate the performance of the proposed method on the NER problem using three application scenarios: pharmacological domain, oncological domain, and knowledge discovery in biomedical texts.
- Conduct a results analysis comparing our system with the state-of-the-art.
- Perform an error analysis to understand the capabilities and drawbacks of our system.
- Identify open issues from the conclusions in order to propose future research.

1.3 Hypothesis

In this thesis, we address the problem of biomedical entity extraction in Spanish through deep learning and combinations of word embeddings using NLP methods. Based on the objectives set out above, our general hypothesis can be summarized as follows:

NLP techniques applied to the NER task can improve biomedical systems.

However, since this hypothesis is very ambitious, we have decided to subdivide it into three specific hypotheses:

Hypothesis 1 *(H1). Deep neural networks in NLP leverage the advantage of existing relevant information from the Spanish biomedical textual data and the NER task, outperforming models that do not integrate this information properly.*

Deep learning methods have emerged in recent years in the area of NLP. This success is due to the ability to tackle complex learning problems through multiple levels of representation and abstraction that help to make sense of texts in tasks such as NER.

Previously with the NER issue, Conditional Random Fields (CRFs) [19] were the primary modeling method for sequential labeling and extracting information from documents. However, deep learning has replaced CRF, causing the focus to shift from feature engineering to neural network design and implementation. The main weakness of CRFs is that they are unable to model the semantic similarity between two words. To overcome this problem, many

CRFs rely on dictionaries (a list of words related to the domain). However, this is a poor solution as dictionaries are intrinsically limited and can be very expensive to develop [20].

More recently, Recurrent Neural Networks (RNNs) are powerful deep learning models for application in NLP. These models usually use a vector representation for each token by reading token by token and "remembering" important information. In other words, they are loop networks that allow for the persistence of information and are capable of handling sequential data such as text sequences [21].

Considering the above, we believe that RNNs could be a potential model for addressing the NER task, in addition to other deep learning methods.

Hypothesis 2 *(H2). Combining different types of word embeddings by concatenating each embedding vector to form the final word vectors is an important part of the biomedical entity recognition task. The probability of recognizing a specific entity in a text should increase as optimal representations of that word are combined because they are more comprehensively represented and integrate relevant knowledge.*

Word embeddings are functions that allow us to map words to an n-dimensional vector, based on the assumption that words in a similar space must be related to each other. These models attempt to capture as much information from the context as possible in a word, and can even contain semantic and syntactic information [22].

Context-independent and context-based word embeddings are the most popular approaches at the moment. On the one hand, context-independent embeddings are static and word-level, which means that each distinct word

receives exactly one pre-computed embedding. Examples of this type are Bag-of-Words, Tf-Idf, Word2Vec, GloVe, and fastText [23, 24, 25]. On the other hand, context-based embeddings are based on the assumption that the same word in different contexts has different meanings. Some examples include transformer-based word embeddings such as Bidirectional Encoder Representations from Transformers (BERT) [26] and RoBERTa [27].

Considering both types of approaches, in this thesis we use different word embeddings, comparing them individually and proposing combinations. We believe that by combining different word representations through concatenation, the system may be able to understand a word more appropriately and then classify it as an entity or non-entity. The main advantage of using a combination of embeddings for each word is that this allows the combination of the knowledge of different embeddings in order to generate better quality word representations, i.e. it obtains the benefits of different word embeddings given their different nature, their training corpus, and their specific purpose. Finally, we consider that given the different nature of the selected word embeddings, in different scenarios the optimal combinations may be different from each other.

Hypothesis 3 *(H3). Integrating domain-specific knowledge into the training corpus can be beneficial for improving the quality of word embeddings. Thus, this resource provides a more accurate representation of words in a particular context and domain.*

Existing word embeddings have been trained on a huge corpus and generally work well, but sometimes fail on specific tasks such as health. The incorporation of generic word embeddings can lead to a difficult training

process where there is a discrepancy among domain words. For instance, in Spanish, the word "*órgano*" (organ) can have different meanings. Depending on the context of an application, it can refer to a collection of tissues or a harmonic instrument.

Therefore, a crucial issue in achieving a robust system is that it is usually dependent on the domain in which the data is represented. The training of word embeddings involves fitting a model to a pre-processed corpus and adjusting the hyperparameters of the model, which are settings whose values are empirically specified before training.

With this hypothesis we aim to address the following question: would it be beneficial to train domain and language-specific word embeddings?

1.4 Methodology

The methodology to be followed in order to achieve the above objectives is as follows:

1. **Study and review of state-of-the-art**. We will study the current state of the literature, compile important sources, ideas, and concepts related to NLP and the NER task. In addition, we will review the most commonly used Machine Learning (ML) methods in the NER task.

2. **Experimental design**. This step consists of creating a set of procedures to test the hypotheses. Good experimental design requires a solid understanding of the system under study. Specifically, experimental design is a way of carefully planning the experiments to be conducted in the NER

task to achieve valuable objectives and results.

3. **Implementation and experimentation**. After designing the experiments, we will proceed to the implementation and experimentation step. In this step, we will test the method in different application scenarios. In our particular case, we will use the proposed system in three health-related experimental frameworks: oncological, pharmacological, and knowledge discovery in biomedical texts.

4. **Analysis of results**. Subsequently, the systems developed will be evaluated and the results obtained will be compared with existing ones. The analysis of the results is another important part of the methodology used, as this step will show the success of our system. Finally, an error analysis is carried out to identify possible improvements to our systems.

1.5 Thesis outline

This thesis is organized into six chapters and an appendix. This first chapter contains an introduction explaining the motivation and objectives that led us to carry out the study. Furthermore, we have presented the hypotheses with the research questions we intend to solve and the methodology we will carry out. The remainder of this thesis is divided into different chapters and is organized as follows.

Chapter 2 presents an overview of the methodologies based on ML commonly used in the NER task and which are necessary to understand the later parts of this thesis. Specifically, this chapter analyzes the ML approaches

carried out to solve the NER problem in the biomedical domain. These techniques have been used throughout this thesis and have been divided into two categories: supervised and unsupervised methods.

Chapter 3 summarizes previous work on NLP tasks based on ML in the biomedical domain and shows an extensive literature review of the NER task with regard to present state-of-the-art studies. Since the interest of this thesis lies in word representation, this chapter details the review of existing methods for word representations up to the moment. Finally, it presents knowledge resources widely used by the Biomedical NLP (BioNLP) community.

Chapter 4 describes the proposed model to solve the biomedical entity extraction problem. After an extensive review of previously applied methodologies, we propose an approach based on a Bidirectional Long Short-Term Memory (BiLSTM) neural network with a final CRF layer. The input to the network will be composed of word embeddings to represent the words of a document consistently. Following this idea, our approach proposes the combination of different types of word embeddings by concatenating each embedding vector to form the final word vector. Finally, this chapter shows the new word embeddings created in this study for the specific domain and language.

Chapter 5 presents the experimentation carried out using the approach proposed previously. The experimental framework was developed in three scenarios belonging to different biomedical sub-domains including pharmacology, oncology, and knowledge discovery. For each scenario, this chapter

contains a description of the problem addressed, the dataset used, the methodology employed, and results obtained. Finally, error analysis and discussions are included for each one.

Chapter 6 contains our conclusion where we summarize our findings and main contributions. Moreover, this chapter provides an outlook into the future, the publications derived from the study, and the research results transferred.

Chapter 2

Machine learning approaches

This chapter provides basic knowledge in order to establish the basis for the following chapters. We will give a brief introduction to NLP to outline its general principles and then we will detail the most frequent approaches used in machine learning in the scope of artificial intelligence. According to McCarthy [28], we can define the Artificial Intelligence as follows:

> *"Artificial Intelligence (AI) is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence."*

Nowadays, AI has become more popular thanks to increased data volumes, advanced algorithms, and improvements in computer and storage power. Specifically, AI in healthcare has led to rapid advancement in digital medicine across multiple clinical specialties, including oncology [29, 30], radiology [31], and neurology [32]. Since a substantial amount of most relevant clinical information is embedded in unstructured data [33], NLP plays an essential role in extracting valuable information that can benefit decision-making, report structuring, report classification, and entity recognition, among others.

For more complex applications of NLP, the systems are based on ML models to improve their understanding of human language. According to Mitchell [34] a scientific field is best defined by the central question it studies. Therefore, the field of ML seeks to answer the question *How can we build computer systems that automatically improve with experience, and what are the fundamental rules that regulate all learning processes?*

Briefly, ML is defined by the Computer Scientist and ML pioneer Mitchell [35] as:

> *"The study of computer algorithms that allow computer programs to automatically improve through experience."*

Over the past decades, many automatic NER systems have been developed and used to identify and categorize biomedical entities using ML approaches. These approaches can be organized into different categories. For example, considering whether the algorithm is trained with labeled data or not, we can classify such algorithms into supervised learning and unsupervised learning.

In the course of this thesis, we have studied some of the approaches included in the previous categories and summarized in Figure 2.1[1]. On the one hand, within the unsupervised methods, we have investigated rule-based methods and dictionary-based methods. On the other hand, in supervised learning methods, we have made a distinction between traditional algorithms, neural networks, and transformer-based models.

[1]It should be noted that this classification has been proposed by ourselves.

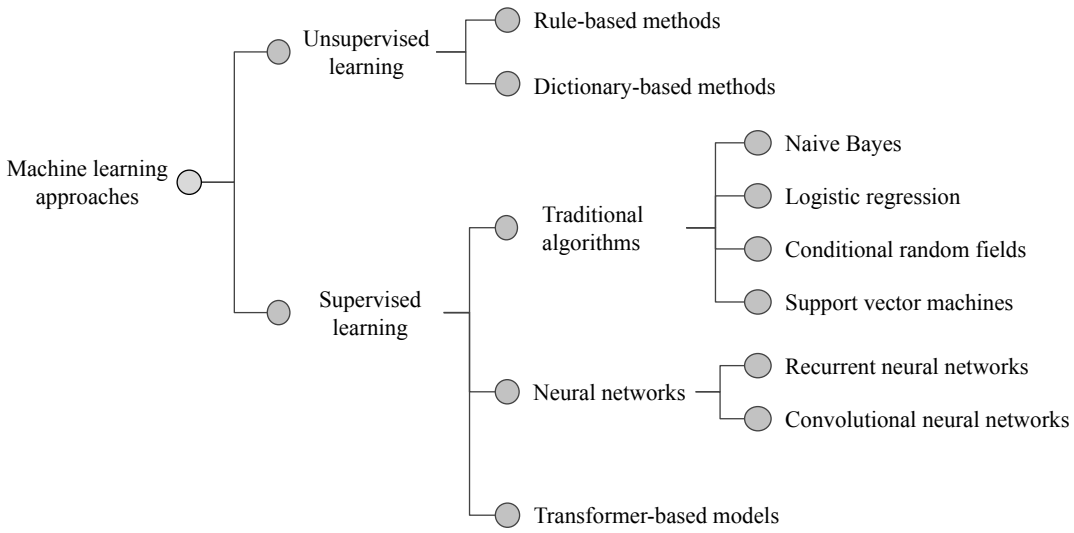

Figure 2.1: Classification of the machine learning methods studied throughout this thesis.

2.1 Unsupervised learning

In this section, we aim to survey two of the unsupervised learning techniques used for NER tasks in the biomedical domain. Unsupervised learning is a machine learning technique in which it is not necessary to supervise the model, i.e. it learns patterns from unlabeled data.

2.1.1 Rule-based methods

Rule-based systems are one of the most widely-used unsupervised methods in ML. These models are very appropriate in situations where the knowledge to be represented comes with a structure of rules.

The handcrafted models are hand-built systems that rely mainly on the intuition of human designers. They generally incorporate human knowledge

in the form of patterns or rules. Usually, patterns use grammatical (e.g. Parts-Of-Speech), syntactical (e.g. word precedence), and other language features to make a more accurate identification.

In rule-based systems, two types of rules are typically used:

- **Pattern-based rules**. These rules are based on checking a given sequence of tokens for the presence of the constituents of some pattern. To understand this type of approach, we present some examples of text features to be considered by the rule designer. In these examples, the aim would be to recognize chemicals and drugs in sentences.

  - May be acronyms such as "ADP" for the "acetylacetonate" componente and "DTT" for "dithiothreitol".
  - May be composed of a molecular formula including alphanumeric characters in capital letters such as "C6N4" referred to as "tetracyanoethylene" and "CD34" which is a transmembrane phosphoglycoprotein protein.
  - May contain a prefix such as "amino-", for example, "aminoglycoside" and "aminoacids".
  - May contain a suffix such as "-nitrile" included in "butanedinitrile" and "succinonitrile".

  The rules described above are not sufficient to identify all occurrences of entities in a document. Frequently, the rules themselves are incomplete and do not cover many examples.

Patterns in a sentence are often described using regular expressions (regexp) and matches. A regular expression consists exclusively of normal characters (such as "abc") or a combination of normal characters and meta-characters (such as "ab*c). Meta-characters describe certain constructions or character patterns, e.g, whether a character should be at the beginning of the line or whether a character should only appear exactly once [36].

- **Context-based rules**. Usually, the relevant information on the named entities is given in the contexts of their mentions. Analyzing a mention by humans or machines is a hard task without any contextual information where we can find the correct meaning of a word through a sequence. For example, if the term "Apple" occurs alone, it is not possible to identify what this term refers to. It could refer to the fruit, a person, a company, or a place. Resolving these ambiguities is often referred to as Named Entity Disambiguation (NED), which is a highly challenging aspect of an entity extraction task.

  Context-based rules establish a higher level of relationship between the tokens and the extracted features, e.g. windows size in a sentence and the use of conjunctions in order to connect words, phrases, clauses, or sentences.

Rule-based systems work very well when the language lexicon is not diverse. The advantages of these systems are that they are relatively easy to understand and the cause-effect relationship is transparent so that a domain expert can check the rule base and make adjustments if necessary. Due to

specific domain rules and incomplete dictionaries, high precision and low recall of such systems are often observed. The main disadvantages of these methods are several: *i*) the manual construction of the rules, which can be a time-consuming task depending on the domain, *ii*) since the rules are created for a very specific scenario, it is not possible to transfer this system to another different domain, and *iii*) they do not handle incomplete or incorrect information very well, i.e. data that does not have an associated rule will be ignored.

2.1.2 Dictionary-based methods

Dictionary-based approaches are other popular techniques in unsupervised ML. These techniques use linguistic resources such as dictionaries, glossaries, empty word lists, taxonomies, and thesauri to analyze the different levels of language: phonetic, lexical, semantic, or pragmatic.

This type of method is really useful in fields where the entities to be recognized may be contained in lists of words. However, these techniques are not always useful, for instance, if the entity to be identified is a first name and surname of a person, due to its nature and diversity, it will be difficult to find them in any resource, either in a dictionary or a list of words.

Dictionary-based approaches require exploration of variations in entity spelling in order to carry out the matching process [37]. Some examples of variations that can be taken into account by the automatic pattern matching process are:

- Special characters such as hyphen, slash, and brackets are used as separators in different combinations by different authors. For instance, the entity "Ki-67" and "Ki67" refers to the same protein but contains a dash between characters, which makes them difficult to match correctly.

- Parts of the names can be spelled in the upper case by some authors and lower case by others. Each author can write their texts without following a formatting guide so it is possible to find mixed capital and small letters such as "Beta-HCG" and "beta-hcg".

- There is a wide variety of acronyms in the human language, we may have problems with matching descriptions and acronyms such as "Ag" (antigen).

Often, the dictionaries used in NLP could contain a large amount of useful information to address the problems mentioned above. On the one hand, it is possible to find them with a simple list of words which our system must match directly, but on the other hand, they could contain synonyms, antonyms, descriptions, acronyms, among others, whereby the system needs to perform a less comprehensive search.

Since a dictionary may have many meanings of an entity, the system has to determine which meaning is used in the context of a document. In this process, note that it is important to disambiguate each word in the right context in order to carry out a correct matching. For instance, if we take into account the word "*cólera*" (cholera) (in Spanish could be related to illness or a bad mood), the entity can be included in a disease dictionary, but in a sentence like "*la chica*

montó en cólera al no encontrar sus llaves" (the girl went into a rage when she could not find her keys) the word "*cólera*" should not be labeled as a disease since it refers to a mood.

2.2 Supervised learning

The main goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictive features using a labeled training dataset. The resulting algorithm is used to assign class labels to test instances.

For a description of these methods, we have performed a division including traditional algorithms, neural networks, and transformer-based models.

2.2.1 Traditional algorithms

In the context of ML, traditional algorithms mean the things we have been doing for years and are often the basis for more advanced ML. In the following section, we review four algorithms that are considered traditional machine learning methods and are widely used by researchers interested in the NER task. These algorithms include Naive Bayes, logistic regression, CRF, and support vector machines[2].

Naive Bayes

Naive Bayes (NB) is a simple probabilistic classifier that makes the naive assumption that all the feature variables are independent. In broad terms, an

[2]Although there are more algorithms included in traditional machine learning, in this section, we wanted to highlight those commonly used in NER and specifically those used in the initial studies of this thesis.

NB classifier assumes that the value of a particular feature is independent of the value of any other feature, given the class variable [38].

The NB technique decides the possibility based on a table of feature vectors. An entity extraction problem can be written as the problem of finding the class with maximum probability given a set of observed feature values [39]. This probability is the posterior probability of the class given the data and is computed using the Bayes theorem, as:

$$P(y, X) = \frac{P(X|y)P(y)}{P(X)} \tag{2.1}$$

Where y is a variable class, X is represented by a vector of features such as $X = \{x_1, x_2, ..., x_n\}$, $P(y)$ is the prior probability and $P(X|y)$ the likelihood function.

As described above, when we compute $P(y, X)$ the feature variables are assumed to be independent which implies that the joint probability can be written as a product of probabilities:

$$P(y, X) = P(y) \prod_{m=1}^{M} P(X_m|y) \tag{2.2}$$

In this equation, the category label of X is predicted as the class y which has the highest $P(y, X)$.

There are different types of NB classifiers such as multinomial Naive Bayes which is mostly used for document classification problems, similar to the multinomial we can find Bernoulli Naive Bayes with boolean variables as predictors, and Gaussian Naive Bayes when the predictors take up a continuous

value and are not discrete.

Logistic regression

Logistic Regression (LR) is another technique that ML has adopted from the statistical field [40]. Specifically, the algorithm LR is a discriminative model that describes the conditional probability as:

$$P(y|X) = \frac{exp(\sum_{m=1}^{M} \lambda_m f_m(y, X))}{\sum_{y'} exp(\sum_{m=1}^{M} \lambda_m f_m(y', X))} \tag{2.3}$$

Logistic regression is a linear method, but the predictions are transformed using the logistic function. This function is also called sigmoid which describes the weight $\lambda_m f_m$ of features f_m defined with respect to y and X in order to generate a class prediction. Moreover, the features are defined for state-observation pairs $f_m(y, X)$ [41].

Conditional Random Fields

The Conditional Random Fields (CRFs) are an important type of ML models motivated by the principle of the Maximum Entropy Markov Model (MEMMs) [42] and used for sequence labeling.

Lafferty, McCallum, and Pereira [19] proposed CRF as probabilistic models to segment and tag sequence data in order to inherit the advantages of previous models, overcome their shortcomings and increase their efficiency.

According to the authors, there are two main differences between CRF and MEMMs are two: MEMMs use exponential state models for the conditional

probabilities of the next states given the current state, and the CRF algorithm has a single exponential model for the joint probability of the entire label sequence given the observation sequence. Thus, the weights of the different features in different states can be exchanged with each other.

The basic principle of CRF is to define the conditional probability distribution over the label sequences in a given observation [20]. More specifically, the conditional probability of a sequence of labels y given a sequence of word X is shown in Equation 2.4.

$$P(y|X) = \frac{exp(\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_j f_j(y_{i-1}, y_i, X))}{\sum_{y} exp(\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_j f_j(y_{i-1}, y_i, X))} \tag{2.4}$$

Where the denominator of the equation is a normalization factor of all state sequences. $f_j(y_{i-1}, y_i, X)$ is one m function which describes a specific feature and λ_j is a learned weight for each feature function.

Using CRF models the sequences could be represented by linguistic features. Typical features for CRFs can be generalized such as the previous word, current word, next word and Part-Of-Speech tag to provide context to the model. Furthermore, other features respond to the syntax of the word such as is it small, is it capitalized, is it a number, among others.

Support Vector Machines

Support Vector Machines (SVMs) are supervised learning models with associated learning algorithms for data classification. SVMs were developed in the 1990s, within the field of computer science. Although they were originally

developed as a binary classification method, their application has extended to multiple classifications and regression problems. SVMs have proven to be one of the best classifiers for a wide range of situations and are therefore considered one of the benchmarks in the field of statistics and machine learning [43].

In a binary classification task, when we have n observations, each with p predictors and whose response variable has two levels, we can use hyperplanes to build a classifier that allows us to predict which group an observation belongs to according to its predictors. This same problem can also be addressed with other methods such as LR, Latent Dirichlet Allocation (LDA), and classification trees.

In a p-dimensional space, a hyperplane is defined as a subspace plane and related to $p - 1$ dimensions. For instance, in a two-dimensional space, the hyperplane is a 1-dimensional subspace, i.e. a straight line. In three-dimensional space, a hyperplane is a two-dimensional subspace, a conventional plane. For $p > 3$ dimensions it is not intuitive to visualize a hyperplane, but the concept of subspace with p-1 dimensions is maintained.

SVMs achieve good results when the boundary between classes is approximately linear, however, their capacity declines if they are not linear. One strategy for addressing this type of scenario is to expand the dimensions of the original space. The fact that groups are not linearly separable in the original space does not mean that they are not separable in a larger space. This scenario is shown in Figure 2.2, where we can see two groups, whose separation in two dimensions is not linear (left of the figure), but are made independent

variables by adding a third dimension (right of the figure).

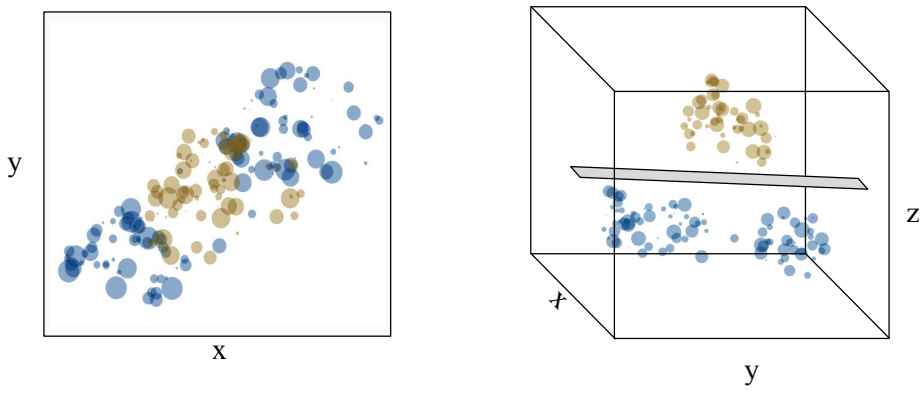

Figure 2.2: Hyperplane extension in SVM

The SVM model allows the expansion of space through kernels [44]. Although we will not go into detail, there are many different kernels, some of the most used being: linear, polynomial, Gaussian Radial Basis Function (RBF), and hyperbolic tangent or sigmoid.

2.2.2 Neural networks

In recent years, deep neural networks have revolutionized many application domains of ML. Deep neural networks are part of a broader family of machine learning methods based on Artificial Neural Networks (ANNs). An ANN employs a hierarchy of layers in which each layer considers information from a previous layer and then passes its output to other layers [21]. While traditional ML algorithms are usually linear, deep learning algorithms are stacked in a hierarchy of increasing complexity and abstraction.

ANNs, usually named neural networks, are inspired by the biological neural networks that constitute brains. An ANN is based on a set of connected

units or nodes called artificial neurons or simply neurons, which attempt to model the neurons in a biological brain. Each neuron has inputs and produces a single output that can be sent to multiple other neurons. The inputs can be the characteristic values of a sample of external data, such as images or documents. Moreover, these characteristics may be the outputs of other neurons. The final outputs of the neural network accomplish the task, for instance, the identification of an entity in the text. Usually, the neurons are organized into multiple layers as we can see in Figure 2.3. In this figure, the neural network structure consists of input, hidden, and output layers. Note that neurons in one layer connect only to neurons in the immediately preceding and following layers.

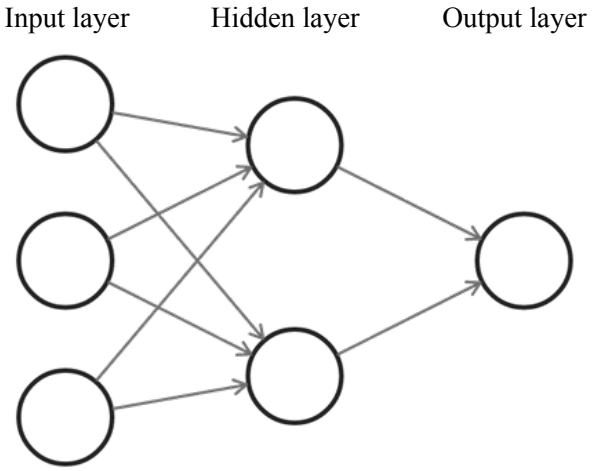

Figure 2.3: Basic ANN architecture.

The Feedforward Neural Network (FNN) was the first and simplest type of ANN [45]. In this particular network, information moves in only one direction: forward from the input nodes, through the hidden nodes (if any), and to the output nodes, so there are no cycles or loops in the network. FNNs can be

divided into two groups: Single-Layer Perceptron (SLP) and Multi-Layer Perceptron (MLP). On the one hand, SLP consists of a single layer of output nodes, whereby the inputs are fed directly to the outputs through a series of weights. On the other hand, MLP consists of multiple layers of computational units, usually interconnected in a feed-forward way. Each neuron in one layer has direct connections to the neurons of the subsequent layer [46].

The learning process of a neural network takes the desired inputs and outputs and updates the internal state accordingly so that the calculated output is as close as possible to the desired output [47]. The prediction process takes an input and subsequently generates (using the internal state) the most probable result according to its past experience. To achieve this, we will briefly discuss the learning process in several steps:

1. **Initialization**. The initialization of the model refers to the first hypothesis that the process intends to start. As with genetic algorithms and the theory of evolution, neural networks can start from anywhere. Therefore, a random initialization of the model is a common practice.

2. **Forward propagate**. This step is concerned with propagating the computations of all the neurons within all the layers that move from left to right. This begins in the input layer and ends with the final prediction. The forward calculations occur during training to evaluate the target and loss function under the current network parameter settings in each iteration, as well as during prediction when applied to new, previously unseen data.

3. **Loss function**. At this step, on the one hand, we have the actual output of our randomly initialized neural network. On the other hand, we have the desired output we would like the network to learn. The loss function is a performance metric on how well the neural network achieves its objective of generating results as close to the desired values as possible. For this reason, ML algorithms aim to minimize the loss function. In classification models, the most common loss functions used are binary cross-entropy, categorical cross-entropy, and cosine similarity, among others. For regression-based problems, we also have functions like Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE).

4. **Differentiation**. In mathematics, differentiation is the step that can help the network to optimize the weights. It is the partial derivative of the loss function. Therefore, by finding the partial derivative of the loss function, we know how much (and in which direction) we must adjust our weights and biases to decrease the loss.

5. **Back-propagation**. In the neural network, any layer can forward its results to many other layers, in this case, to perform back-propagation. Essentially, this step evaluates the expression for the derivative of the loss function as a product of derivatives between each layer from left to right using the gradient of the weights.

6. **Weight update**. The updates of the weights of the neurons will reflect the importance of the error propagated backward after a forward pass has been completed. The methods of updating weights are called optimizers.

7. **Iterate until convergence**. Since we update the weights with a small step, the network will need several iterations to learn the most optimal output.

In an ANN, the activation functions determine the output of a deep learning model, its accuracy, and also the computational efficiency of training a model. Softmax and sigmoid functions are common functions used in the final or output layer of a neural network to obtain a categorical and Bernoulli distribution respectively (function for binary classification) [48]. For hidden layers, the most popular activation functions are Rectified Linear Unit (ReLU) and a hyperbolic tangent or tanh function [49]. Additionally, some functions applied to the neural network have different configuration parameters. For instance, the learning rate is an adjustment parameter in the optimization function that determines the size of the step in each iteration while moving towards a minimum loss function.

As mentioned above, neural networks often use optimizers to reduce losses. These are algorithms or methods used to change the attributes of a particular neural network, such as weights and learning rate. Some examples of the most commonly used optimizers are Adam, Stochastic Gradient Descent (SGD), Adadelta, RM-Sprop, Adamax, and Adagrad [50, 51, 52].

ANNs have limitations in remembering sequences when they are large. For example, suppose a 90-word sentence in which the penultimate word refers to the beginning of the sentence. ANNs tend to forget information on time steps that are far behind schedule. To address this problem, the attention mechanism emerged to deal with time-varying data (sequences). Attention is

considered one of the most influential ideas in the deep learning community because it maintains relevant information over time. With this mechanism, each word of the sentence contains a hidden state with past values that will be taken into account in each iteration [53].

In conclusion, deep learning represents a set of techniques based on ANN. The two main architectures designed for classification are the RNNs and Convolutional Neural Networks (CNNs) [45]. These two models differ in the kind of input that they predict: RNNs are designed to classify temporal signals and CNNs are designed to classify spatial signals. The following sections describe the particularities of these types of ANN.

Recurrent Neural Networks

RNNs were first studied in 1986 [54] and they are a class of ANNs in which the connections between nodes form a directed graph along a time sequence. RNNs are commonly used for ordinal or temporal problems, such as NLP, speech recognition, and image subtitling; they are also incorporated into popular applications such as "Apple's Siri" and "Google Voice Search".

RNNs use a memory cell that takes information from previous input to influence the current input and output. While traditional deep neural networks assume that inputs and outputs are independent of each other (no feedback), the output of RNNs depends on the above elements within the sequence. Figure 2.4 illustrates how RNN (on the left of the figure) has a recurrent connection on the hidden state and FNN contains no feedback (on the right of the figure). This loop restriction ensures that sequential

information is captured in the input data [21].

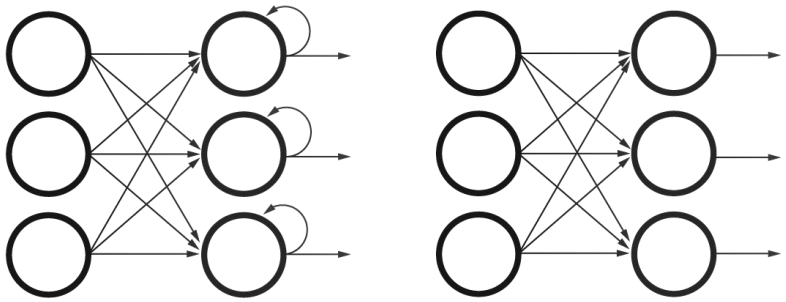

Figure 2.4: Comparison of RNNs (left) and FNNs (right).

In order to deepen the knowledge about RNN, we describe two variants of RNN architectures that we have used in the course of this thesis:

- **Long Short-Term Memory** (LSTM) networks are a type of RNN capable of learning order dependence in sequence prediction problems. LSTM was introduced by Hochreiter and Schmidhuber [55] in 1997 as a solution to the vanishing gradient problem.

  A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates (input, output, and forget) regulate the flow of information into and out of the cell. Therefore, the LSTM can remove or add information to the cell status. To better understand this performance Figure 2.5 shows an LTSM network cell. As we can see, the input gate (i_t), the forget gate (f_t) and the output gate (o_t) in the current step (t), transform the input vector x_t taking the previous output h_{t-1} using its corresponding weight and bias computed with a sigmoid

function. Moreover, the cell state c_t takes the information given from the previous cell state c_{t-1}. Finally, the current output h_t is defined by the function of the cell state and regulated by the output gate.

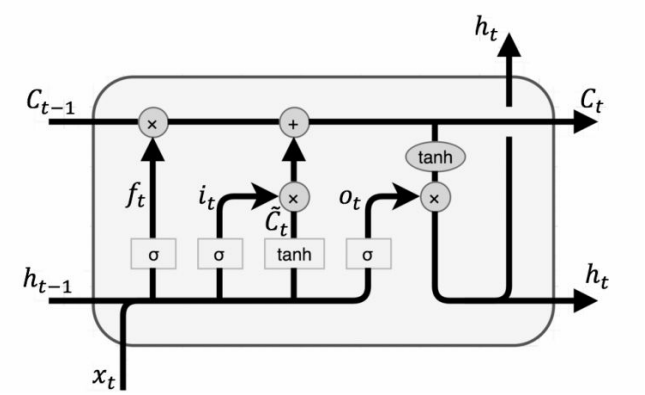

Figure 2.5: Long Short-Term Memory cell.

- **Gated Recurrent Units** (GRU) are a gating mechanism in recurrent networks, introduced in 2014 by Cho et al. [56]. GRU is considered similar to the LSTMs as it also works to address the short-term memory problem of RNN models.

  The main differences between LSTM and GRU networks are two: *i*) GRU instead of using a cell state to regulate information uses hidden states, and *ii*) GRU instead of three gates, has only two, a restart gate and an update gate. Similar to the gates within LSTMs, the restart and update gates control how much and what information is retained.

  As we can observe in Figure 2.6, this type of RNN defines the input gate z_t and the reset gate r_t which take the input vector x_t of the current time (t) and the previous output h_{t-1} using its corresponding weight and bias computed with a sigmoid function. The current cell information is

captured by using the input and the previous output regulated by the reset gate with the function. Then, the output vector h_t is regulated by the reset gate by choosing the most significant information between the current and the previous cells.

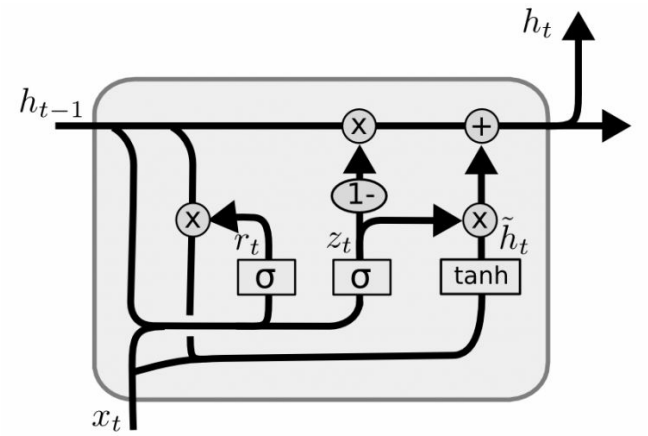

Figure 2.6: Gated Recurrent Units cell.

Convolutional Neural Networks

In cases where the inputs are large, RNNs involve a large number of training parameters. The main idea for overcoming this problem is to take the local representation that describes the entire input rather than taking the global representation. CNN uses layers with convolution filters that are applied to local features in order to represent local information [57]. In this type of neural network, the connections between nodes do not form a loop but use a variation of MLP designed to require minimal preprocessing.

Originally invented for computer vision, CNN models have subsequently been shown to be effective for NLP and have achieved excellent results in text classification [58]. CNNs can also be applied to NLP tasks using textual data

because the inputs are the vector representation of each word in a sentence.

Three types of layers make up the CNN: convolutional layers, pooling layers, and fully-connected layers. As we show in Figure 2.7, the convolutional is the first layer that is used to extract the various features from the input. In this layer, the mathematical operation of convolution is performed between the input and a filter of a particular size MxM. In most cases, the convolutional layer is followed by the pooled layer. The main objective of this layer is to decrease the size of the map of convolutional features to reduce computing costs. Finally, the fully connected layer consists of the weights and biases along with the neurons and is used to connect the neurons between two different layers. These layers are normally placed before the output layer and form the last layers of the CNN architecture.

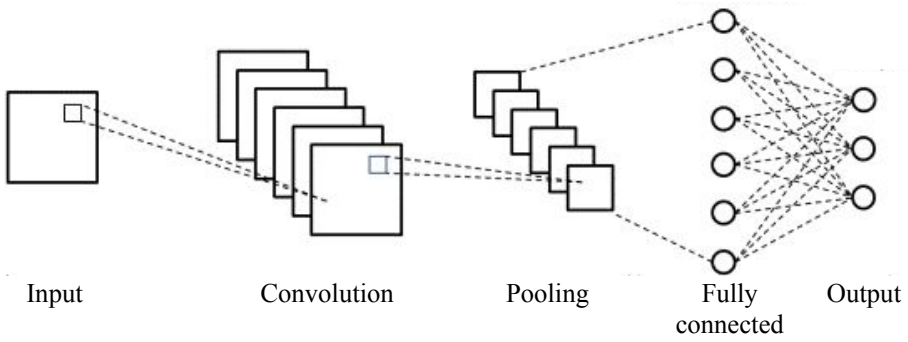

Figure 2.7: Basic CNN architecture.

2.2.3 Transformer-based models

The Transformer is a deep learning model introduced in 2017, used primarily in the field of NLP [59]. Similar to RNNs, Transformers are designed to handle sequential data such as natural language for tasks like NER and text

classification. Contrary to RNNs, Transformers do not require sequential data to be processed in order, therefore, Transformers do not need to process the beginning of a sentence before processing the end. Due to this aspect, this technique allows much more parallelization than RNNs and therefore reduces training times.

The most important part of the Transformer is the attention mechanism. The attention mechanism represents the importance that other tokens of an entry have for the coding of a given token. In other words, the attention mechanism allows the Transformer to focus on certain words on both the left and the right to treat the current word according to the NLP task we are addressing.

A further advantage of the Transformer architecture is that learning in one language can be transferred to other languages via transfer learning. In broad terms, transfer learning is the idea of taking the knowledge acquired when performing a task and applying it to a different task. Transformers rely on this technique in order to achieve state-of-the-art results [60].

There is a major difference between the traditional approach of building and training ML models, and using a methodology that follows the principles of transfer learning. Figure 2.8 illustrates the difference between traditional ML and the new idea based on transfer learning. In this figure, we can see that traditional ML (left of the figure) is isolated and occurs strictly task-specifically (task 1 and task 2) and dataset (dataset 1 and dataset 2). Therefore, the training models are independent of each other and do not retain any knowledge that can be transferred from one model to another. On the contrary, by using

transfer learning models (right of the figure) it is possible to take advantage of the knowledge (characteristics, weights, etc.) of the previously trained models to train new systems.

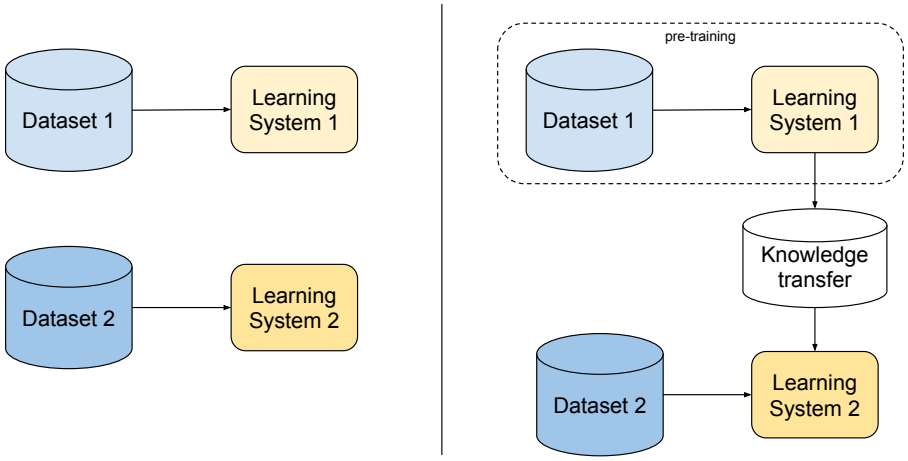

Figure 2.8: Traditional machine learning vs transfer learning models.

As we can observe in Figure 2.8, Transformer models are first trained in huge amounts of text (dataset 1) in a step called pre-training. During this step, the models are expected to learn the words, structure, morphology, grammar, and other linguistic characteristics of the language. In this step, the text is represented by tokens through a tokenization process converting the actual text into a numerical representation that can be used with neural network models.

Once the text is converted to an ANN-compatible format, we can train the model to understand the language. Masked Language Modelling (MLM) is a technique used to perform this task [26]. With this technique, a certain percentage of the tokens in a sequence is replaced with a mask token, and the model is asked to predict the token that was previously there. When trained in

a task using this technique, a model is able to learn good representations of the various vocabulary tokens. Other techniques for learning text representation are Causal Language Modeling (CLM), which predicts the probability of a word given the previous words in a sentence, and Translation Language Modeling (TLM) designed for cross-lingual data [61].

The result of this prior training process is a model that is capable of modeling a language accurately, understanding the different characteristics and linguistic rules of the language [62].

Transfer learning is popular in deep learning, given the enormous resources needed to train deep learning models and the large and challenging datasets on which deep learning models are trained. Fortunately, many pre-trained models are already available for reuse and serve as a starting point for different tasks. Hugging Face[3] is the most popular Python library that contains these models. Among the most well-known pre-trained models, we can find T5, GPT-3, GPT-2, BERT, XLNet, and RoBERTa, which demonstrate the ability of Transformers to perform a wide variety of such NLP-related tasks and have the potential to find real-world applications [63, 64, 65, 26, 61, 27].

In essence, Transformers changed by NLP offering the following benefits:

- It introduced a revolutionary attention mechanism that replaces convolutional or recurrent architectures.
- It produced a shift in transfer learning from pre-training (word vectors)

[3]Hugging Face: `https://huggingface.co/transformers/`

for feature extraction to the training of generic language models (pre-trained models).

- It provided fine-tuning where the model may need to be adapted for the task of interest.
- It resulted in an exponential growth in the size of pre-trained language models, which led to high performance in a series of NLP tasks involving understanding the language.
- It provided pre-trained models for tasks where we lack large datasets to train on.

Chapter 3

Related work

This chapter summarizes previous work covering NLP tasks in the biomedical domain. Specifically, current attempts to address the NER task will be presented taking into account the architectures and methodologies followed by the authors. For a comprehensive review, we have divided this chapter into several sections including important aspects of current literature such as the biomedical domain, approaches applied to the NER task, word representation, and biomedical knowledge resources.

3.1 Biomedical domain

BioNLP refers to the methods and study of how text mining may be applied to texts and literature of the biomedical field and other more specific subdomains such as radiology, oncology, and pharmacology [66]. Moreover, BioNLP is frequently used by health services since it has benefits such as reducing uncertainty, supporting evidence-based decision-making, and offering interoperability with health systems. All these potential benefits are briefly

described below in order to show some methods and possible applications in the biomedical field.

Many studies involved with developing BioNLP approaches have been dedicated to uncertainty detection. For instance, in Information Retrieval (IR) tasks, uncertainty detection improves the results of information extraction from radiology reports [67]. Following this idea, Vincze et al. [68] created the BioScope corpus, which is an open-access resource for research on uncertainty management in biomedical texts. The corpus consists of three parts, namely medical free texts, biological full papers, and biological scientific abstracts. Due to their prevalence and high level of biomedical uncertainty, breast cancer [69] or pneumonia [70] are important cases for analyzing the impact of biomedicine on illness identity.

Regarding support for evidence-based decision making, it refers to a health information technology system designed to provide physicians and other healthcare professionals with Clinical Decision Support (CDS), i.e. assistance in clinical decision-making tasks. ML and NLP researchers can play a key role in making evidence more actionable, for example, by making it easier to seek out and extract reported findings [71]. Peiffer-Smadja et al. [72] focused on the evaluation of the use of decision support systems concerning various ML techniques, the evaluation of the results, and the implications of these decision support systems at the real-time clinical level for the diagnosis of cardiac problems.

Currently, the new COronaVIrus Disease 2019 (COVID-19) is creating an important and urgent threat to global health. In this way, many efforts are

being focused on developing automated solutions to support medical experts in the early detection of the disease based on medical images and text [73, 74]. Prediction models that combine variables or features to estimate the risk of people becoming infected are helping clinicians to deal with the COVID-19 outbreak [75]. As mentioned above, these models require innovative approaches that provide immediate and real-time results. For instance, López-Úbeda et al. [16] developed an ML model that automatically extracted radiological findings consistent with COVID-19 in chest Computed Tomography (CT) reports. This system is currently used in real-life scenarios by radiologists as a decision support tool to detect suspected cases of COVID-19. In this study, they also extract information using an unsupervised automatic system and add extra information to the system. Specifically, they detect virus-related disorders such as bilateral pneumonia and ground-glass opacities.

The last benefit mentioned is interoperability. In this case, BioNLP models can contribute to solving the problems of semantic interoperability and knowledge reuse in clinical information systems. Following the definition of Miranda et al. [76]:

> *"Interoperability is the ability of independent systems to exchange meaningful information and initiate actions from each other, in order to operate together to mutual benefit."*

Interoperability is currently a major issue within the scientific community because electronic health information systems used in healthcare organizations have developed independently with tools, methods, processes, and procedures that result in a large number of unique and proprietary models

that represent and record patient information [77]. In order to ensure semantic interoperability between healthcare systems, it is necessary to use standards that allow the exchange of data, as well as the use of normalized and curated vocabularies, which unify the data used in different institutions resulting in the correct exchange of information. Some of the standardized vocabularies are detailed in the next Section 3.4.1. In this context, the NER task is one of the most widely used because it allows the extraction of knowledge that can be shared in a standard and understandable way [78].

This thesis focuses on the extraction of named entities using biomedical texts as a source of information. NER adapts to any situation where a high-level overview of a large amount of text is useful. Moreover, the NER task can be applied to a variety of healthcare systems to perform automatic knowledge extraction. For example, with the entities identified in a text, a user could understand the topic of the text and quickly group texts according to their relevance or similarity; they could also improve the speed and relevance of the IR system through text summaries and using meaningful entities; and finally, NER could be standardized through ontologies and controlled vocabularies to manage and exchange information [79].

3.2 Biomedical Entity Recognition

Sophisticated information processing methods are required for the efficient acquisition and integration of data from a corpus of biomedical literature. Effective identification of terms is key to accessing stored information since it is the terms that represent the knowledge in the texts. Due to the complexity

of dynamically changing biomedical terminology, term identification has been recognized as a challenge within text mining, and as a consequence, has become an important research topic in both the NLP and biomedical communities.

Since there is currently a great growth in demand for understanding and extracting information from medical texts, the NLP community has organized a series of open challenges focused on biomedical entity extraction. These challenges usually have several advantages: they provide a corpus available in different languages; they propose a baseline system of experimentation; they provide the participants with an evaluation method, and they offer the state-of-the-art in a specific task and dataset. Some of the most popular NER task-focused challenges are briefly described below.

On the one hand and focused on English, **DDIExtraction** [15] was presented at the SemEval 2013. The task concerned the recognition of drugs and the extraction of Drug-Drug Interactions (DDI) included in the biomedical literature. This challenge was divided into two subtasks: the recognition and classification of pharmacological substances and the extraction of DDI where participants could submit their systems. The **N2C2 - National NLP Clinical Challenges shared task** [12] was focused on the extraction of ADEs from clinical records and three subtasks were evaluated: concept extraction, relation classification, and end-to-end systems. Other workshops have also been proposed in the past to address the ADEs task in texts other than biomedicine, more specifically using tweets [13]. In 2015, the **CHEMDNER** challenge [10]

was organized by BioCreative and promoted the development of novel, competitive and accessible chemical text mining systems.

On the other hand, the use of Spanish as the main language of a challenge has emerged in recent years providing important workshops. In 2018 at IberEval, the **DIANN** [80] (Disability Annotation Task) task consisted of detecting disabilities in English and Spanish texts, independently of each other. The **PharmaCoNER** [9] (Pharmacological Substances, Compounds and proteins Named Entity Recognition) track was also proposed as an entity recognition task in the pharmacological domain. The main objective was to find mentions of chemicals and drugs in clinical cases. The challenge was composed of two sub-tasks: *i*) NER offset and entity classification, and *ii*) concept indexing using SNOMED-CT as vocabulary. **Cantemist** [11] (CANcer TExt MIning Shared Task) was the first task focused on entity recognition in the oncology domain. Participants in this task could submit systems to the three sub-tasks proposed by the organizers named cantemist-NER, cantemist-NORM, and cantemist-CODING. Focused on the recognition of diagnoses and procedures, the **CodiEsp** [81] track was born in CLEF eHealth 2020. Other challenges related to the biomedical domain, such as **eHeatlh-KD** [14] (eHealth knowledge discovery), instead of using entities specific to the medical field use general-purpose entities.

Researchers interested in entity extraction tasks have explored a variety of ML approaches. As we described in Chapter 2, the ML approaches formulate the clinical NER task as a sequence labeling problem that aims to find the best labeling sequence from clinical text. Many previous studies applied

the CRF method [19] in order to perform the identification and subsequent classification of entities. CRF is the most popular solution among conventional ML algorithms. A typical CRF model usually uses features from different linguistic levels, including ontologies, lexicons, syntactic information, or word embeddings [82]. SVM is another useful algorithm used in traditional ML to identify biomedical entities [83, 84]. For instance, Takeuchi and Collier [85] focused on identifying entities from the domain of molecular biology. They used a text collection of MEDLINE[1] abstracts in order to perform the experiment. In addition, they add to the system a set of word-level linguistic features including word surface forms, Part-Of-Speech tags, and orthographic features. The study showed that the combination of some features achieves high results (about 74% F1-score) in this specific domain.

Early studies of the NER task primarily aimed at RNNs to produce promising results. These studies have demonstrated the great effectiveness of RNN applied to biomedical entity extraction using complex network architectures [86, 82, 87]. Compared with traditional ML methods, RNNs usually use an embedding layer as input in order to learn the vector representation of words [24, 88, 89, 22].

In 2015, Huang, Xu, and Yu [90] proposed a variety of LSTM models for sequence labeling, including LSTM networks, BiLSTM networks, LSTM with a CRF layer (LSTM-CRF), and BiLSTM with a CRF layer (BiLSTM-CRF). Their research found that the BiLSTM-CRF model made effective use of both past and future input features. The models presented produce state-of-the-art

[1]MEDLINE: `https://www.nlm.nih.gov/medline/index.html`

accuracy on Part-Of-Speech labeling, chunking, and NER datasets. More recently, Hong and Lee [91] introduced DTranNER, a novel CRF-based framework incorporating a deep learning-based label-label transition model into biomedical NER tasks. They performed experiments on five benchmark corpora to compare the state-of-the-art methods in each of them. The DTranNER model achieves the best F1-score on four datasets, including BC2GM [92], BC4CHEMD [10], BC5CDR [93] on chemical and disease datasets surpassing the popular BioBERT model [63] based on Transformers. However, BioBERT outperforms DTranNER on the NCBI-Disease [93] corpus.

Although RNNs have obtained high results and a wide range of related literature on the NER task in recent years, the pre-training of Transformer-based language models such as BERT [26] has also led to impressive gains in NER systems [94]. Some pre-trained models based on BERT are even specific to the biomedical domain such as BioBERT [63], which is pre-trained on large-scale biomedical corpora and ClinicalBERT, which specializes in clinical texts and showing improvement in some NLP tasks in the clinical domain [95]. SciBERT is a pre-trained language model on scientific text that has demonstrated similar results in the biomedical field but improves in the computer science domain compared to BioBERT [64].

Given the increasing number of available pre-trained models, the related literature, like the state-of-the-art results, is constantly changing. Thus, all of the described models above are frequently compared by fine-tuning them to different domains and corpora [96, 64, 97].

Concerning Spanish, there are pre-training models for this language. BETO

[65] is a BERT model trained on a large Spanish corpus and SpanBERT [98] is a pre-training method that is designed to better represent and predict spans of text. Contrary to English, this language does not yet have a specific model for the biomedical domain. However, with the growing attention being paid to these language models in languages other than English, multilingual models such as mBERT [26], XLM-RoBERTa [27] and XLM [61] have been generated and can be used for the development of systems in different languages.

3.3 Word representations

In the field of NLP, researchers needed to find a way to represent textual data as input into ML systems. This process consists of transforming a set of categorical features in the raw text (words, letters, Part-Of-Speech tags, word position, word order, among others) into a series of vectors.

First, beyond converting words into a numerical representation, we ask the following questions: what are we interested in knowing about the text when performing this data encoding? and more specifically, what exactly do we want to encode? In this section, we discuss the options most commonly used by the NLP community to represent words that ML systems can understand.

The first approach emerged as a simple method in which each word in the vocabulary was given a unique identifier. A vocabulary in a corpus or text consists of the unique words included in it. **Dictionary lookup** methods are a simple way of representing text by checking whether an input string appears in a dictionary. Otherwise, if the word does not appear, the string is marked as a misspelled or Out-Of-Vocabulary (OOV) word [99].

A dictionary-based approach stores as many named entities as possible in a list called a gazetteer, which offers high accuracy in correctly identifying these entities [100]. Currently, such methods are outdated for simple word representation because they have limitations but they are still used as additional features for word representation in neural networks [101, 102, 82]. The main shortcomings of these techniques can be summarized as follows: *i*) a short dictionary may not be sufficient to find the word in context, *ii*) a dictionary may not contain all the words we want to represent, and *iii*) a large dictionary greatly increases the cost of the search.

The second approach to carrying out word representation we studied is named **one-hot encoding**. This technique uses a representation of categorical variables as binary vectors. The main idea is to create a vocabulary size vector filled with all zeros except one position. Then, for a word, only the corresponding column is filled with the value 1 and the rest has value zero. Moreover, this method uses a vector position to indicate that the word is OOV [103]. In order to better understand this type of representation, Figure 3.1 shows an example where the sentence "*anomalías del sistema nervioso*" (nervous system abnormalities) is encoded through zeros and ones. As we can see, the encoded words consist of a vector of dimension $N + 1$, where N is the size of the vocabulary and the extra 1 is added for OOV words.

One-hot vectors are frequently used as word representations for NER tasks. For instance, Kuru, Can, and Yuret [104] developed a neural network-based model that took as input a one-hot vector representation to encode the input characters. Additionally, one-hot vectors can be used as input features for ML

	anomalías	del	sistema	nervioso	central	...	OOV
anomalías	[1,	0,	0,	0,	0,	0, 0,	…, 0]
del	[0,	1,	0,	0,	0,	0, 0,	…, 0]
sistema	[0,	0,	1,	0,	0,	0, 0,	…, 0]
nervioso	[0,	0,	0,	1,	0,	0, 0,	…, 0]
central	[0,	0,	0,	0,	1,	0, 0,	…, 0]

Figure 3.1: Example of one-hot encoding. English translation: nervous system abnormalities.

methods [16] by entering extra information into the algorithms. Following this idea, Li et al. [105] concatenated the one-hot Part-Of-Speech encoding with other word representations.

The concept of word similarity is also difficult to extract since the word vectors mentioned so far are statistically orthogonal. For example, the word pairs "tumor" and "tumors", or "drug" and "medicine", are similar but are represented in different ways. Therefore, we need a more robust approach to address the discovery of similarities between words. To address early word similarity problems, distributional approaches to word representation were born.

One of the most widely used methods in the family of **distributional representations** is named Tf-Idf (Term frequency – Inverse document frequency). The Tf-Idf is a Bag-Of-Words (BOW) weighting model used to give weights to the words in a document collection by measuring how often a word is found within a document (Tf), offset by the frequency with which the word is

found in the entire collection (Idf) [106]. Often, ML models also simply use the weighted Tf to assign the number of occurrences of a word in a document. The main idea behind this approach is that words typically appearing in a similar context and document would have a similar meaning. Tf, Idf, and Tf-Idf have been compared in studies related to biomedical entity recognition as word weights. Zhang and Elhadad [107] showed that Idf and Tf-Idf weights provide improvements over using only term frequency as a weight.

One of the main drawbacks of the previous word encodings shown is the lack of meaning representation. With approaches such as one-hot or distributional, we represent the presence and absence of words in a particular text, however, we cannot determine any meaning from the simple presence/absence of these words. Part of this problem is that we lose the positional relationships between words and word order. This order in the sequence of words ends up being crucial in representing the meaning of the words and is discussed below.

Word embeddings are a family of NLP techniques that focus on mapping the semantic meaning of a word in a geometric space. For this purpose, a numerical vector is associated with each vocabulary word, so that the distance between any two vectors captures part of the semantic relationship between the two associated words. Moreover, word embeddings play a fundamental role in transfer learning, as they are trained on large amounts of corpus by using neural networks [22]. Figure 3.2 shows an example of the mapping of each word of the corpus to a dense representation vector. Thus, word vectors are positioned in the vector space so that words that share semantic meaning in

the corpus are placed very close to each other in the vector space. Specifically, the figure shows how the words "*ojo*" (eye), "*ocular*" (ocular) or "*pupila*" (pupillary) are represented in an 8-dimensional vector (size of word embeddings). Subsequently, the word embeddings can be displayed in a 2-dimensional vector space through dimensionality reduction using t-distributed Stochastic Neighbor Embedding (t-SNE). The t-SNE algorithm calculates a similarity measure between pairs of instances in the high dimensional space and the low dimensional space [108].

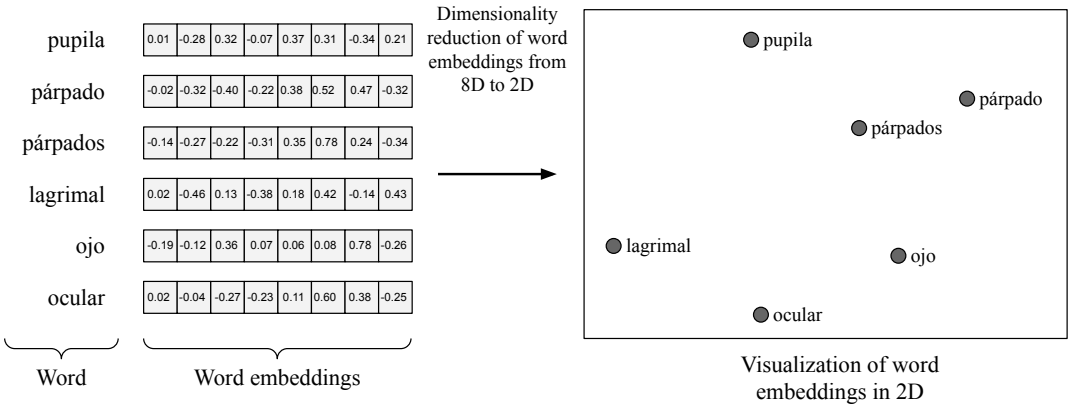

Figure 3.2: Example of word embeddings by mapping each word into a vector space.

Word embeddings were popularized by Word2Vec in 2013 [23]. Afterward, Pennington, Socher, and Manning [24] created the algorithm GloVe which aims to perform the meaning embeddings procedure of Word2Vec explicitly. Although the vocabulary of a word embedding space is large, we can find situations where a word is OOV. FastText was designed to resolve this situation by improving Word2Vec [25].

Recently, contextual word embeddings such as Embeddings from Language Models (ELMo) and BERT have emerged. These techniques generate

embeddings for a word according to the context in which the word appears, thus generating slightly different embeddings for each occurrence of the word [109]. On the one hand, ELMo is derived from a bidirectional LSTM that is trained with a coupled language model objective on a large text corpus [110], in this way, ELMo looks at the entire sentence before assigning a vector to each word. On the other hand, BERT representations are jointly conditioned on both the left and right context and use the Transformer [59], a neural network architecture based on a self-attention mechanism (cf. Section 2.2.3).

Since there are pre-trained models specific to Spanish, we can also extract the pre-trained word embeddings to represent the words in context. The most commonly used word embeddings are taken from the BETO model [65]. Furthermore, there are a variety of static word embeddings based on Word2Vec, GloVe, and FastText. Specific to the biomedical domain, Soares et al. [111] develop a word embedding using the fastText model and two sources of data: *i*) the SciELO (Scientific Electronic Library Online) database [112], which contains full-text articles primarily in English, Spanish, and Portuguese, and *ii*) the Wikipedia, with a subset which we call Wikipedia health, comprising the categories of Pharmacology, Pharmacy, Medicine and Biology. Santiso et al. [113] developed word embeddings to perform negation detection in health records written in Spanish. As a corpus, they used both biomedical domain and general domain data. For the specific domain, they used unannotated electronic medical records from a hospital in Spain. For the general domain, they used the Spanish Billion Word Corpus and Embeddings (SBWCE) corpus [25].

The integration of contextual and non-contextual word embeddings in

neural architectures has shown impressive gains in a wide variety of natural language tasks ranging from sentence classification to sequence tagging [114, 115, 116, 88, 89]. Akhtyamova and Cardiff [117] showed that contextualized word embeddings outperform other types of embeddings on a variety of tasks including NER. To do this, they used five benchmark datasets leading to significant improvements over the baseline they propose.

To conclude this section, we point out that word representation plays an important role in NLP tasks as it can be used in many different applications that require them as a resource, especially those using RNNs [90, 118, 119]. In this way, neural networks will be able to learn how a language model is represented through words, as the philosopher Wittgenstein [120] stated:

> *"The meaning of a word is its use in the language."*

3.4 Knowledge resources

Nowadays, the use of appropriate knowledge resources is essential for the development of NLP systems. In this section, we review the most popular terminology resources in the field of BioNLP and automatic entity detection tools. On the one hand, terminological resources include ontologies, controlled vocabularies, and lexicons that are available to researchers to better represent knowledge through concepts, structures, and relationships between them. On the other hand, computational frameworks have been developed to rapidly build tools for biomedical entity extraction tasks.

3.4.1 Terminological resources

The identification of entities in the biomedical literature is one of the most challenging research topics of recent years, both in NLP and in the biomedical communities. Fortunately, there are numerous manually-corrected and curated terminology resources available to the scientific community where researchers can find relevant biomedical entities.

Curated terminology resources provide the common medical language necessary for interoperability and efficient exchange of clinical data. To maximize the value of health information, these resources must be used appropriately according to their purpose, domain, and design. They are designed to serve a variety of purposes with the following benefits: *i*) health knowledge is included in an easily accessible resource, *ii*) concepts are often categorized for searching, *iii*) they facilitate data normalization, and *iv*) they allow interoperability between systems through unique concept identifiers.

Among the difficulties in successfully identifying terms are wide lexical variations, which prevent some terms from being recognized in the biomedical text, the synonymy of terms, and the homonymy of terms (when a term has several meanings), which create uncertainty as to the exact identity of the term [121]. In order to address this problem, different ontologies, controlled vocabularies, terminologies, and dictionaries containing a variety of terms have been designed. Showing the open issues and challenges, Freitas, Schulz, and Moraes [79] provided a survey of terminologies and ontologies applied to biology and medicine.

An ontology is an explicit specifications of conceptualizations [122]. The

term is borrowed from philosophy, where ontology is a systematic account of existence. However, in the computer science field, the view of ontology is somewhat narrower. A definition of ontology was given by Uschold et al. [123] describing ontology thus:

> *"An ontology may take a variety of forms, but necessarily it will include a vocabulary of terms and some specification of their meaning. This includes definitions and an indication of how concepts are inter-related which collectively impose a structure on the domain and constrain the possible interpretations of terms."*

In ontologies, the characteristics associated with the names of entities (e.g. descriptions, relations with other objects, functions, among others) describe more specifically the meaning of each one of them. Also, the relationships between entities make the ontologies well-structured [124, 125].

Currently, no ontology captures the entire range of concepts in the biomedical domain. However, despite the concerns mentioned, there are several well-designed biomedical ontologies, such as the UMLS (Unified Medical Language System) [126], the Gene Ontology (GO) [127], the EcoCyc ontology [128] and TAMBIS Ontology (TaO) [129]. UMLS and GO are the most popular biomedical ontologies in the BioNLP community since they currently involve the largest number of concepts.

- **UMLS** [126] is a collection of files and software that brings together many biomedical and health vocabularies and standards to enable interoperability between computer systems. The UMLS integrates over 2

million names for some 900,000 concepts from more than 60 biomedical vocabularies, as well as 12 million relations among these concepts. UMLS is based on three sources of knowledge:

- Metathesaurus is the major component of the UMLS containing concepts and codes from many different vocabularies, hierarchies, definitions, relationships, and attributes. Some vocabularies such as MedDRA (Medical Dictionary for Regulatory Activities) [130], SNOMED-CT (Systematized Nomenclature of Medicine — Clinical Terms) [131] and MeSH (Medical Subject Headings) [132] can be found in the UMLS Metathesaurus.
- Semantic Network which provides high-level categories (semantic types) used to categorize each concept in the Metathesaurus and relationships between concepts semantic relations.
- SPECIALIST lexicon and lexical tools containing a large syntactic lexicon of biomedical and tools for normalizing strings, generating lexical variants, and creating indexes.

The UMLS ontology has been a reference for generating corpora available to the scientific community interested in the NER task. Some corpora annotated with UMLS identifiers include MedMentions [133], MIMIC [134] and MANTRA corpus [135]. These corpora have been used by various researchers to carry out their experiments. For instance, Soldaini and Goharian [136] presented an unsupervised system that extracts medical concepts from unstructured text. Given a document,

the system extracts sections of the documents that have an approximate match in the set of UMLS strings by returning associated concepts. Song, Yu, and Han [137] also examined the impact of the dictionary on the performance by combining three different terminology resources including UMLS.

- **GO** [127] provides structured and computational knowledge about the functions of genes and gene products. GO was created in 1998 and has been widely adopted in the life sciences. It is considered an ontology that facilitates the biologically significant annotation of genes and their products in a wide variety of organisms. A great number of concepts are annotated in the GO. Specifically, this ontology includes 1,561,738 genetic products in its last version (January 2021[2]). Fortunately, GO publishes official versions of the ontology monthly, which means it is constantly updated and freely available[3].

Other terminological resources used in the biomedical domain are called controlled vocabulary. A controlled vocabulary is a hierarchically organized list of related terms. Terms included in a controlled vocabulary must have an unambiguous and non-redundant definition that makes them unique in the vocabulary. We can find relevant vocabularies in the biomedical field such as MeSH and SNOMED-CT.

- **MeSH** [132] is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine (NLM). This vocabulary is

[2]Gene Ontology releases: `http://geneontology.org/stats.html`
[3]Gene Ontology downloads: `http://geneontology.org/docs/downloads/`

used to index, catalog, and search for biomedical and health-related information. In addition, MeSH includes the subject headings that appear in MEDLINE/PubMed, the NLM catalog, and other NLM databases. In this structure, the descriptors include additional information about the attributes of the concepts and their relationships. Existing MeSH headings associated with a MEDLINE document have been used in combination with NLP methods to identify relationships between entities, including cancer types, treatments, and drug classes, among others [138, 139].

- **SNOMED-CT** [131] is the most comprehensive, clinically validated, semantically rich, and multilingual clinical healthcare terminology in the world. This vocabulary is used by physicians and other healthcare providers in the EHR to capture, retrieve and analyze clinical data. This clinical terminology is currently maintained and distributed by the International Health Terminology Standards Development Organisation (IHTSDO) and contains more than 400,000 different concepts in English. Moreover, it is a multilingual terminology that handles different languages including English, Spanish, and Swedish, and several dialects such as the German spoken in Austria.

  Some terminology embraces a formal framework that is discouraged when the semantics are overloaded. For instance, the concept "influenza" could mean the disease influenza or the virus that causes the disease. The development of SNOMED-CT ensures that a clinician is only allowed to choose the concept for influenza that "is a" respiratory tract infection

and "has a" causative agent of influenza virus.

For many years, SNOMED-CT has been used by the NLP community to identify and normalize entities within textual data. Thanks to the categorization of the resource, it has allowed researchers to focus on specific entities such as chemicals and drugs [82], disorders [140, 16], anatomical phrases [141], or even for general purposes [142]. In some of the studies mentioned above, researchers used a subset of SNOMED-CT to perform concept normalization. Furthermore, SNOMED-CT has been included as a feature in the ML systems to improve concept identification [143, 144].

Finally, other terminological resources included in the biomedical domain such as ICD and RadLex contain more specific knowledge according to the area or problem to be solved.

- **ICD** (International Classification of Diseases) defines the universe of diseases, disorders, injuries, and other related health conditions, listed in a comprehensive and hierarchical fashion. The ICD is published by the World Health Organization (WHO) and is presented in different versions. ICD-6 was first published in 1949 and today ICD in its 11th version is the most updated with 80,000 concepts. Moreover, the WHO works with specialized user groups to develop ICD-derived classifications that allow greater depth in a specialized area such as oncology (ICD-O) or primary care (ICPC). ICD-10 has been used as a benchmark in different challenges and studies. On the one hand, the previously mentioned CodiEsp track [81] required the analysis and transformation of Spanish

medical narratives into a structured or coded format using ICD-10. On the other hand, the Multilingual Information extraction task at CLEF 2018 (as in previous versions [145]) used this terminology resource to code death certificates in French, Hungarian, and Italian [146].

- **RadLex** is a complete set of radiology terms for use in radiology reporting, decision support, data mining, data logging, education, and research. The comprehensive lexicon is currently only available in English but a full translation of RadLex into German has been completed by the German Radiology Society. RadLex already contains over 8,000 anatomical and pathological terms. Many of these terms are not currently available in any other medical terminology system [147].

  Recently, literature reviews have shown that the use of RadLex is growing [148]. For instance, Datta, Godfrey-Stovall, and Roberts [149] carry out the mapping of entities by proposing two deep learning-based NLP methods based on a pre-trained language model and RadLex as terminology.

3.4.2 Mapping tools

As described above, the range and diversity of ontologies, terminologies, and lexicon have increased dramatically over the years. The demand for mapping textual data with these terminological resources has led to the creation of automatic systems and tools. Each of these systems has common characteristics, all of which employ one or more of the following features: lexical analysis, often using a specialized lexicon; syntactic analysis, a mapping procedure

that takes into account partial matching; and the use of knowledge sources to make the match.

In the biomedical domain, we can find some interesting tools for English such as MicroMeSH [150], CHARTLINE [151], and EDGAR [152]. Most of them use the UMLS ontology to match entities recognized in the textual data with the largest Metathesaurus known so far. Among the tools most commonly used by the BioNLP community, we can include cTAKES and MetaMap.

- **cTAKES** [153] is a popular system that aims to build and evaluate an open-source NLP system for the extraction of information from the textual EHR written in English. This system provides mappings to the UMLS using different components: cTAKES tokenizer, normalizer, Part-Of-Speech tagger, and NER annotator. Relevant studies evaluated cTAKES as a NER system for peripheral artery disease case discovery [154], disorder identification [155, 156] and diagnostic knowledge extraction [157].

- **MetaMap** [158] is a highly configurable application developed by NLM to map biomedical text to the UMLS Metathesaurus or, equivalently, to identify Metathesaurus concepts referred to in an English text.

  This tool employs a knowledge-intensive approach, NLP methods, and computational-linguistic techniques to identify concepts more accurately. MetaMap is used in many studies as a benchmark. For instance, Jimeno et al. [100] compared three solutions. On the one hand, they used a dictionary-based model, on the other hand, a statistical model, and finally the mapping tool MetaMap. The study showed that dictionary

searches already provide competitive results compared to the other methods. He and Kayaalp [159] used the CRF framework to compare it with the MetaMap output. The results revealed that the features the authors included in CRF were able to identify entities more accurately.

Since MetaMap and cTAKES are the most commonly used tools, there is research comparing the two systems [157, 160, 155], while others combine them to achieve better accuracy [156].

As for Spanish, there have been a few attempts to process biomedical reports in this language. To address this issue, Carrero, Cortizo, and Gómez [161] proposed a "Spanish MetaMap" that combines automatic translation techniques with the MetaMap tool. Castro et al. [162] developed a system for the identification of biomedical concepts in the Spanish language using SNOMED-CT as a knowledge source. Moreover, other systems use pipelines to address the NER task, such as that proposed by Perez, Cuadros, and Rigau [163]. In this tool, they use a sequential pipeline that retrieves mapping candidates from an indexed UMLS Metathesaurus. Then, they use the IXA pipeline [164] for basic language pre-processing, and finally, they employ UKB [165] for Word Sense Disambiguation (WSD).

- **FreelingMed** [166] system recognizes concepts included in clinical documents written in Spanish. This tool uses the Freeling analyzer [167] and extends its linguistic data with various knowledge sources including SNOMED-CT, a list of medical abbreviations, Bot PLUS, and ICD-9. Using the analyzer, the system output consists of the tokenized text, the

corresponding lemma, Part-Of-Speech tag, and also the semantic tag of the clinical entities.

- **BSB** (*Buscador Semántico Biomédico* - Biomedical Semantic Search Engine) is a prototype that includes a biomedical entity recognition system and semantic search engine [168]. The systems mentioned above often use a single source of knowledge to perform the entity identification task, however, BSB integrates different terminologies and dictionaries. BSB proposes an integrated architecture with a core focused on terminology. This architecture is composed of four main components and is illustrated in Figure 3.3:

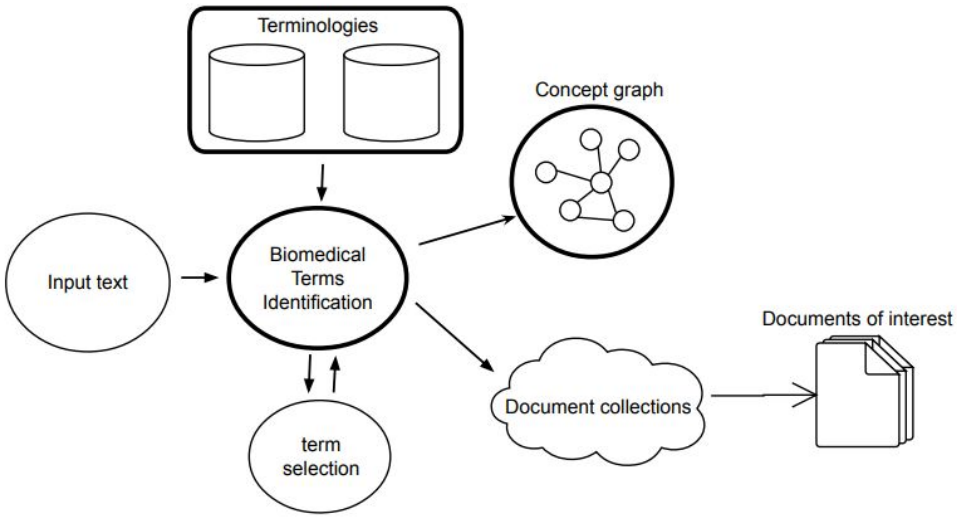

Figure 3.3: BSB architecture.

1. **Terminology knowledge bases**. SNOMED-CT, UMLS, and ICD-10 are used by this architecture to leverage information retrieval and browsing.

2. **The term identification engine**. This is the central part of the system. Starting from any free text, the system identifies the specialized concepts by carrying out several steps such as text normalization, Part-Of-Speech tag with the CoreNLP tool [169], tokenization, and partial matching
3. **The information retrieval module**. This component retrieves from different sources and collections those documents closer to the concepts identified.
4. **Concept exploration**. Semantic links between concepts provided in terminology thesauri and concept graphs, like UMLS, allow for term navigation and gathering. This is an important tool when searching for information since the semantic search is performed by exploring concepts that move in the graph of semantic relationships.

The main objective of the BSB is to serve as a proof of concept in the development of entity recognition, analysis, and medical text search systems for both expert and non-expert user communities. Tthe developed system is freely available at `http://sinai.ujaen.es/demo/bsb/`.

Finally, we can highlight that the BSB system has been one of the developments accomplished during the thesis as a knowledge discovery tool.

Chapter 4

Proposed model: combining word embeddings

In this chapter, we present the methodology applied for the recognition of biomedical entities in Spanish. This methodology will be applied subsequently in several sub-domains of biomedicine (cf. Chapter 5). The proposed system consists of three steps: *i*) we carry out a pre-processing of the text using NLP resources and methods; *ii*) we analyze, study, and create different word embedding approaches as a words representation to introduce them into the neural network; and *iii*) we describe the neural network used to finally identify relevant entities by integrating different word embeddings in a combined way.

4.1 Text pre-processing

The initial step in data science is the data preparation or text pre-processing. In textual NLP tasks, this means that any raw text needs to be carefully preprocessed before the algorithm can process it. Text pre-processing usually consists of several steps that depend on a specific task and the type of text to

be handled. In our particular case, we work with texts written in Spanish and related to a different medical sub-domain. The pre-processing carried out in all the texts is the following:

- **Sentence tokenization**. This process consists of splitting the text into individual sentences. For this purpose, we use the FreeLing library [170] that incorporates analysis functionalities for a variety of languages, including Spanish.
- **Word tokenization**. This step converts text strings to streams of token objects, where each token object is a separate word, punctuation sign, number/amount, date, e-mail, URL/URI, etc. In this step, we also use the FreeLing library. This library offers an optimal result for clinical texts since in some cases the entities, institutions, and organizations do not separate them as individual tokens, e.g. the sentence "*El síndrome de Cornelia de Lange (SCdL)*" is separated into the following tokens: "*El*", "*síndrome*", "*de*", "*Cornelia_de_Lange*", "*(*", "*SCdL*", "*)*".
- **Lowercase**. The texts have been converted to lowercase.

4.2 Word embeddings

After performing the textual pre-processing, we assign a representation vector to each word. To do this, we used an NLP technique called word embeddings. Word embedding is a representation vector that contains semantic information, which allows it to be associated with other vectors (words) according to different grammatical contexts. The vectors can be entered into neural

networks, making it easier to establish complex relationships between words because their semantics are already known.

Word embeddings are usually classified as predictive models because they are computed through language modeling objectives such as the prediction of the next word or a missing word. Moreover, word embeddings are commonly used and evaluated in two types of NLP tasks: intrinsic and extrinsic [171]. For intrinsic tasks, word embeddings are used to calculate or predict semantic similarity between words. On the contrary, for extrinsic tasks, word embeddings are used as the input for various NLP tasks, such as NER and text classification, among others. Our current focus is to study NER tasks in the biomedical domain considered as an extrinsic task.

As previously mentioned, this type of word representation was popularized by Word2Vec [23]. Afterward, Pennington, Socher, and Manning [24] created the algorithm GloVe which tries to perform the meaning embeddings procedure of Word2Vec in an explicit manner. On the one hand, Word2Vec takes texts as training data for a neural network and the resulting embedding captures whether words appear in similar contexts. On the other hand, the GloVe algorithm is focused on the word's co-occurrences over the whole corpus. Its embeddings are related to the probabilities of two words appearing together.

Although the vocabulary of a word embeddings space is large, we can find situations where a word is OOV. FastText was designed in order to solve this situation improving Word2Vec [25]. In the case of a word OOV, the corresponding word embeddings are induced by averaging the vector

representation of its character n-grams. This allows the model to compute word representations for words that did not appear in the training data

Mikolov et al. [172] propose two model architectures for learning distributed representations of words that try to minimize computational complexity: CBOW and the skip-gram model. CBOW is learning to predict the word by the context or maximizing the probability of the target word by looking at the context, while in the skip-gram model the distributed representation of the input word is used to predict the context.

Several resources related to word vector representations in Spanish are available[1] [173, 25, 111, 113], although they are still scarce compared to other internationally known languages such as English.

Since 2018, pre-trained language models showed a paradigm shift in the way NLP models were being built. The new intuition using pre-trained language models was to initialize an entire model architecture with pre-trained weights.

	Language	Word embeddings	Algorithm/Model
Classic WE	SPA	Wikipedia	FastText
	SPA	Spanish medical embeddings	FastText
Contextual WE	SPA	Pooled contextualized embeddings	Flair
	SPA	BETO	BERT
	Multi-Lang	XLM-RoBERTa	RoBERTa
	Multi-Lang	mBERT	BERT

Table 4.1: Overview of the different embeddings used. WE: Word embeddings. SPA: Spanish.

In order to address the task of word representation, we use various word embeddings: classic word embeddings and contextualized word embeddings

[1]Spanish word embeddings: `https://github.com/dccuchile/spanish-word-embeddings`

based on pre-trained language models. All these word representations are summarized in Table 4.1 showing an overview of the different embeddings, the training language, and the model or algorithm used. Also, we describe each of them below.

4.2.1 Classic word embeddings

Classic word embeddings can capture semantic and syntactic essentials of words from a large number of raw text corpora without human intervention or language-dependent processing. These word embeddings are also known as static word embeddings since the same word will always have the same representation regardless of the context where it occurs. The classic word embeddings we used for this study are presented below.

FastText embeddings from Spanish Wikipedia

The first embeddings used have been pre-trained on Wikipedia texts in Spanish. Wikipedia is a multilingual online encyclopedia created and maintained as an open collaboration project. In addition, Wikipedia is one of the largest and most popular general reference works on the World Wide Web because it offers free content on many diverse topics.

Wikipedia word embeddings are trained using the fastText method with a skip-gram model. The dimension of the word vectors is 300 and they contain 985,667 word vectors.

To clarify the word vector representation through the Spanish Wikipedia word embeddings we provide Figure 4.1 showing a graphical representation

of the word "*tumor*" (tumor) and its most similar words. This graphic uses the t-SNE algorithm in order to visualize high-dimensional data.

Since the embeddings are simply vectors of numbers such as $A = [a_1, a_2, ..., a_n]$ and $B = [b_1, b_2, ..., b_n]$, we can compute the cosine similarity between two vectors. Cosine similarity is a metric widely used to determine how similar two vectors are. This trigonometric function gives a value equal to 1 if the angle understood is zero, i.e. if both vectors point to the same place. Taking into account this measure, the words "*tumor*" (tumor) and "*carcinoma*" (carcinoma) obtain a similarity value of 77.99% using Spanish Wikipedia embeddings.

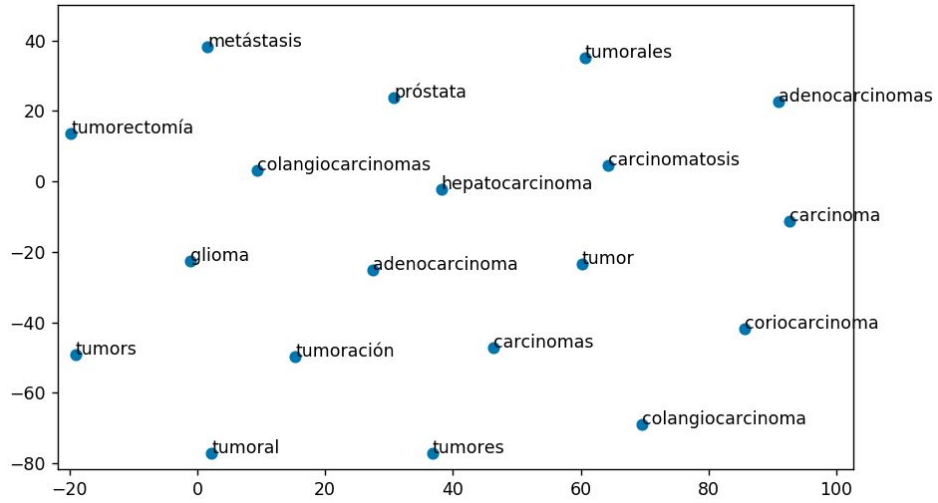

Figure 4.1: Vector space of word embeddings trained on Wikipedia in Spanish.

Spanish medical embeddings

The disadvantage of pre-trained word embeddings available in Spanish is that the words containing them may not capture the peculiarities of the language in a specific application domain. For example, Wikipedia may not have much

word coverage of particular aspects of biomedical literature or clinical cases, so the results may not be optimal due to the generality of the downloaded word embedding model. Although there are word embeddings trained for this domain [173, 25, 111, 113], they are not always available to the scientific community, therefore, we have generated the Spanish Medical Embeddings (SME)[2]. The steps to train word embeddings in a large biomedical corpus are detailed below.

1. Corpus and resource collection.

   This resource consists of an unannotated corpus of the Spanish language compiled from different corpora and resources from the web related to the biomedical domain. This corpus covers different scopes and content types, including:

   - **IBECS** (*Índice Bibliográfico Español en Ciencias de la Salud*) [174] is a bibliographical database that collects scientific journals covering multiple fields in health sciences. This database is maintained by the Spanish National Health Sciences Library (BNCS).
   - **SciELO**[112] is a bibliographic database, digital library, and cooperative electronic publishing model of open access journals. This initiative gathers electronic publications of complete full-text articles from scientific journals of Latin America, South Africa, and Spain.

[2]SME: `http://bit.do/fLTt3`

- **Pubmed** [175] is a free search engine accessing primarily the MEDLINE database, a bibliographical database of references and abstracts on life sciences and biomedical topics.
- **MedlinePlus** [176] is an online information service provided by the U.S. NLM. Moreover, this resource contains records related to Spanish health topics.
- **The OPUS - EMEA corpus** is a parallel corpus compiled from documents from the European MEdicines Agency (EMEA) [177]. The corpus includes documents related to medicinal products and their translations into 22 official languages of the European Union.
- **Portion of Wikipedia health**. Wikipedia is a free online encyclopedia, created and edited by volunteers around the world and hosted by the Wikimedia Foundation. This encyclopedia is organized hierarchically by category. We downloaded Wikipedia pages that belong to the category "*Medicina*" [178] up to the fourth sub-level of hierarchy. To do this, we employed the Wikipedia Application Programming Interface (API) for Python.
- **Web resources**. Finally, we also downloaded additional content from different websites such as Webconsultas [179], WebMd [180], Organización Mundial de la Salud [181], Mujer y Salud [182], Mejor con Salud [183], Mayo Clinic [184], Diario Médico [185], and EFE Salud [186].

2. Corpus pre-processing.

As previously mentioned, text pre-processing is a fundamental step in the NLP field. This step aims to transform the text into a more suitable form so that machine learning algorithms can perform better. The steps we have carried out to clean and prepare the texts extracted from the different resources are as follows:

(a) Removed scripts.

(b) Removed HTML tags.

(c) Lowercase.

(d) Removed URLs.

(e) Replaced multiple spaces with a single one.

3. Parameters for embeddings training.

   Once we have the text prepared, we use the fastText algorithm for the training phase of word embeddings. To train the word embeddings we used the following parameters:

   (a) The selected algorithm is the skip-gram model.

   (b) The minimum word frequency to 3.

   (c) The learning rate to 0.05.

   (d) The number of epochs to 10.

   (e) The dimension of the final word embedding is 300.

   Finally, the number of word vectors obtained was 1,704,151. Comparing these generated word embeddings and the embeddings trained on

Wikipedia, we can see that our trained word embeddings cover more vocabulary (almost double).

4.2.2 Contextual word embedding

Contextual word embeddings are considered powerful embeddings because they capture latent syntactic-semantic information that goes beyond standard word embeddings. The main goal of contextual word embeddings is to prove more than one representation for each word depending on the context in which it appears. Contextual embedding methods are used to learn sequence-level semantics by considering the sequence of all words in the documents. Thus, such techniques learn different representations for polysemous words.

The contextual word embeddings studied in our approach are pooled contextual embeddings and Transformer-based embeddings.

Pooled contextualized embeddings

Pooled contextualized embeddings are based on character-level language modeling and their use is particularly useful when the NER task is approached as a sequential labeling problem for several reasons: *i*) the embeddings are pre-trained on large unlabeled corpora, *ii*) they can capture the meaning of the words in context producing different embeddings for polysemous words depending on their usage, and *iii*) both help better handle rare and misspelled words and model sub-word structures such as prefixes and endings [187]. According to their author, these word embeddings are also known as contextual string embeddings, although researchers also refer to them as Flair

embeddings.

This type of word embeddings is based on character-level tokenization rather than word tokenization. In other words, it converts a sentence into a sequence of characters as illustrated in Figure 4.2. Each sentence is passed as a sequence of characters to a bidirectional character-level neural language model. Taking the example of the figure, the BiLSTM model allows "*de*" to retrieve information from the previous word "*cáncer*" and the next word "*mama*" so that it can compute the vectors in the context of a sentence.

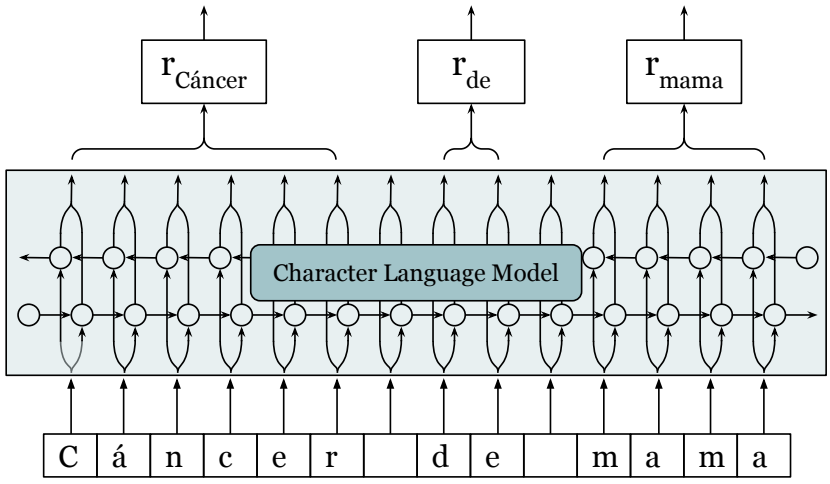

Figure 4.2: Extraction of a contextual string embedding for a sentence in a sentential context. English translation: breast cancer.

In pooled contextualized embeddings, Akbik, Bergmann, and Vollgraf [188] added a pooled operation to distill a global word representation from all contextualized instances. Clinical texts tend to be ungrammatical and limited in context, lacking even complete sentences. For example, the sentence "*La enfermedad de Carrión.*" is very short and does not provide a context for the reader. If we consider that the word "*Carrión*" is rare, i.e. this is the first time

this word has occurred in the corpus used, the underspecified context allows this word to be interpreted as a disease or a person. In addition, clinical notes make heavy use of acronyms and abbreviations, making them highly ambiguous.

Pooled contextual embeddings were created to address this problem by dynamically creating a "memory" of contextualized embedding and applying a pooling operation to distill a global contextualized embedding for each word. This means that it uses the original word embedding and the previous contextual information of each word with a pool operation to combine embedding vectors. Finally, the resulting pooled contextualized embedding has twice the dimensionality of the embedding. Figure 4.3 displays an example using the pooled operation of the word "*Carrión*" in the current sentence "*el río Carrión*" taking into account the memory of this word.

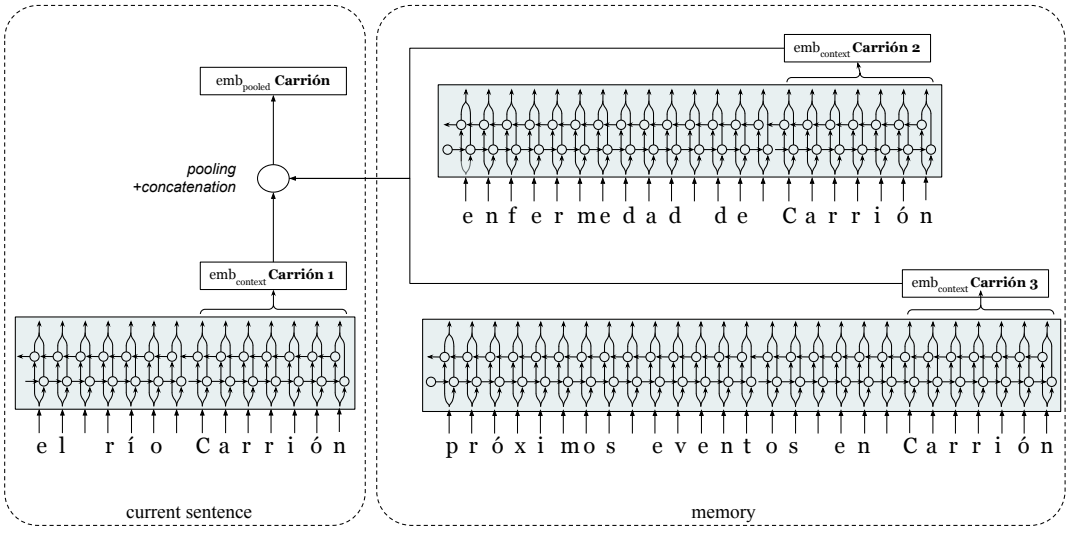

Figure 4.3: Example using the pooled operation (emb_{pooled}) for the word "*Carrión*" in the current sentence.

Embeddings based on Transformer

The idea of training a separate language model to produce better contextual word representation has proved very successful in many NLP tasks, but RNN language models, due to their recurrent and sequential nature tend to be slow to train and very hard to parallelize. For this reason, Vaswani et al. [59] developed a non-recurrent alternative to RNNs at the heart of which there is the Transformer block.

Transformer architecture replaces RNN cells with self-attention and point-wise fully connected layers, which are highly parallelizable and more cost-effective to compute. Together with positional encoding, Transformers can capture long-range dependencies. This architecture provides a more comprehensive representation of the sequence at the level of the sentence.

Unlike other language representation models, BERT [26] is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers. BERT proposes an MLM objective, where some of the tokens of an input sequence are randomly masked, and the goal is to predict these masked positions taking the corrupted sequence as input. BERT applies a Transformer encoder to attend to bi-directional contexts during pre-training.

There are several multilingual models available that also include the Spanish language. For instance, multilingual BERT (mBERT) [26] has been trained jointly on Wikipedia data on 104 languages. Spanish is also included in the cross-lingual language model (XLM-100 and XLM-17) [61], which was trained on 100 Wikipedia languages, and cross-lingual RoBERTa (XLM-RoBERTa) [27],

which was trained on much larger CommonCrawl corpora including 100 languages. Finally, BETO [65] is a BERT model trained on a big Spanish corpus.

In our experiment frameworks, we have used three specific pre-trained word embeddings. Two of them are focused on multilingual models (mBERT, XLM-RoBERTa) and the remaining one (BETO) is specifically focused on Spanish.

- **mBERT** (Multilingual BERT) released by Devlin et al. [26] is a single language model pre-trained from monolingual corpora in 104 languages including Spanish. The training set is a concatenation of monolingual Wikipedia corpora from all languages. We used the *bert-base-multilingual-cased* model with 12 self-attention layers and 110M parameters in total.

- **XLM-RoBERTa** makes a few changes to the released BERT model and achieves substantial improvements. The changes include *i*) training the model longer with larger batches and more data, *ii*) training on longer sequences, and *iii*) dynamically changing the masked positions during pre-training. For our research, we test the *xlm-roberta-base* model containing 12 self-attention layers and 125M parameters in total.

- **BETO** trained a model similar in size to a BERT-base [26]. In order to train the model, texts were collected from different sources such as Spanish Wikipedia and all the texts included in the OPUS project [189] related to the Government, subtitles, and stories, among others. The total size of the corpora gathered is comparable to the corpora used in

the original BERT. The corpus for training has about 3 billion words and it is freely available[3]. This model has 12 self-attention layers and 110M parameters in total.

In order to process the corpus, all Transformer models use a hybrid between word-level and character-level tokenization called subword tokenization. There are different models of subword tokenization, among the most important of which we can include: Byte Pair Encoding (BPE) [190], WordPiece [191], Unigram Language Model [192], and Sentence Piece [193].

The models described above have been trained using the Sentence Piece model (SPM). This tokenization technique has been created to train multilingual models since not all languages use spaces to separate words. SPM is a language-independent subword tokenizer and detokenizer designed for neural-based text processing. Other subword tokenizations assume that the input sentences are pre-tokenized. SPM treats the input as a raw stream, includes the space in the set of characters to use, then uses unigram to construct the appropriate vocabulary.

Some examples of sub-word tokenization using mBERT, XLM-RoBERTa and BETO are shown in Figure 4.4, 4.5 and 4.6 respectively. These figures show how upper and lower case is treated (examples 1 and 2), how its tokenizer treats hyphens between words (example 3), and how it deals with language errors (example 4). We can highlight that BETO and XLM-RoBERT models correctly separate complete and lowercase words such as "*hepatitis*" (hepatitis) and "*crónica*" (chronic). Furthermore, BETO and mBERT break the word into

[3]BETO corpus: `https://github.com/josecannete/spanish-corpora`

two subwords using the ## simbol, unlike XLM-RoBERTa that replaces blank spaces with underscores even at the beginning of the sentence.

(1) hepatitis B crónica ⟶ ['hep', '##ati', '##tis', 'B', 'c', '##rónica']
(2) Hepatitis B crónica ⟶ ['He', '##pati', '##tis', 'B', 'c', '##rónica']
(3) Hepatitis B-crónica ⟶ ['He', '##pati', '##tis', 'B', '-', 'c', '##rónica']
(4) hepatitis B cronica ⟶ ['hep', '##ati', '##tis', 'B', 'c', '##ronica']

Figure 4.4: Example of subword tokenization using the mBERT model. English translation: Chronic hepatitis B.

(1) hepatitis B crónica ⟶ ['_hepatitis', '_B', '_crónica']
(2) Hepatitis B crónica ⟶ ['_He', 'pati', 'tis', '_B', '_crónica']
(3) Hepatitis B-crónica ⟶ ['_He', 'pati', 'tis', '_B', '-', 'cr', 'ónica']
(4) hepatitis B cronica ⟶ ['_hepatitis', '_B', '_cro', 'nica']

Figure 4.5: Example of subword tokenization using the XLM-RoBERTa model. English translation: Chronic hepatitis B.

(1) hepatitis B crónica ⟶ ['hepatitis', 'B', 'crónica']
(2) Hepatitis B crónica ⟶ ['He', '##patitis', 'B', 'crónica']
(3) Hepatitis B-crónica ⟶ ['He', '##patitis', 'B', '-', 'crónica']
(4) hepatitis B cronica ⟶ ['hepatitis', 'B', 'cron', '##ica']

Figure 4.6: Example of subword tokenization using the BETO model. English translation: Chronic hepatitis B.

In the field of NLP, contextual word embeddings can represent each word according to the context in which it occurs. Therefore, they may eliminate some problems of ambiguity in words. It is well known that in medicine many terms can be ambiguous according to their meaning in Spanish. Some

examples of ambiguous words in the biomedical domain are "*cólera*", "*radio*", and "*frente*", among others. "*Cólera*" can be used as a mood (anger) or as an illness (cholera) and "*radio*" could be referred to broadcasting (radio), anatomy (radius) or geometry (radius). Moreover, "*frente*" could be a weather condition (front) or an anatomical part (forehead). For the last case ("*frente*"), we wanted to see an example of similarity using the same word with different meanings. To do this, we extracted the word embeddings from the BETO model for each word. Figure 4.7 illustrates the cosine similarity between word vectors. As we can see, the word "*frente*" referred to a meteorological condition has a similarity of 73% with "*frente*" related to the anatomy of the body. In contrast, the last two instances of the word "*frente*" are referred to as an anatomical part of the body, so they obtain a higher cosine similarity (90%).

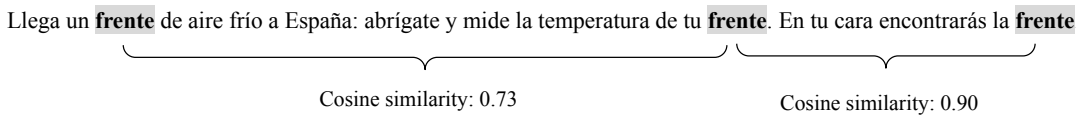

Figure 4.7: Similarity cosine in the word "*frente*" with different meanings by using BETO word embeddings. English translation: A cold air front arrives in Spain: wrap up and measure your forehead temperature. You will find your forehead on your face.

4.3 BiLSTM-CRF architecture

Once the text has been represented and the different word embeddings described, in this section we present the final step to assembling our system. Our approach will combine word embeddings using a BiLSTM neural network with a final CRF layer.

RNNs have been employed and produced promising results on a variety of tasks including language model [194, 195] and NER [196]. An RNN maintains a memory based on historical information, which enables the model to predict the current output conditioned on long-term features.

Figure 4.8 shows the basic RNN [197] with an input layer x, hidden layer h and output layer y. In the NER task, x represents input features and y represents tags or labels. This example shows the recognition of proteins in a sentence. In this case, the words that constitute the PROTEÍNA entity are "*proteína C reactiva*". In addition, the figure illustrates outputs with a special scheme. Annotations are coded using the BIO tagging scheme [198]. Thus each token in a sentence was labeled with B (beginning token of an entity), I (inside token of an entity), and O (non-entity). In the example of Figure 4.8, the annotated entity is "*proteína C reactiva*" (C-reactive protein), the category or class of the entity is PROTEÍNA, the beginning of the entity (B-PROTEÍNA) is the word "*proteína*" and "*C*" and "*reactiva*" are tokens inside the entity (I-PROTEÍNA). The BIO scheme and variants are the most popular in the NER task [199].

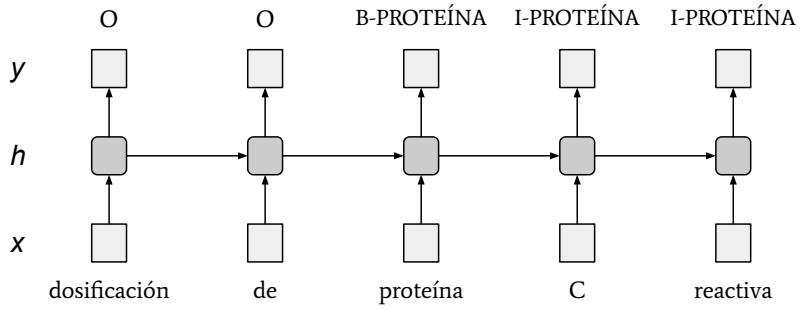

Figure 4.8: Simple architecture of an RNN model. English translation: C-reactive protein dosage.

LSTM is a variant of the above RNN that solves long-term memory. LSTMs have a more complex cell structure than a normal RNN because it allows them to better regulate how to learn or forget efficiently from the different input sources [55].

With the definition of LSTM described above, we can see that the hidden state at the time only captures information from the past. However, both past (left) and future (right) information could also be beneficial for our task. For instance, in the sentence "*Hemoglobina de 10,9 g/dl y glucosa de 55mg/dl.*", it helps to tag the word "*Hemoglobina*" as B-PROTEÍNA, if the LSTMs know the following word refers to the dose.

One shortcoming of conventional LSTM is that they are only able to make use of the previous context. In order to incorporate the future and past information in the sentence, we extend LSTM with a bidirectional approach, referred to as the BiLSTM [200]. The idea is to divide the state of the neurons of a regular RNN into one part that is responsible for the positive time direction (forward states) and one part for the negative time direction (backward states).

After applying a BiLSTM in our model, we need to predict the correct label for each token. To accomplish this task, we use the popular algorithm for NER named CRF. In the NER task, the CRF algorithm is beneficial in considering the correlations between labels in neighborhoods and jointly decoding the best chain of labels for a given input sentence [19].

Similar to the model proposed by Huang, Xu, and Yu [90], in our final architecture we combine a BiLSTM network and a CRF layer to form the BiLSTM-CRF model. As an input layer to the BiSLTM-CRF model, we use the

word vectors studied above. Current NER models often combine different types of embedding by concatenating each embedding vector to form the final word vectors. We similarly experiment with different stackings of embeddings vectors to form the input of the network. In this way, the probability of recognizing a specific word (entity) in a text should be increased as different types of representation of that word are combined. For instance, in many configurations, it may be beneficial to include classic word embeddings to add potentially greater latent word-level semantics to our proposed embeddings. The final word representation is given by Equation 4.3, where the word w_1 is represented as a concatenation of pooled embedding and a precomputed fastText embedding of that word.

$$w_i = \begin{bmatrix} w_i PooledEmbedding \\ w_i FastTexEmbedding \end{bmatrix} \tag{4.1}$$

The final architecture employed is shown in Figure 4.9. Given a sentence, the model predicts a label corresponding to each of the input tokens in the sentence. Firstly, through the embedding layer, the sentence is represented as a sequence of vectors X=(x_1,x_2,...,x_n) where n is the length of the sentence. The **combination of embeddings** are taken as input of each time step of the BiLSTM. The implicit state output sentence of forward LSTM ($\overrightarrow{h_1}, \overrightarrow{h_2}, ..., \overrightarrow{h_n}$) and the output sequence of reverse LSTM ($\overleftarrow{h_1}, \overleftarrow{h_2}, ..., \overleftarrow{h_n}$) are concatenated to obtain $h_n = [\overrightarrow{h_n}; \overleftarrow{h_n}]$ (BiLSTM output) [201]. Finally, instead of modeling tagging decisions independently, the CRF layer is added in order to decode the best tag of all the possible tags.

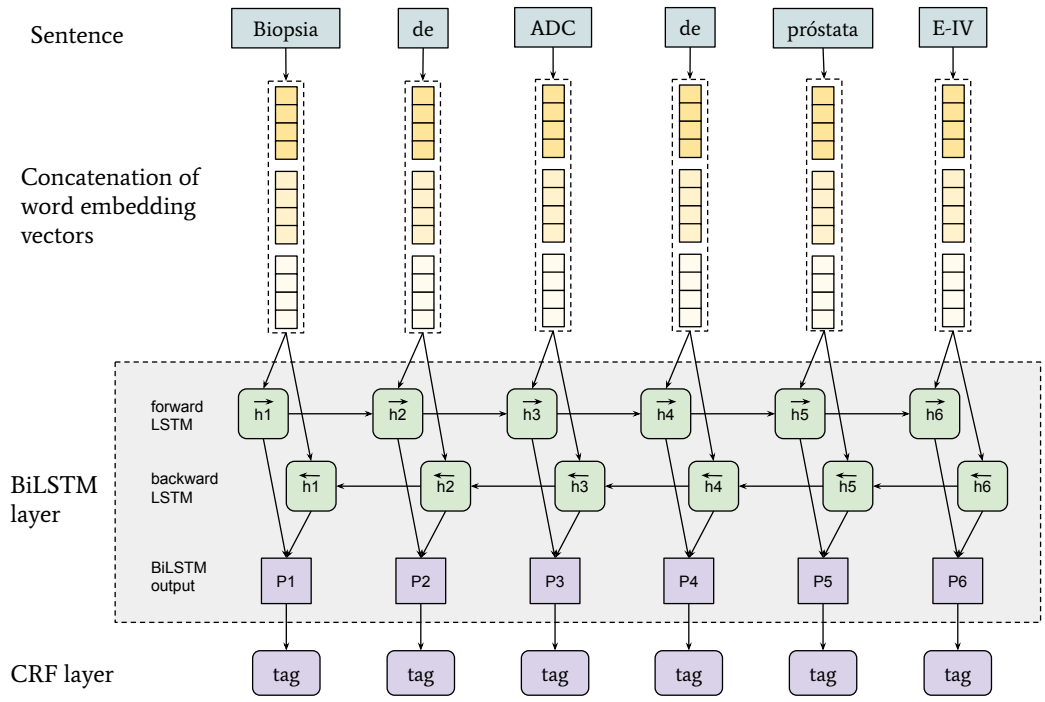

Figure 4.9: Architecture of the BiLSTM-CRF model.

4.4 Closing remarks

In this chapter, we have described the general approach we will take to the NER task in the biomedical domain. This approach is divided into several steps: first, we pre-process the textual data, then we describe the word embeddings that will be used to represent each document. Moreover, we also generate specific word embeddings for the biomedical domain because we consider that it can be very helpful to have a representation of words that have been trained with specific biomedical documents. Lastly, our approach uses a neural network composed of a BiLSTM network with a CRF layer. It is important to highlight that our methodology proposes the combination of different types of word embeddings by concatenating each embedding vector

to form the final word vectors. In this way, the probability of recognizing a specific entity in a text should be increased as different types of representation of that word are combined.

For the implementation of our system, we have employed Flair [202]. Flair is an NLP library developed by Zalando Research. Flair is built on PyTorch, as PyTorch is considered one of the best deep learning frameworks. After testing several hyperparameters, in most of our experiments Flair is used with the following configuration: learning rate as 0.1, dropout as 0.5, maximum epoch as 150, 300 neurons with *tanh* activation function, and a batch size of 32. In addition, we use early stopping to control overfitting in deep learning neural network models by stopping training before the weights have converged [203]. Finally, all experiments were performed on a single Tesla-V100 32 GB GPU with 192 GB of RAM.

Chapter 5

Experiments and results

In this chapter, we study different scenarios for applying entity recognition. All these scenarios have in common the approach described above (cf. Chapter 4). Each sub-section of this chapter contains a description of the problem addressed, the dataset used, the methodology employed and results obtained. Finally, error analysis and discussions are included for each one.

5.1 Named Entity Recognition in the pharmacological sub-domain

5.1.1 Problem description

Chemical and drugs named entity recognition is a fundamental step for further biomedical text mining and has received much attention recently. This task aims to automatically detect chemical and drug mentions in biomedical literature and is a great challenge to the scientific community for several reasons: there are several ways to refer to the same chemical or drug, abbreviations and acronyms are commonly used, symbols are often included

in scientific publications and new chemicals and drugs are constantly and rapidly reported [204].

NLP can be a solution that gives fast, accurate, and automated concept detection that can provide important advances for the NER scientific community [4].

NER in the pharmacological domain has been studied by many important researchers recently [205, 206, 207]. In addition, specific challenges such as chemical compound or drug name recognition (CHEMDNER) [10] and the extraction of drug-drug interactions from biomedical texts task (DDIExtraction) [15] have been designed to address this issue. Most of the studies that conduct the chemical and drug extraction task employ biomedical terminologies including SNOMED-CT and UMLS [208, 142].

The first challenge in the extraction of chemical and drug mentions in clinical cases written in Spanish is named PharmaCoNER [209]. PharmaCoNER was presented at the 5th Workshop on BioNLP Open Shared Tasks in 2019. The main goal of PharmaCoNER is to promote the development of NER tools of practical relevance, that is chemical and drug mentions in non-English content. This challenge proposes two sub-tasks for interested participants:

- **NER offset and entity classification**. The first evaluation scenario consists of the classical entity-based evaluation that requires that the system outputs match exactly the beginning and end locations of each entity tag, as well as the entity annotation type.

- **Concept indexing**. The second evaluation scenario consists of a concept indexing task where for each entity the list of unique SNOMED-CT

concept identifiers has to be generated.

Following the focus of this study, we propose an approach based on neural networks using a combination of word embeddings to address the first subtask of PharmaCoNER: NER offset and entity classification.

5.1.2 Corpus description

The dataset used in this challenge is the Spanish Clinical Case Corpus[1] (SPACCC). The SPACCC corpus was created by collecting 1,000 clinical cases from SciELO [112]. This type of narrative shows properties of both the biomedical and medical literature, as well as clinical records. Moreover, clinical cases involve a variety of medical disciplines such as oncology, cardiology, urology, infectious diseases, and pneumology, and these medical disciplines cover a diverse set of chemicals and drugs [209].

The annotation of the entire set of entity mentions was carried out by medicinal chemistry experts and it includes the following four entity types or categories:

- **Normalizables**: mentions of chemical compounds and drugs that can be normalized or standardized in a unique identifier in the SNOMED-CT knowledge base (e.g. vincristine, dactinomycin, and doxorubicin).
- **No-Normalizables**: mentions of chemical compounds and drugs that cannot be standardized (e.g. pyrazolone and triptans).

[1]SPACCC: `https://github.com/PlanTL-SANIDAD/SPACCC`

- **Proteinas** (proteins): peptides, proteins, genes, peptide hormones, and antibodies (e.g. transaminases, S-100, HMB-45, and PSA).
- **Unclear**: pharmaceutical formulations, general treatments, chemotherapy programs, vaccines, and a predefined set of general substances (e.g. cigarettes, alcohol, and tobacco). Mentions of this class will not be part of the entities evaluated by this challenge.

The SPACCC corpus is composed of a training set (train), a development set (dev), and a test set (test gold). Table 5.1 shows some statistics about the number of documents, sentences, and the number of entities in each set, among others.

	Train	Dev	Test gold
# of docs	500	250	250
# of sentences	8,071	3,748	3,930
# of tokens	202,901	96,869	100,963
# of unique tokens	18,623	12,170	12,442
# of Normalizables	2,304	1,121	973
# of Proteinas	1,405	745	860
# of No-Normalizables	24	16	10
# of Unclear	89	44	34

Table 5.1: Basic analysis of SPACCC corpus documents.

Entities annotated as Normalizables have an average of 1.2 words per entity although they reach a maximum of 5 words, i.e. "*heparina de bajo peso molecular*" (low molecular weight heparin). Moreover, Proteinas contain a mean of 1.4 words per entity and the entity with the largest number of words

is "*anticuerpos inmunoglobluina M (IgM) para parvovirus B19*" (immunoglobulin M (IgM) antibodies for parvovirus B19). In contrast, the entities with category No-Normalizables and Unclear have an average of 2 words. In all cases, entities are formed by words in sequential order, which means that there are no discontinuous entities.

The SPACCC corpus is distributed in Brat standoff format [210] and plain text format with UTF-8 encoding. The annotations are included in a separate document (ANN file) with the same name as the plain text document name following the standards defined in Brat. In the NER task, every line of the ANN file contains the mention string of the annotation, its start character offset, and its end character offset, which uniquely locate the mention in the text document. Figure 5.1 displays a fragment of a sample annotation.

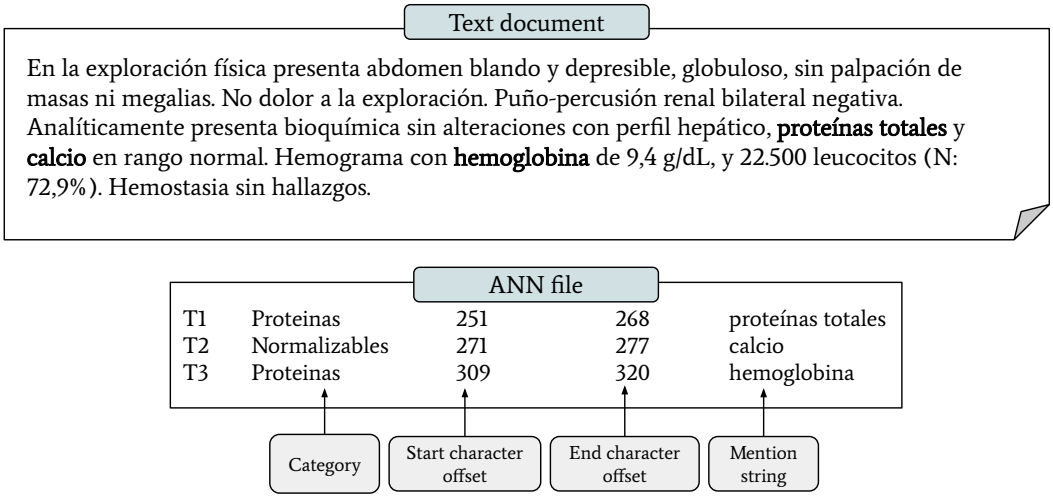

Figure 5.1: Example of annotation file for PharmaCoNER task.

5.1.3 Methodology

The methodology followed is described in detail in Chapter 4. Specifically, the architecture applied is based on a BiLSTM-CRF and a combination of word embeddings. This methodology is chosen because it facilitates the processing of arbitrary length input sequences and enables the learning of long-distance dependencies, which is useful in the case of the NER task. Furthermore, our method proposes the combination of different types of word embeddings by concatenating each embedding vector.

For experiments in this scenario, we first performed a pre-processing of the corpus (cf. Section 4.1) and then it was converted to CoNLL format [211]. The CoNLL format consists of a text file with one word per line with sentences separated by an empty line. The first word in a line should be the word and the last word should be the label. Figure 5.2 shows an example of a SPACCC corpus fragment using the CoNLL format and the BIO tagging scheme. The numbers between the first and last positions correspond to the start and end positions of the word in the clinical document.

Once we have the correct format, it is entered into the proposed neural network composed of a BiLSTM and a CRF layer (cf. Section 4.3) to train the system and subsequently predict the correct label for each word.

5.1.4 Results

In this section, we report the performance of the proposed method along with the comparison with the latest state-of-the-art studies.

To compute the metrics we used the evaluation library proposed by the

```
serología   195 204  O
positiva    205 213  O
para        214 218  O
veb         219 222  O
(           223 224  O
vca         224 227  B-Proteinas
igg         228 231  I-Proteinas
,           231 232  O
vca         233 236  B-Proteinas
igm         237 240  I-Proteinas
y           241 242  O
ebna        243 247  B-Proteinas
positivas   248 257  O
)           257 258  O
```

Figure 5.2: Example of annotation file in SPACCC corpus.

organizers of the PharmaCoNER challenge[2]. The primary evaluation metrics used consisted of standard measures from the NLP community, namely micro-averaged precision, recall, and balanced F-measure:

$$Precision = TP/(TP + FP) \tag{5.1}$$

$$Recall = TP/(TP + FN) \tag{5.2}$$

$$F - measure(F1) = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.3}$$

where TP (True Positive) is the set of samples that have exactly matched the start and end locations of each entity label, as well as the type of entity

[2]PharmaCoNER evaluation script: `https://github.com/PlanTL-SANIDAD/PharmaCoNER-Evaluation-Script`

annotation with the gold standard, FP (False Positive) refers to a system response that does not exist in the gold annotation, and FN (False Negative) is a golden annotation that is not captured by a system.

	Precision (%)	Recall (%)	F1 (%)
Wikipedia	86.65	84.8	85.71
SME	88.66	89.65	89.15
Pooled	90.56	87.03	88.76
BETO	85.9	83.37	84.61
XLM-RoBERTa	80.21	76.92	78.53
mBERT	85.06	80.87	82.91
Wikipedia + SME	90.58	90.25	90.41
Pooled + BETO	79.58	80.47	80.02
XLM-RoBERTa + mBERT	84.35	79.91	82.07
Wikipedia + SME + Pooled	**92.71**	**90.83**	**91.76**

Table 5.2: Micro-averaged performance for the NER pharmacological domain in Spanish using the BiLSTM-CRF approach.

The results of using different types of word embeddings comparing them individually and proposing different combinations are presented in Table 5.2.

For each of the experiments carried out in this thesis, we will compare the results by applying the word embeddings individually. Then, a combination of embeddings is performed following three types of grouping: *i*) classic word embeddings, *ii*) contextual word embeddings trained on Spanish, and *iii*) contextual word embeddings trained on a large multilingual corpus. Finally, further combinations of two or more embeddings are performed. Only the best result achieved is reflected in Table 5.2. However, all the results obtained can be found in Appendix A Table A.1.

As mentioned above, we first carried out an experiment using each word embedding explained in Section 4.2 individually. In terms of precision, we achieved 90.56% by applying Pooled embeddings. However, recall and F1-score obtained better results using self-trained embeddings (SME), reaching 89.65% and 89.15% respectively. The use of SME in our system is already yielding significant results, almost 90% of F1.

Subsequently, we propose a combination of word embeddings to represent the words in the corpus. We provide a combination of them according to their type: classical word embeddings (Wikipedia + SME), contextual word embeddings trained on Spanish (Pooled + BETO), and contextual word embeddings trained on a multilingual corpus (XLM-RoBERTa + mBERT). The best result with these combinations has been obtained using classical word embeddings, specifically, 90.58% precision, 90.25% recall, and 90.41% F1-score. We found a small improvement by making a combination exceeding 90% of F1.

Since the combination of classic word embeddings has obtained a significant result, we have added the pooled word embeddings to this concatenation (Wikipedia + SME + Pooled). We have selected the pooled embeddings because they also perform well in the individual evaluation. For this scenario, we obtain the best results in all metrics: 92.71% precision, 90.83% recall, and 91.76% F1-score.

The runtime of an RNN is often due to the size of the corpus and the number of unique words contained. We wanted to analyze the runtime for the PharmaCoNER corpus using different word representations to obtain time-complexity according to the size of the word embeddings. Table 5.3

	Runtime	Embeddings size
Wikipedia	02:12:07 $\pm$ 26.98	300
SME	02:08:10 $\pm$ 5.60	300
Pooled	04:38:26 $\pm$ 36.18	8,192
BETO	04:24:25 $\pm$ 33.74	3,072
XLM-RoBERTa	02:38:10 $\pm$ 51.43	768
mBERT	02:59:50 $\pm$ 16.24	3,072
Wikipedia + SME	02:10:30 $\pm$ 47.24	600
Pooled + BETO	05:18:14 $\pm$ 43.1	11,264
XLM-RoBERTa + mBERT	04:40:03 $\pm$ 46.79	3,840
Wikipedia + SME + Pooled	04:01:59 $\pm$ 53.25	8,792

Table 5.3: Running time results and word embedding sizes using the PharmaCoNER corpus.

provides both parameters: the execution time of the neural network and the size of the word embeddings used. In order to obtain these results, we use the GPU benchmark detailed in Section 4.3. Moreover, we ran each experiment three times and have shown the average and Standard Error of the Mean (SEM) in seconds. On the one hand, traditional word embeddings obtain the best runtime if they are used individually because they are the smallest embeddings (300). Afterward, Transformer-based word embeddings also obtain efficient times, except when using BETO. Pooled embeddings are the most time-consuming vectors to run and they are also the largest vectors (8,192). On the other hand, the combination of traditional embeddings (Wikipedia + SME) is still time efficient since the size of each vector is small in comparison with others. Using BETO or Pooled embeddings increases the runtime in all cases.

System	Word embedding	P (%)	R (%)	F1 (%)
BiLSTM-CRF (Our model)	Wikipedia + SME + Pooled	92.71	90.83	91.76
BERT [212]	Char emb + POS tagger + Word shape	91.23	90.88	91.05
Rule and dictionary-based [213]	-	90.62	91.31	90.97
BiLSTM-CRF [118]	SciELO Flair + SciELO fastText + BPE + Char emb	91.97	89.74	90.84

Table 5.4: State-of-the-art results for the NER pharmacological domain in Spanish. P: Precision, R: Recall, POS: Part-Of-Speech.

Since the corpus is available in the PharmaCoNER challenge, we have summarized the most relevant scientific contributions to perform a comparison of results. Table 5.4 shows the comparison of the state-of-the-art in the SPACCC corpus.

On the one hand, the study proposed by Akhtyamova et al. [118] applied a method based on BiLSTM-CRF similar to our approach. Their model is trained using the custom SciELO Flair embeddings, SciELO fastText embeddings, BPE embeddings [214] and character embeddings. On the other hand, León and Ledesma [213] used a regexp contextual rule system with a previously developed lean rule formalism [215]. Xiong et al. [212] developed a system based on BERT concatenating several features as input such as character-level representation, Part-Of-Speech tagging representation, and word shape representation of each word.

We can observe that we are superior in terms of precision and F1, although relatively weaker in terms of recall. Taking into account the F1-score, our results are 0.71 points higher than the method proposed by Xiong et al. [212], 0.79 points higher than León and Ledesma [213] and 0.92 than Akhtyamova et al. [118]. Therefore, we obtain state-of-the-art values by recognizing entities

in the Spanish pharmacological sub-domain.

5.1.5 Error analysis

In order to gain deeper insight into the proposed method performance, we have conducted an error analysis. We mainly analyze the instances in the test set that were wrongly labeled by our system in the SPACCC dataset. In terms of presence or absence of correctness entity recognition, we use the metrics TP, FP, FN, precision, recall, and F1-score. In a general test, precision is a measure of how well a test can identify TP. Sensitivity can also be referred to as recall and it is the percentage or proportion of TP out of all the samples that have the condition (TP and FN). As we detailed above, F1-score is obtained as a consequence of precision and recall.

Labeling every token as out-of-entity would be useless, but would still give a high classifier accuracy. In these scenarios, the very large class of negatives is not so interesting, and obtaining a True Negative (TN) is typically unremarkable [216]. For this reason, we do not include this metric in our error analysis.

Label	TP	FP	FN	Precision (%)	Recall (%)	F1 (%)
Normalizables	900	54	73	94.34	92.5	93.41
Proteinas	774	77	84	90.95	90.21	90.58
No-Normalizables	2	0	8	100	0.2	0.33
Overall	1,676	131	165	92.71	90.83	91.76

Table 5.5: Analysis of entity results using the BiLSTM-CRF model with Wikipedia + SME + Pooled embeddings.

A more in-depth analysis of results is shown in Table 5.5. This analysis summarizes the results obtained for each annotated entity in the SPACCC corpus using the BiLSTM-CRF model with Wikipedia + SME + Pooled embeddings. As we see, the entity Normalizables obtains the best value of F1-score (93.41%), this entity is correctly annotated 900 times (TP), although it has been mislabeled 54 and 73 times (FP and FN). The category Proteinas also reaches high values with all metrics, specifically 90.58% of F1-score. In contrast, the No-Normalizables entity is the complex one to identify since there are 24 mentions in the training set and 10 mentions in the test set. The system has been able to recognize only two (TP) obtaining a 0.33% of F1. The results of the table suggest that the system is more accurate in identifying entities the more mentions the training corpus contains.

	requiriendo	analgésicos	no	esteroideos	(	ketorolaco	)	para	el	control	del	dolor
Gold	O	O	O	O	O	B-Proteinas	O	O	O	O	O	O
System	O	B-Normalizables	I-Normalizables	I-Normalizables	O	B-Proteinas	O	O	O	O	O	O

Figure 5.3: Example of FP in the PharmaCoNER corpus comparing the gold output and the output of our system. English translation: requiring non-steroidal analgesics (ketorolac) for pain control.

	determinación	de	vimentina	,	citoqueratina	7	y	citoqueratina	de	amplio	espectro
Gold	O	O	O	O	O	O	O	B-Proteinas	O	O	O
System	O	O	B-Proteinas	O	B-Proteinas	I-Proteinas	O	B-Proteinas	I-Proteinas	I-Proteinas	I-Proteinas

Figure 5.4: Example of FN in the PharmaCoNER corpus comparing the gold output and the output of our system. English translation: determination of vimentin, cytokeratin 7 and broad-spectrum cytokera.

Regarding some errors produced by our system, we wanted to show some examples of fragments of the SPACCC corpus in which our system misclassified. An example of FP is displayed in Figure 5.3, in this figure you

can see how our system identifies the words "*analgésicos no esteroideos*" as a Normalizables entity but on the gold output it is not labeled. In Figure 5.4 we show an FN since our system identifies the entity "*citoqueratina de amplio espectro*" as Proteinas but in the gold system the correct entity is "*citoqueratina*". This is a clear example of how our system, although it is correct with the label (Proteins), does not match well the beginning and the end of the entity. Errors such as the latter shown (no matching start or end of the entity but the matching type of entity) occur 81 times. Another error of this type is found with the entity annotated on the gold as "*antigangliósidos GM1 y GD1b*" where our system recognizes "*antigangliósidos GM1*" and "*D1b*" independently. This means that the system produces three error types: one FN and two FP, because the originating entity has not been annotated by the system (one FN) and our system has produced two entities that are not in the gold standard (two FP).

Moreover, we found some errors produced by our system related to the tokenizer used. Specifically, we found three incorrect occurrences when tokenizing a protein that ends with an abbreviation and is followed by a dot. For instance, in the sentence "*Suplementos de vitamina D. No refería hábitos tóxicos.*", our tokenizer split the sentence into the following tokens: "*Suplementos*", "*de*", "*vitamina*", "*D.*", "*No*", "*refería*", "*hábitos*", "*tóxicos*" and ".". The token "*D.*" has probably been confused with the Spanish abbreviation *Don* and the tokenizer has joined the abbreviation with the dot. For this reason, our system has produced a negative output since the correct entity was "*vitamina D*" and we have identified "*vitamina D.*".

Finally, other errors are associated with the golden standard. We found two

errors annotating the entity "*TG*" (thyroglobulin). In both cases, our system has assigned the Proteinas label to that mention, but in the gold standard, it is marked as Normalizables. It should be noted that this entity is found 6 times in the training set and all instances are labeled as Proteinas.

5.1.6 Discussion

Drug and chemical name recognition aims to recognize types of mentions in unstructured medical texts and classify them into pre-defined categories. These types of tasks are fundamental to medical information extraction and medical relation extraction systems [205, 206, 207].

Given the large growth of the scientific community researching the pharmacological subdomain, the clinical NLP community has organized a series of open challenges with the focus on identifying chemical and drug entities from narrative clinical notes. These workshops are very useful because the participants use innovative and updated systems, offering a state-of-the-art approach to the tasks.

Following the neural network proposed by Huang et al [90], our approach uses the BiLSTM-CRF network to detect chemicals and drugs in Spanish biomedical literature using the PharmaCoNER [209] challenge and SPACCC corpus as a starting point. Our main goal was to prove the performance of different types of word embeddings for the NER task: classic word embeddings trained with fastText on the Spanish Wikipedia corpus, contextual embeddings that provide extra information about the context, and other word embeddings trained by ourselves adding more sources of information related

to the biomedical domain.

Our model shows that the combination of traditional word embeddings together with contextual word embeddings shows improvements over those used individually. In our experimentation, we have carried out several combinations taking into account the results of each individual embedding. Finally, we obtained the best results by combining two traditional (Wikipedia and SME) and one contextual embedding (Pooled). The results are promising since we obtain 92.71% precision, 90.83% recall, and 91.76% F1-score.

It should be noted that the results obtained in the corpus were already high, so achieving improvements allows for further research in this field. In the recognition of drugs and chemicals in Spanish, we achieved state-of-the-art results. Specifically, we showed a 0.71% improvement in F1-score and 1.48% precision compared to the best research so far [212].

The analysis of results is an important step in our study. With this analysis, we can consider the future improvements of our system. We have shown several cases of error in which our system has not labeled correctly. In particular, our error cases cover a low percentage of false negatives and false positives. Overall, we obtained 131 FP and 165 FN corresponding to the 0.07% and 0.09% of the total number (1,843 annotated entities). Our analysis suggested that we could improve on the tokenization of our texts since it sometimes did not separate the tokens most effectively. We could also treat entities that are marked as consecutive but are independently identified by our system or the opposite, for instance, our system recognizes "*isoenzimas*" and "*FA*" but the correct entity is "*isoenzimas de FA*". The entities labeled as No-Normalizables

involved only 24 instances in the training set so our system has not been able to identify this type comprehensively.

5.2 Extracting neoplasms morphology mentions from literature and electronic health records

5.2.1 Problem description

Cancer has caused nearly 10 million deaths worldwide by 2020, according to a study by the WHO International Agency for Research on Cancer (IARC) [217]. The most common cancers in recent years are female breast cancer (11.7% of the total new cases), followed by lung cancer (11.4%), colorectal cancer (10%), prostate cancer (7.3%), and stomach cancer (5.6%). Specifically in Spain, according to the Spanish Society of Medical Oncology, the number of new cancer cases diagnosed in 2020 will reach 277,394 [218]. For this reason, scientists and medical experts are making an effort to study this type of disease in depth.

Observations resulting from cancer evaluations are often reported by pathologists and documented in pathology reports. The observations described in the pathology reports are used by clinicians to guide decision-making and determine appropriate treatment and prognosis of the tumor.

Text mining has emerged as a potential solution for bridging the gap between free-text and structured representation of cancer information [219, 220]. It uses NLP, knowledge management, data mining, and ML techniques to efficiently process large document collections in order to support information

retrieval, document classification, information extraction, terminology extraction (which collects domain-relevant terms from a corpus of domain-specific documents), and named entity recognition, among other activities. In summary, NLP can facilitate the use of information from the literature and EHR in biomedical data analysis.

In the case of cancer text mining approaches, most efforts were exclusively focused on medical records in English [221, 222]. Moreover, due to the lack of high-quality manually-labeled clinical texts annotated by oncology experts most previous efforts relied mainly on customized dictionaries of names or rules to recognize clinical concept mentions. More recently, advanced technologies based on deep learning offer promising results [221, 223].

Due to the special relevance of cancer as one of the main causes of death and the increasing health costs for oncological treatments, a challenge has been created to identify entities related to oncology named Cantemist [30] (Iberian Languages Evaluation Forum - IberLEF 2020). Cantemist is the first shared task specifically focusing on the NER of a critical type of concept related to cancer, namely tumor morphology. The Cantemist task is structured into three independent subtasks, each one taking into account a particularly important usage scenario:

- **Cantemist-NER**. This subtask consists of automatically finding tumor morphology mentions.
- **Cantemist-NORM**. The second subtask requires returning all tumor morphology entity mentions together with their corresponding ICD-O codes (Spanish version: eCIE-O-3.1), i.e. finding and normalizing tumor

morphology mentions. This subtask is also known as clinical concept normalization or named entity normalization.

- **Cantemist-coding**. The last subtask requires returning for each document a ranked list of its corresponding ICD-O-3 codes.

Since the goal of our study concerns the extraction of entities, we will address the first proposed task (Cantemist-NER) focused on the biomedical domain and more specifically on the oncological field.

5.2.2 Corpus description

Cantemist corpus is a collection of 1,301 oncological clinical case reports written in Spanish. All documents of the corpus have been manually annotated by clinical experts following the guidelines[3] with mentions of tumor morphology (in Spanish, "*morfología de neoplasia*").

The corpus is composed of three sets: training, development, and testing. The task organizers have also provided two development sets (dev 1 and dev 2) so that participants can train their systems more accurately. Some statistics of the Cantemist corpus can be found in Table 5.6. In this table, we can see information related to the number of annotated entities, sentences, and vocabularies, among others.

The clinical cases were distributed in plain text in UTF-8 encoding, where each clinical case would be stored as a single file. These clinical case reports were carefully selected to represent records reflecting as much as possible the

[3]Cantemist guidelines: `https://zenodo.org/record/3878179`

	Train	Dev 1	Dev 2	Test gold
# of docs	501	250	250	300
# of sentences	22,022	10,847	9,917	12,739
# of tokens	441,993	219,172	177,574	240,562
# of unique tokens	22,280	15,391	13,921	16,551
# of Morfología_neoplasia	6,396	3,341	2,660	3,633

Table 5.6: Cantemist corpus statistics.

clinical narrative related to electronic clinical reports. Figure 5.5 illustrates an example text snippet corresponding to a short sample record that includes the following entities: "*neuroblastoma poco diferenciado*" (poorly differentiated neuroblastoma), "*metástasis*" (metastasis), "*peritumorales*" (peritumoral) and "*tumoral*" (tumoral). Note that in the Cantemist corpus the entities are always one or more consecutive words.

The Cantemist challenge provides an annotation file for each clinical case report with the entities. An example of this file is shown in Figure 5.6. This figure illustrates three entities with different identifiers (T1, T2, and T7), the start and end positions, and the text string of the entity.

```
Resultado histopatológico de piezas quirúrgicas: neuroblastoma
poco diferenciado, pobre en estroma schwaniano. Metástasis en 3
ganglios  linfáticos retroperitoneales peritumorales con
infiltración/metástasis del parénquima hepático y afectación de
márgenes quirúrgicos de la  tumorectomía retroperitoneal y
hepatectomía. Diámetro tumoral 16 cm.  Índice mitótico/cariorexis
(IMK): intermedio.
```

Figure 5.5: Example plain text Cantemist corpus document.

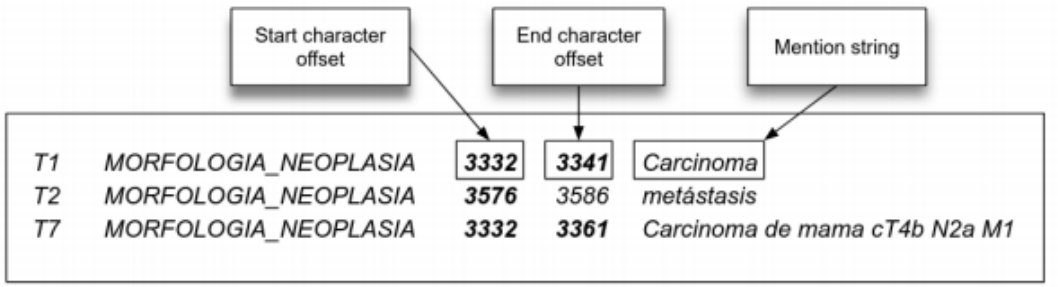

Figure 5.6: Example of annotation file for Cantemist corpus.

The size of the entities is an important aspect to take into account in the corpus statistics because a large entity is probably more difficult to detect than a simple entity of one or two words. An example of a large entity is "*carcinoma intraductal predominantemente grupo 3 de Van Nuys (micropapilar y sólido de alto grado citológico con necrosis) con cancerización lobulillar*" (intraductal carcinoma predominantly Van Nuys group 3 -micropapillary and solid high grade cytological with necrosis- with lobular cancerization) since it contains 20 words in Spanish. Table 5.7 summarizes some statistics on the number of words contained in the entities for each set of the corpus. All sets contain entities with only one word and the average number of words in an entity is usually between 2.25 and 2.3 which is reasonable. However, the maximum number of words in all sets is above 16.

	Train	Dev 1	Dev 2	Test gold
Minimum of words within the entity	1	1	1	1
Mean of words within the entity	2.30	2.27	2.30	2.25
Maximum of words within the entity	16	17	20	25

Table 5.7: Statistics on the number of words within the entities in the Cantemist corpus.

To conclude this section, we would like to emphasize another important issue regarding the entities annotated in the corpus. The corpus Cantemist contains large entities and on some occasions, we have verified they have identified entities inside other entities. Figure 5.7 presents an example of a large entity that starts at position 3,058 and ends at position 3,122, it also shows how there is another entity mention ("*metastásica*") at position 3,069, which means that one entity is annotated within another. As we have been able to analyze, we have found 30 entities included within other mentions in the training set, 15 entities in development set 1, 8 entities in development set 2, and 17 in the test set.

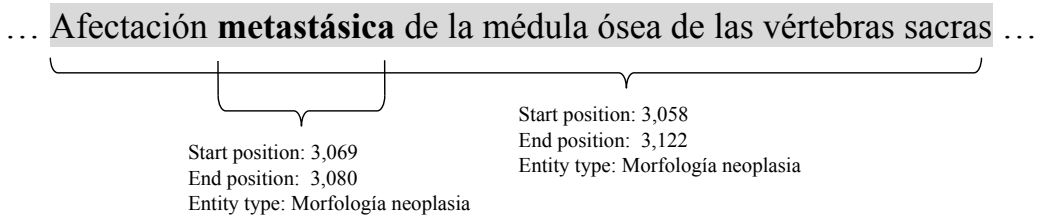

Figure 5.7: Example of a long annotated entity with another annotated entity inside it in the Cantemist corpus. English translation: Metastatic involvement of the bone marrow of the sacral vertebrae.

5.2.3 Methodology

The methodology followed to address this scenario is described in Chapter 2. We first perform a pre-processing of the Cantemist corpus (cf. Section 4.1). Afterward, the corpus is converted to the CoNLL format. This format admits one word for each line of the file along with the start and end positions and the label assigned to each word using the BIO tagging scheme (see Figure 5.8).

Finally, we select the word embeddings we are going to use and we carry out the experimentation in the BiLSTM-CRF neural network.

```
metástasis   3044 3054   B-Morfología_neoplasia
pulmonares   3055 3065   I-Morfología_neoplasia
de           3066 3068   I-Morfología_neoplasia
cáncer       3069 3075   I-Morfología_neoplasia
de           3076 3078   O
páncreas     3079 3087   O
intervenido  3088 3099   O
con          3100 3103   O
extensión    3104 3113   B-Morfología_neoplasia
abdominal    3114 3123   I-Morfología_neoplasia
.            3123 3124   O
```

Figure 5.8: Example of annotation file in the Cantemist corpus.

Since the organizers of the challenge delivered two development sets (dev 1 and dev 2), we decided to use training and development set number 1 as training, while number 2 development set was used to validate our system. The reason is that development set 1 contains a greater number of annotated entities and more tokens than development set 2.

5.2.4 Results

In order to represent the results obtained, this section has been distributed as follows: *i*) we show the metrics used for the evaluation of the system, *ii*) we present the results achieved using a combination of word embeddings, *iii*) we show the runtime of the neural network according to the embeddings used, and *iv*) we conduct a comparison with the state-of-the-art studies using this corpus.

The evaluation metrics used are proposed by the organizers of the Cantemist challenge. In this case, the main evaluation metric has been micro average F1-score. In addition, precision and recall have been computed as follows:

$$Precision = TP/(TP + FP) \tag{5.4}$$

$$Recall = TP/(TP + FN) \tag{5.5}$$

$$F - measure(F1) = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.6}$$

where TP is the set of samples that have exactly matched the start and end positions of each entity label, as well as the type of entity annotation with the gold standard, FP refers to a system response that does not exist in the gold annotation, and FN is a golden annotation that is not captured by a system.

In order to extract entities related to malignant neoplasm and as in the previous experiments, we have presented several combinations of embeddings. As in the previous experiment, we present three evaluation scenarios in Table 5.8. On the one hand, we have used the embeddings individually. On the other hand, we have combined them according to their type and language (classic word embeddings, contextual embeddings trained on Spanish, and contextual embeddings trained on multilingual corpus). Finally, other different combinations of more than two word embeddings are presented.

Firstly, considering that word embeddings are used independently, we

	Precision (%)	Recall (%)	F1 (%)
Wikipedia	81.2	80.8	81
SME	83.4	**86.6**	85.2
Pooled	84.9	83.8	84.4
BETO	81.9	75.5	78.3
XLM-RoBERTa	80.7	73.7	77
mBERT	81.9	74.8	78.2
Wikipedia + SME	84	86.1	85
Pooled + BETO	81.8	75	78.6
XLM-RoBERTa + mBERT	81	72.6	76.6
Wikipedia + SME + Pooled	**85.9**	85.1	**85.5**

Table 5.8: Micro-averaged performance for NER in the oncological domain in the Cantemist corpus using the BiLSTM-CRF approach.

obtain the best precision value by using Pooled embeddings. However, the best recall and F1-score value have been achieved with SME embeddings (86.6% and 85.2% respectively). The worst result in all metrics has been obtained using XLM-RoBERTa Transformer-based embeddings (77% F1-score).

Secondly, we test several combinations according to their origin and whether they are multilingual or monolingual. In this case, we find that these combinations (the second part of the table) obtain slightly worse results in their combination than in the use of them independently, e.g. SME embeddings perform better individually than in combination with Wikipedia. But this is not always the case, for example, the combination of Wikipedia + SME improves the word embeddings of Wikipedia individually, and the combination of Pooled + BETO improves BETO.

Lastly, to make other combinations, we take into account the word representations that obtained high results individually. Wikipedia, SME, and Pooled embeddings perform better in an individual way, therefore, we have decided to implement the combination of Wikipedia + SME + Pooled. They obtain the highest value achieved, specifically 85.9% precision, 85.1% recall, and 85.5% F1-score. Other combinations and their results using the Cantemist corpus can be found in Appendix A Table A.1.

	Runtime	Embeddings size
Wikipedia	05:02:41 ± 44.54	300
SME	06:19:01 ± 1:24.12	300
Pooled	15:56:25 ± 9:10.38	8,192
BETO	08:21:24 ± 28.38	3,072
XLM-RoBERTa	08:51:58 ± 1:55.08	768
mBERT	07:47:05 ± 1:04.42	3,072
Wikipedia + SME	06:24:10 ± 1:11.28	600
Pooled + BETO	1 day 4:09:17 ± 1:06.07	11,264
XLM-RoBERTa + mBERT	10:47:03 ± 1:21.26	3,840
Wikipedia + SME + Pooled	20:54:41 ± 1:52.19	8,792

Table 5.9: Running time results and size of word embeddings using the Cantemist corpus.

The execution time and size of the word vectors are detailed in Table 5.9. As in previous experiments, the least time spent has been using traditional word embeddings (SME and trained on Wikipedia). Even when combining traditional word embeddings (Wikipedia + SME) we obtain better response time due to the number of vectors used (300 + 300). As we can see, Pooled embeddings have long execution times since one word is represented with a vector size of 8,192 elements. On the other hand, word embeddings based

on Transformers obtain the same runtime approximately (between 7 and 8 hours). The longest time obtained has been using Pooled + BETO embeddings as together each word is represented with a vector size of 11,264. According to Table 5.8, our best result has been obtained by using the combination of Wikipedia + SME + Pooled with an execution time of 20h and a standard error of the mean of 1 minute and 52 seconds. With this configuration, each word is represented with a vector of 8,792 numbers.

System	Pre-trained model/word embedding	P (%)	R (%)	F1 (%)
BERT [224]	Multilingual BERT	87.1	86.8	87
Ensemble model based on Transformer [225]	BETO and SciBERT [64]	86.8	87.1	86.9
BiLSTM-CRF (Our model)	Wikipedia + SME + Pooled	85.7	85.2	85.5
BERT [226]	Multilingual BERT, character and fastText embeddings.	85.4	85.2	85.3

Table 5.10: State-of-the-art results for the extraction of neoplasm morphology mentions in Spanish. P: Precision, R: Recall.

The comparison of results with other studies has been carried out thanks to making the dataset and the evaluation script available to the scientific community. Moreover, this challenge has been a reference point for researchers focused on this domain, language, and task since it had great popularity in 2020 [30]. Specifically, a total of 23 teams (mostly from the academic community) participated in the task by sending in up to 62 runs. Although the task is in Spanish, teams from up to 16 countries participated, which shows considerable attention from the community.

Table 5.10 summarizes the four top results of the Cantemist challenge. Xiong et al. [224] have obtained the highest F1-score (87%). Their system is

highly balanced: 87.1% precision and 86.8% recall. It is almost equivalent to the F1-score obtained by García-Pablos, Perez, and Cuadros [225] (86.9%). We rank third using our methodology and reaching an 85.5% F1-score [119]. Finally, Lange et al. [226] obtained 85.4% precision, 85.2% recall and 85.3% F1.

Most successful teams employ machine learning approaches for NER tasks. For instance, Xiong et al. [224] propose a model considered as a Machine Reading Comprehension (MRC) problem inspired by Li et al. [227] and Levy et al. [228], whose task is to answer questions regarding different types of entities based on given passages. A Transformer-based language model and a classification layer are employed by García-Pablos, Perez, and Cuadros [225] using voting ensembles. Our method employs a BiLSTM-based model with a CRF layer by combining different word embeddings and Lange et al. [226] use multilingual BERT, character and fastText embeddings [111] similar to Yu, Bohnet, and Poesio [229].

5.2.5 Error analysis

This section presents the results in a more comprehensive way. For this purpose, we first use the measures TP, FP, and FN to provide the number of entities that our system correctly detected and the number of entities that it misclassified.

The Cantemist corpus, unlike the previous scenario, has only one type of annotated entity (morphology of neoplastic), therefore, we show the results according to that category. Note that we present the results of this section using the best method, which means that it uses a BiLSTM-CRF network with

the Wikipedia + SME + Pooled word embeddings.

Based on the analysis of the Cantemist corpus, Table 5.6 showed that the dataset contained 3,633 entities of malignant neoplasm type in the test set. Taking into account this number, our method has obtained 3,096 TP, 521 FP, and 537 FN. Thus, the number of entities correctly annotated by the system is 3,096 (85.22% of the total). It also identified 521 entities (14.34% of the total number) that were not included in the gold standard (FP), and finally, the proposed system did not recognize 537 entities (14.78% of the total) that were labeled in the gold standard (FN). The gold standard contains a total of 3,633 entities, which is obtained by adding the number of successes (TP) and the number of FN.

For further error analysis, we show examples where the proposed approach has recognized incorrect entities. Figure 5.9 presents a fragment of the test dataset in which our system has classified incorrectly. Specifically, the figure shows how the gold standard contained the entity "*lesión a nivel hipofisario*" (lesion at pituitary level) labeled as cancer morphology, but our method has not been able to identify it. If we take into account the number of occurrences of some of the words included in that entity, the word "*hipofisario*" (pituitary) occurs 9 times in the Cantemist training corpus, of which 6 times it appears with the word "macroadenoma" (which is more related to cancer). In the test set, the proposed approach correctly detected "pituitary macroadenoma" but incorrectly "lesion at the pituitary level" as shown in Figure 5.9, demonstrating that the system is locating words associated with the oncology domain.

Often, cancer-related entities in the Cantemist corpus detail a part of the

	engrosamientos	a	nivel	de	la	duramadre	y	lesión	a	nivel	hipofisario
Gold	O	O	O	O	O	O	O	B-Morfología_neoplasia	I-Morfología_neoplasia	I-Morfología_neoplasia	B-Morfología_neoplasia
System	O	O	O	O	O	O	O	O	O	O	O

Figure 5.9: Example of FN in the Cantemist corpus comparing the gold output and the output of our system. English translation: thickening at the dura mater and injury at the pituitary.

body along with indications of tumors, i.e. "*carcinoma de ovario izquierdo*" (left ovarian carcinoma), "*tumor urotelial*" (urothelial tumor), "*linfoma folicular*" (follicular lymphoma), and "*adenocarcinoma intestinal*" (intestinal adenocarcinoma), among others. For this reason, it is important to focus on those words that are related to cancer along with parts of the body.

An example of FP is illustrated in Figure 5.10. In this case, the proposed method has labeled the entity "*pseudotumor inflamatorio*" (inflammatory pseudotumor) as a neoplastic morphology when in the gold standard it was tagged as a non-entity. Although it seems that the entity is referred to as a malignant tumor, if we look at the full context of the sentence, it claims the following: "*Otras opciones de patología no maligna incluyen infecciones virales por CMV o un pseudotumor inflamatorio*" (Other non-malignant pathology options included viral infections from CMV or an inflammatory pseudotumor). So, non-malignant pathologies are being listed in this document. Pseudotumor is a word that does not occur once in the training corpus and only once in the test set. Probably, our system has misclassified this word because of the close relationship it has with the word tumor.

The positions of the entity mentions are also an aspect to take into account to identify them correctly. Indicating a start or end position incorrectly means an FP and an FN at the same time because we annotate an entity that does

	incluyen	infecciones	virales	por	CMV	o	un	pseudotumor	inflamatorio
Gold	O	O	O	O	O	O	O	O	O
System	O	O	O	O	O	O	O	B-Morfología_neoplasia	I-Morfología_neoplasia

Figure 5.10: Example of FP in the Cantemist corpus comparing the gold output and the output of our system. English translation: included viral infections from CMV or an inflammatory pseudotumor.

not exist in the gold standard (FP) and also forgets to recognize an entity included in the gold (FN). In order to conduct this analysis, we carried out a semi-automatic test and discovered that our system marked the end position incorrectly 88 times in the entities and 282 times in the start position. To better clarify these errors, Figure 5.11 shows an example of the gold notation included in the test set with the start and end position. In addition, Figure 5.12 illustrates the output of entities extracted by our system. As we can see, in the gold standard the entity ranges from position 3,042 to 3,090 and contains 7 words ("*cáncer de mama localmente avanzado o metastásico*"). However, our system only recognized 2 words independently ("*cáncer*" and "*metastásico*"). Therefore, this error results in one FN and two FP.

```
MORFOLOGÍA_NEOPLASIA    3042 3090 cáncer de mama localmente
                                  avanzado o metastásico
```

Figure 5.11: Example of a gold standard annotation in the Cantemist test set. English translation: locally advanced or metastatic breast cancer.

```
MORFOLOGÍA_NEOPLASIA    3042 3048 cáncer
MORFOLOGÍA_NEOPLASIA    3079 3090 metastásico
```

Figure 5.12: Example of our system's annotation in the Cantemist test set. English translation: cancer and metastatic.

Example	Test set	Entity
# 1	Gold	1. *progresión tumoral a nivel hepático* 2. *tumoral*
	System	1. *progresión tumoral a nivel hepático*
# 2	Gold	1. *afectación metastásica pulmonar bilateral* 2. *metastásica*
	System	1. *metastásica*
# 3	Gold	1. *adenocarcinoma (ADC) infiltrante* 2. *ADC*
	System	1. *adenocarcinoma* 2. *ADC*

Table 5.11: Examples of misclassification of our system by having entities within other entities. English translation: #1 tumor progression to the liver level, #2 bilateral pulmonary metastatic involvement, and #3 infiltrating adenocarcinoma (ADC).

The last error analysis performed has been about recognizing entities included within other entities. As we know, the Cantemist corpus contained entities that were included in other entities according to their start and end positions (see more details in Figure 5.7). Our proposed approach and the labeling scheme carried out did not take into account these types of entities, so it has not been able to detect them independently from each other. Table 5.11 lists errors obtained by the proposed approach when there are entities that are part of other entities. This table presents three examples indicating the mentions that the system has to detect (gold) and those that our system has identified (system). For instance, in example one the gold standard contains two entities: "*progresión tumoral a nivel hepático*" (tumor progression to the liver level) and "*tumoral*" (tumoral), but our system has only detected the first one, ignoring "*tumoral*" as a separate entity. In example two, gold standard included two entities: "*afectación metastásica pulmonar bilateral*" (bilateral pulmonary

metastatic involvement) and "*metastásica*" (metastatic), however, the method identified only "*metastásica*" as an entity. Finally, in example three, the gold test also contained two entities: "*adenocarcinoma (ADC) infiltrante*" (infiltrating adenocarcinoma - ADC) and "*ADC*", but in this case, the system independently detected "*adenocarcinoma*" and "*ADC*".

5.2.6 Discussion

Due to the large amount of information stored in a clinical report related to the oncology domain, the NLP community proposes challenges such as Cantemist. Cantemist is a new shared task on Spanish NLP included in the conference of the Spanish Society of Natural Language Processing (SEPLN). Cantemist is composed of three sub-tasks: NER, concept normalization, and clinical coding specifically. Our study focuses mainly on the detection and classification of mentions of biomedical entities, so we have the opportunity to employ our methodology in the oncological domain using the Cantemist challenge.

The Cantemist corpus is a pioneer work on the distribution of domain-specific medical NLP corpus and in languages other than English. The dataset contains a total of 1,301 oncological clinical case reports written in Spanish and contains a total of 16,030 entities related to neoplastic morphologies. As we have studied, these entities have relevant characteristics described below:

- **The entities are very different in size**. The size of the annotated entities in this corpus is diverse. We can find entities with one word or entities of up to 25 words. The most frequent size of an entity is around two to

three words.

- **The entities contain extra information**. Since some entities in the corpus are large, they contain more relevant information within them. Frequently, entities include adjectives that modify or enhance the noun of the entity, for example, the entities "*carcinoma epitelial maligno*" (malignant epithelial carcinoma) and "*carcinoma de células pequeñas*" (small cell carcinoma) contain the words "malignant" and "small". In addition, they often include parts of the human body to differentiate where the tumor is located. For example, in the entities "*carcinoma renal*" (renal carcinoma) and "*carcinoma de pulmón derecho*" (right lung carcinoma), carcinomas are found in the renal part of the patient and the right lung.

- **The entities include other entities within them**. Finally, another important feature within the annotated entities is that they may contain other entities within them. As we have analyzed, in many cases the entities are usually large and the entities located inside are one word. Also, several of the entities found inside are usually acronyms, for instance, the entity "*metástasis de carcinoma - M1* " (carcinoma metastasis - M1) contains the separate entity "M1" and "*tumor estromal gastrointestinal maligno (GIST)*" (malignant gastrointestinal stromal tumor - GIST) includes "GIST".

Regarding the results obtained, we did not reach the state-of-the-art using the Cantemist corpus, but we are satisfied to be in the top three. Considering the large number of participants registered in this challenge, we achieved 85.7% precision, 85.2% recall, and 85.5% F1-score. Xiong et al. [224] obtained

the best result reaching 87% F1-score using BERT and a multilingual pre-trained model.

5.3 Knowledge extraction and discovery from health texts

5.3.1 Problem description

The large amount of clinical text available online has motivated the development of automated knowledge discovery systems to analyze this data and discover relevant evidence. These discoveries or findings may serve as the basis for new treatments, understanding of disease and drug-drug interactions.

The eHealth Knowledge Discovery (eHealth-KD) challenge leverages a semantic model of human language that encodes the most common expressions of factual knowledge by using general-purpose entity types and thirteen semantic relations among them.

To address the task of biomedical information extraction, the eHealth-KD shared task has been conducted for two years [230, 231]. The organizers provide different NLP tasks to automatically extract a variety of knowledge from electronic health documents written in Spanish. In 2020 [232], eHealth-KD proposes two subtasks related to capturing the semantic meaning of health-related sentences: *i*) entity recognition (subtask A), whose goal is to identify all the entities in a document and their types; and *ii*) relation extraction (subtask B), which seeks to identify all relevant semantic relationships between the entities recognized. Even though this challenge is oriented to the health domain, the structure of the knowledge to be extracted is general-purpose.

The challenge has been included at IberLEF 2020 (SEPLN) and involved the participation of eight research teams from different institutions.

Using this scenario, we focus on the NER task (subtask A). The main objective of this task is to identify all the entities included in the document and also to determine the category of this entity. These entities are relevant terms (single word or multiple words) that represent semantically important elements in a sentence.

5.3.2 Corpus description

The corpus is composed of a list of health documents written in Spanish extracted from MedlinePlus [176]. MedlinePlus is an online information service provided by the NLM that includes health information and a medical encyclopedia covering hundreds of diseases, conditions, and general wellness topics [233]. Recent research has used MedlinePlus as a source of information to create their approaches to NLP and continue their studies in the biomedical domain [234, 160].

MedlinePlus files contain several entries related to health and medical issues and have been processed to remove all XML markup extracting the textual content. Once the convenient text was extracted, the organizers applied additional preprocessing to remove undesirable phrases, such as headers, footers, and similar elements, and to flatten HTML lists into simple sentences. The final documents are manually tagged using Brat standoff format [210] by a group of annotators. After tagging, the Brat output file (ANN format) was processed to obtain the output files in the desired format.

In this corpus, the annotated entities always consist of one or more complete words and never include any surrounding punctuation symbols, parenthesis, etc. The types of entities annotated in this corpus are:

- **Concept**: identifies a relevant term, concept or idea, in the knowledge domain of the sentence.
- **Action**: identifies a process or modification of other entities. It can be indicated by a verb or verbal construction, such as "*afecta*" (affects), but also by nouns, such as "*problemas*" (problems) and "*análisis*" (analysis).
- **Predicate**: identifies a function or filter of another set of elements, which has a semantic label in the text such as "*parte posterior*" (backside) and is applied to an entity such as "*ojo*" (eye).
- **Reference**: identifies a textual element that refers to an entity (of the same sentence or different one) which can be indicated by textual clues such as "*estos*" (these), "*su*" (your), etc.

Figure 5.13 shows an example of three sentences with the relevant entities annotated. In this example, we can see some entities that span more than one word such as "*vías respiratorias*" (airway) and "*60 años*" (60 years).

The corpus is divided into different sets including training, development, transfer, ensemble, and testing. The development and transfer sets contain sentences related to the biomedical domain and an alternative domain (Wikinews) respectively. This alternative domain is proposed by the organizers to evaluate other scenarios outside the medical domain. Moreover, the unreviewed

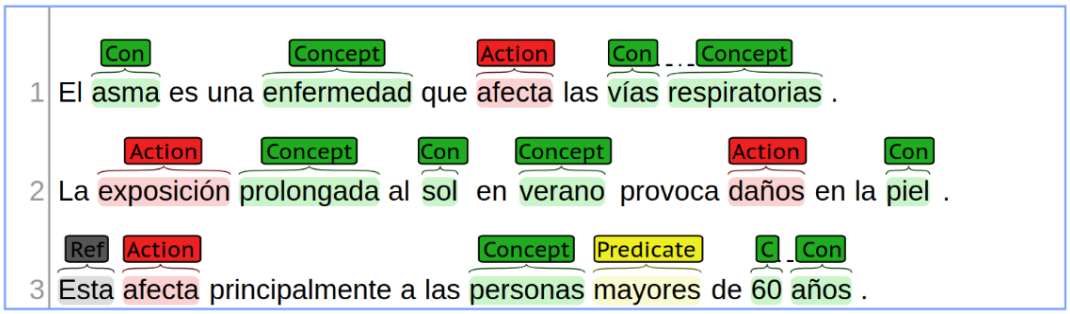

Figure 5.13: Example of sentence annotation for the eHealth-KD challenge.

dataset named "ensemble" contained the submissions from past editions. Table 5.12 shows a statistical summary of the eHealth-KD corpus.

	Train	Dev	Ensemble	Transfer	Test gold
No. sentences	800	200	3,000	100	100
No. tokens	12,867	3,350	38,965	3,016	1,455
No. unique tokens	2,484	1,085	3,667	1,245	593
No. Concept	3,112	797	9,513	841	377
No. Action	1,319	340	3,866	278	121
No. Predicate	412	124	963	104	49
No. Reference	169	44	403	19	9

Table 5.12: Summary statistics of the eHealth-KD Corpus.

In our case, since the organizers offer different datasets, the training, ensemble and transfer sets (800 + 3,000 + 100 sentences respectively) constitute the training set, while the development set (200 sentences) was used to validate the system.

An important feature of this corpus relates to discontinued entities. Discontinuous entities are referred to as those mentions that contain non-continuous words, in other words, they are not sequential entities since they contain

words that are not labeled as an entity. The number of entities concerned is reduced: in the training set 8 entities, in the validation set 2, and in the ensemble set 1. The other groups did not contain such entities. An example of this type is presented in Figure 5.14. In this figure, we can see the sentence and the annotated entities. The entity "*sentidos del olfato*" (olfactory senses) is not an entity with sequential words, moreover, in this entity, some words are overlapped with words from another entity, in this case, the word "*sentidos*".

```
- Sentence:
Los sentidos del gusto y el olfato nos brindan gran placer.

- Annotated entities:
sentidos del gusto
sentidos del olfato
```

Figure 5.14: Example of a discontinued annotated entity in the eHealth-KD corpus. English translation: The senses of taste and smell give us great pleasure.

The entities included in the eHealth-KD corpus have different sizes depending on the category of the entity being labeled. We have conducted a simple analysis counting the number of words that make up the entity. This analysis is shown in Table 5.13. As we can appreciate, the entities labeled with the category Concept include a larger number of words than the other categories. Specifically, these entities can contain up to 9 words such as "*Centros para el Control y la Prevención de Enfermedades*" (Centers for Disease Control and Prevention). Entities classified as References are usually the shortest ones containing only one word. Examples of this type of entity are "*esta*" (this) and "*ambas*" (both).

		Train	Dev	Ensemble	Transfer	Test gold
Concept	Min	1	1	1	1	1
	Mean	1.24	1.26	1.20	1.37	1.23
	Max	9	6	7	7	5
Action	Min	1	1	1	1	1
	Mean	1	1.01	1	1	1.02
	Max	3	3	1	2	3
Predicate	Min	1	1	1	1	1
	Mean	1.04	1.04	1	1.04	1.16
	Max	4	2	4	2	4
Reference	Min	1	1	1	1	1
	Mean	1	1	1	1	1
	Max	1	1	1	1	1

Table 5.13: Statistics on the number of words within the entities in the eHealth-KD corpus.

5.3.3 Methodology

The steps followed to carry out the experimentation in this corpus have been the following:

1. Pre-processing of the text included in the medical documents (cf. Section 4.1).

2. Convert each set to CoNLL format. For this purpose, a file is generated for each set of the corpus in which it includes mainly one word per line together with the assigned tag. In order to clarify this step, we show in Figure 5.15 an example in which we see a fragment of the training corpus. This figure shows three of the four existing categories in the corpus. The entities labeled as Concept are three: "*padres*" (parents),

"*hijos*" (children) and "*peligros*" (dangers); there are two entities of the Action category in this fragment: "*deben*" (must) and "*conocer*" (know); and finally, the entity "*estos*" (these) is a Reference type, in reference to the problems mentioned in the previous sentence.

```
los         3029 3032   O
padres      3033 3039   B-Concept
y           3040 3041   O
los         3042 3045   O
niños       3046 3051   B-Concept
deben       3052 3057   B-Action
conocer     3058 3065   B-Action
estos       3066 3071   B-Reference
peligros    3072 3080   B-Concept
.           3080 3081   O
```

Figure 5.15: Example of annotation file in eHeatlh-KD corpus. English translation: Parents and children should be aware of these dangers.

3. Select the word embeddings for the performance of the experiments (Section 4.2).

4. Train the system with the corpus by using the proposed methodology (Section 4.3) and then evaluate the output.

The eHeatlh-KD dataset contains two development sets (*transfer* and *development*) and an *ensemble* set with the annotations produced by previous tasks. In our particular case, we used as training set the combination of the training, transfer and ensemble, and the development set as the validation set.

5.3.4 Results

The metrics defined by the eHealth-KD challenge to evaluate these experiments are commonly used for some NLP tasks such as NER or text classification, namely precision, recall, and F1-score. Contrary to previous experiments this challenge provides other metrics in an evaluation considering different categories of errors. These metrics can be defined in terms of comparing the response of a system against the golden annotation, i.e. correct, incorrect, partial, spurious, and missing entities [235]:

- Correct (C): matches are reported when a text in the system output file exactly matches a corresponding span of text in the gold file in the start and end positions, and also the entity type.
- Incorrect (I): matches occur when the start and end positions match, but not the type of the entity.
- Partial (P): matches are reported when two intervals (start and end) have a non-empty intersection, such as the case of "*vías respiratorias*" (airways) and "*respiratorias*" in the previous example (and matching the correct category). Notice that a partial phrase will only be matched against a single correct phrase. For example, "*cáncer de mama*" (breast cancer) could be a partial match for both "*cáncer*" and "*mama*", but it is only counted once as a partial match with the word "*cáncer*" while the word "*mama*" is counted as Missing. This aims to discourage a few large text spans that cover most of the document from receiving a very high score.

- Missing (M): matches are those that appear in the gold file but not in the system file.
- Spurious (S): matches are those that appear in the system file but not in the gold file.

From these definitions, the organizers compute precision, recall, and F1 measure as follows:

$$Precision = \frac{C + \frac{1}{2}P}{C + I + P + S} \tag{5.7}$$

$$Recall = \frac{C + \frac{1}{2}P}{C + I + P + M} \tag{5.8}$$

$$F - measure(F1) = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.9}$$

The results obtained by following the proposed methodology and the described word embeddings are shown in Table 5.14. As in the previous scenarios, we have shown the results in a grouped form to better describe them. First, we evaluate each word embedding independently (first section of the table). Second, we propose a combination of word embeddings according to their type (contextual and non-contextual) and by language (Spanish or multilingual). Finally, we show other results of different combinations (third section of the table).

Initially and taking into account the use of word embeddings individually, we obtained results of between 80% and 82% of F1-score. Specifically, the best value of F1 was obtained using the embeddings extracted from the BETO

	Precision (%)	Recall (%)	F1 (%)
Wikipedia	82.87	80.04	81.43
SME	83.21	80.22	81.68
Pooled	82.31	79.95	81.21
BETO	83.93	80.31	82.08
XLM-RoBERTa	82.87	79.59	81.19
mBERT	83.05	78.87	80.9
Wikipedia + SME	83.24	80.4	81.79
Pooled + BETO	83.99	80.22	82.06
XLM-RoBERTa + mBERT	84.06	80.13	82.04
SME + BETO + mBERT	**85.59**	**80.67**	**83.06**

Table 5.14: Performance results for the NER task in health documents using the BiLSTM-CRF approach.

pre-trained model (82.08%). Following BETO, the use of SME obtained values close to F1, reaching 81.68%.

Using traditional embeddings together (Wikipedia + SME) we obtain an 81.79% F1-score, which shows a slight improvement in the use of them independently. By combining the BETO and Pooled embeddings, we do not achieve an improvement because BETO improves individually rather than in combination. On the other hand, the combination of the multilingual models achieves an enhancement when compared to mBERT or XLM-RoBERTa, as it obtains 84.06% precision (the highest precision so far analyzed) and 82.04% of F1.

Other interesting results are shown in the last section of the table. After making several combinations (cf. Appendix A Table A.1), we wanted to show the most successful ones. The combination of SME + BETO + mBERT

embeddings achieves the highest values in all the proposed metrics: 85.59% for precision, 80.67% for recall, and 83.06% for F1. Unlike the previous scenarios, with this corpus, we obtain good results by combining Transformers-based word embeddings with SME.

	Runtime	Embeddings size
Wikipedia	37:41 $\pm$ 1.03	300
SME	37:29 $\pm$ 5.52	300
Pooled	01:12:11 $\pm$ 5.25	8,192
BETO	01:19:15 $\pm$ 18.64	3,072
XLM-RoBERTa	01:11:02 $\pm$ 0.47	768
mBERT	01:26:05 $\pm$ 2.97	3,072
Wikipedia + SME	34:30 $\pm$ 2.04	600
Pooled + BETO	01:50:21 $\pm$ 6.04	11,264
XLM-RoBERTa + mBERT	01:40:07 $\pm$ 9.48	3,840
SME + BETO + mBERT	01:43:23 $\pm$ 11.18	6,444

Table 5.15: Running time results and size of word embeddings using the eHealth-KD corpus.

The time employed in training our approach with each word embedding studied is presented in Table 5.15. We consider this a key point since we see the time spent on each corpus with different word embeddings. Furthermore, we show the size of the vector used for each word in order to determine whether it is a variable that affects the training time. The runtime has been obtained using the same GPU for each experiment and then averaging it (three runs). We also show the SEM in seconds knowing how much those times differ. As in previous experiments, we can observe that the time spent using non-contextual word embeddings is much less than the time spent

with contextual embeddings. We also highlight the large size involved in using Pooled embeddings. However, the use of embeddings obtained from models based on Transformers consumes approximately the same time as Pooled embeddings. Our best result taking into account the F1-score has been achieved by using the combination of SME + BETO + mBERT embeddings and has consumed an average of 1 hour and 43 minutes.

System	Model/word embedding	P (%)	R (%)	F1 (%)
BiLSTM-CRF (Our model)	SME + BETO + mBERT	85.59	80.67	83.06
BERT [236]	BETO	82.16	82.01	82.09
BiLSTM-CNN-CRF [237]	mBERT + medical embedding [111] + SUC embedding [25]	80.72	82.46	81.58
BiLSTM-CRF [238]	mBERT + character embedding + POS tagger embedding	82.03	80.85	81.43

Table 5.16: State-of-the-art results for entity extraction using the eHealth-KD corpus in Spanish. P: Precision, R: Recall, SUC: Spanish Unannotated Corpora.

A comparison of the best publications and results using the eHealth-KD corpus is summarized in Table 5.16. Other studies such as those proposed by Medina and Turmo [237] and Rodríguez-Pérez et al. [238] use similar architectures to our model. Rodríguez-Pérez et al. [238] use a deep learning model (BiLSTM-CRF) concatenating word embeddings including mBERT, character embedding, and Part-Of-Speech tagger embedding. The results of this study obtained an 81.43% F1-score. Medina and Turmo [237] employ a BiLSTM network with a CNN and a CRF layer. On this occasion, the word representation vectors used are different. On the one hand, they use contextual embeddings extracted from the mBERT pre-trained model. They also concatenate medical embeddings trained on Spanish [111]. On the other hand, they added the general domain Spanish Unannotated Corpora (SUC) [25].

With the described methodology, they reach 81.58% F1. Finally, García-Pablos et al. [236] presented an end-to-end deep neural network with pre-trained BERT models as the core for the semantic representation of the input texts using the BETO model, which ranks second after our results with an 82.09% of F1-score.

The eHealtk-KD challenge attracted a significant number of researchers from the NLP community interested in this task. A total of 8 teams participated in this challenge and the average F1-score was 70.17%, so we are satisfied with the methodology used. Specifically, our study finished in the first position and considerably above the average achieved.

5.3.5 Error analysis

An in-depth analysis of results and errors is presented in this section. This allows us to find improvements for our future work. For this purpose, we use the metrics described above: correct, incorrect, partial, spurious, and missing. Other experiments such as those discussed in previous sections (Section 5.1 and 5.2) have a simple classification evaluation that can be measured in terms of TP, TN, FP, and FN, and subsequently compute precision, recall, and F1-score for each named-entity type. These challenges involve discarding partial matches and other scenarios when the NER system evaluates the named-entity surface string correctly but the type wrongly. However, the eHealth-KD share task evaluates these scenarios again at a full-entity level.

A fine-grained evaluation of these systems can be defined in terms of comparing the response of our best system (BiLSTM-CRF using SME + BETO

+ mBERT embeddings) against the golden annotation. Then, the evaluation of our system considering these different categories of errors is shown in Table 5.17.

Entity	Correct	Incorrect	Partial	Spurious	Missing
Concept	313	6	30	10	28
Action	95	9	1	22	16
Predicate	20	7	0	6	22
Reference	5	0	0	0	4
Total	433	22	31	38	70

Table 5.17: Evaluation results obtained by combining word embeddings in terms of comparing the response of the system against the golden annotation in the eHealth-KD corpus.

As we can see, Table 5.17 summarizes the corpus entity types (rows) and the evaluated measures (columns). The results obtained are remarkable thanks to the great number of correct entities that we obtain. For example, our system can match 433 identifying entities of 556 (which means 77.9%): 313 entities are of the Concept category, 95 of the Action type, 20 of the Predicate type, and 5 References. This first row is consistent with Table 5.12 since the more entities contained in the training set, the greater the probability of classifying them correctly. The number of incorrectly identified entities (the type of entity category did not match) is 22, representing a low value (4% of the total).

Moreover, our system has incorrectly annotated partial mentions 31 times, 30 of which were in Concept type entities. 38 entities were recognized by the system but were not in the gold file (spurious), and on the contrary, 70 entities appeared in the gold annotation and were not identified by our system (missing). Some of the errors frequently detected by our system are analyzed below.

In order to carry out an in-depth study of the incorrectly labeled entities, we performed a confusion matrix. Figure 5.16 illustrates the matrix providing the number of errors committed by our approach when compared to the gold annotation. As we can see, the number of entities annotated as Action in the gold standard and labeled by our system with the Concept category is 9 including words like "*trasplante*" (transplant) and "*recuperación*" (recovery). These errors usually occur due to verb conjugation problems. In other words, if they refer to a noun they should be labeled as a Concept, or on the contrary if they refer to an Action or verb labeled as an action. For instance, the word "transplant" appears 14 times in the training set where 10 times it is labeled as Action.

On the contrary, 6 mentions with the Concept label have been classified as Action by our system. Examples of this type are: "*problemas*" (problems), "*dolores*" (pains) and "*anomalías*" (anomalies). Note that 6 mentions of the predicate type have been incorrectly identified by our system, assigning them the Concept category. Some words associated with this type of error are: "*edad*" (age), "*uno*" (one) and "*dos*" (two).

Concerning the spurious and missing errors, we have collected some mislabeled entities as shown in Table 5.18. The proposed method has not been able to identify the examples of entities presented in the table. We would like to highlight the missing Concepts such as LASIK, "*Equeratomileusis in situ asistida con láser*" (laser assisted "in situ" keratomileusis), "*Consejo Institucional de Revisión*" (Institutional Review Board) and "*DE*" (*disfunción eréctil*-Erectile Dysfunction). All of them form names of institutions, organizations,

Gold \ System	Concept	Action	Predicate	Reference
Concept	x	6	0	0
Action	9	x	0	0
Predicate	6	1	x	0
Reference	0	0	0	x

Figure 5.16: Confusion matrix for entities incorrectly classified by our system using the eHealth-KD corpus

or acronyms which the system has not been able to recognize.

Measure	Category	Examples
Spurious	Concept	*Inmunitaria* (Immunity) and *papel* (paper)
	Action	*Evitar* (avoid) and *ayudar* (help)
	Predicate	*Muchos* (many) and *pocos* (few)
Missing	Concept	LASIK (laser assisted “in situ" keratomileusis) and *faciales* (facials)
	Action	*Causar* (cause) and *análisis* (analysis)
	Predicate	*Después* (then) and *varios* (several)

Table 5.18: Examples of errors caused by our systems using the spurious and missing measurements in the eHealth-KD corpus.

Partial matches are produced when the system is not able to correctly identify the start or end position of the mention. Usually, this type of error may produce other spurious and missing errors. Some examples are illustrated in the following figures. The figures include the gold annotation and the annotation obtained by our system. For instance, figure 5.17 shows the gold entity "*menor de edad*" (underage), and as we can see, the system has identified two entities: "*menor*" and "*edad*" which has resulted in two errors. On the one hand, "*menor*" has been identified as a partial error, and on the other hand, the

entity "*edad*" as spurious because they were not in the gold file. The same error with different entity types is presented in Figure 5.18. In this case, the golden entity is "*glucosa-6 fosfato-deshidrogenasa*" (glucose-6-phosphatedehydrogenase) but our system has identified two independent entities and consequently has produced two errors.

```
- Gold annotation:
Entity: menor de edad

- System annotation:
Entity: menor
Error type: partial

- System annotation:
Entity: edad
Error type: spurious
```

Figure 5.17: Example 1. Partial and spurious errors produced by the system against the golden entity in the eHealth-KD corpus. English translation: underage.

```
- Gold annotation:
Entity: glucosa-6 fosfato-deshidrogenasa

- System annotation:
Entity: glucosa-6
Error type: partial

- System annotation:
Entity: fosfato-deshidrogenasa
Error type: spurious
```

Figure 5.18: Example 2. Partial and spurious errors produced by the system against the golden entity in the eHealth-KD corpus. English translation: glucose-6-phosphate dehydrogenase.

Other types of errors caused by the partial entities can produce missing

mentions. Figure 5.19 illustrates an example of this type of error. In this case, the entity annotated in the gold standard is "*ataque de vértigo*" (dizzy spell) but our system has only identified one mention ("*ataque*"). This has meant that the entity found has been classified as partial and a missing entity ("*vértigo*") since the system has not been able to recognize it.

```
- Gold annotation:
Entity: ataque de vértigo

- System annotation:
Entity: ataque
Error type: partial

- Other error:
Entity: vértigo
Error type: missing
```

Figure 5.19: Example of partial and missing errors produced by the system against the golden entity in the eHealth-KD corpus. English translation: dizzy spell.

Accomplishing a manual count of the entities annotated incorrectly as partial, we have found that 85% of them have produced other errors and have supposed an increase of spurious and missing entities. So we conclude that our system must be improved in terms of the start and end positions of an entity, in particular, we have to focus on the Concept type entities since they are the ones that have classified our system the worst.

5.3.6 Discussion

The eHealth-KD challenge, in its third edition, leverages a semantic model of human language that encodes the most common expressions of factual

knowledge, through a set of four types of general-purpose entities. Although this challenge contained several scenarios for the participants, our main goal was to put into practice our approach to the NER task (scenario 2 and task A).

In this section, we have applied the methodology outlined above (cf. Chapter 4) based on neural networks. Specifically it is composed of a BiLSTM with a CRF layer to predict mentions of entities. In this particular case, the entities are diverse: concepts, actions, references, and predictions making the challenge to be considered as the general domain. However, to carry out this task, the documents that compose the corpus are extracted from the MedlinePlus online library containing health information.

In the challenge, a total of eight teams of researchers from different institutions have been involved in the NER task by proposing novel systems. According to the organizers of the eHealth-KD 2020 task, the most significant change from previous ones is the use of contextual embeddings (i.e. Transformer architectures and specifically BERT) as a replacement for traditional word embeddings.

Following the idea of the previous experiments, we evaluated different types of word embeddings taking into account their nature. The results achieved have been satisfactory. We obtain 85.59% precision, 80.67% recall, and 83.06% F1-score by combining multiple word representations. The combination of word embeddings that produced the best results was using SME, BETO, and mBERT. It is important to highlight in this scenario the good performance of the embeddings extracted from BETO, reaching 82.08% F1-score individually. Until now, these embeddings had obtained positive results but

not as close to the state-of-the-art as in the corpus. This is probably due to the general purpose of the entities to be recognized, which in previous experiments was strongly focused on biomedicine and word embeddings such as SME performed better.

A direct comparison with other participants was carried out to see the enhancement produced by our system. Furthermore, the comparison summarized the approaches employed by the authors, thus demonstrating that the novel BERT model offers results as good as the deep learning methods. More specifically, García-Pablos et al. [236] employed the Transformer-based BERT architecture using the BETO model trained on Spanish. They also achieved encouraging results with an 82.09% F1-score. In the final evaluation, we reached the first position in terms of results and therefore the state-of-the-art in the eHealth-KD corpus.

An analysis of results and errors has been conducted in more detail in this chapter discovering possible improvements to our system in future work. With this error analysis we have found the following important elements: *i*) our system does not usually fail to find the category, however, the most confusing entities are Concept and Action; *ii*) finding the start and end position of an entity remains a major challenge, specifically in entities annotated as Concept; *iii*) sometimes, making a mistake in partial matching causes an increase in the number of errors of spurious and missing entities.

Chapter 6

Conclusions

The huge growth in electronically stored biomedical data has made knowledge extraction an important task in this domain. Health documents may include relevant evidence such as findings, diseases, and treatments that can help health professionals in their decision-making. However, this information is difficult to process manually by professionals due to the time and cost involved, so the generation of automatic resources is necessary. One of the goals of NLP is to facilitate these tasks by enabling the use of automated methods that extract knowledge from a text with high validity and reliability. Specifically, the NER task applied to the biomedical domain aims to extract and identify entities of interest that can be used by health professionals.

Biomedical entity recognition is an important task that is still unresolved but can help in other medical-related systems. For example, the NER can identify important findings that are essential for safe and effective health care. Moreover, this task can be applied to other NLP tasks such as text classification by serving as a support point. Finally, NER can be applied to several sub-domains of healthcare, such as oncology, recognizing cancer-related findings,

and pharmacology, identifying drugs, medicines, or adverse events.

Researchers have invested significant effort into developing NLP methods and tools for the NER task from narrative clinical texts. Several architectures of machine learning approaches have been applied to addressing this task. With the emergence of deep learning, models can use feature representations (i.e. word embeddings) of large volumes of unlabeled clinical text. A word embedding is a numerical vector for representing words in a text and it allows us to leverage knowledge about language semantics more precisely. In addition, many pre-trained word embeddings are publicly available, such as the GloVe, fastText, and contextualized word embeddings based on Transformers models.

The advances in deep learning and the representation of words through word embeddings have motivated us to apply them in the NER task for the Spanish biomedical domain. Most of the research on the NER task has been conducted in English. Therefore, this thesis aims to advance the study of entity recognition in Spanish, the second most spoken language in the world and the third most used on the Internet[1].

This thesis is focused on the use of a sophisticated neural network architecture based on BiLSTM with a CRF layer to predict each word as an appropriate entity (or non-entity). The BiLSTM network is composed of two LSTM networks that read sentences from right to left and vice versa. This ensures that it is able to understand the context of each sentence. In addition,

[1]Spanish language: `https://en.wikipedia.org/wiki/Spanish_language`

the first layer of the neural network consists of a combination of word embeddings based on vector concatenation improving the final classification of each word. For this purpose, we have used several word embeddings available for Spanish. Among those used, we highlight traditional word embeddings such as fastText trained on Spanish Wikipedia and contextual word embeddings such as pooled contextualized embeddings and BETO. Moreover, we have generated new ones specific to the language and domain in order to be able to better understand the clinical language.

To test our proposed model, we have used three corpora available thanks to the challenges proposed in different national and international conferences. Specifically, we used the datasets included in the PharmaCoNER [209], Cantemist [30] and eHealth-KD [232] challenges. Since the workshops provide the results of the participants as well as their approximations, we have been able to compare our systems with the state-of-the-art results showing the strengths of our approach.

6.1 Main contributions

This research has carried out a series of studies, analyses, and development of NLP techniques designed to address the task of NER in Spanish biomedical texts. This has resulted in several contributions to the research that we have considered on the basis of the hypotheses outlined in Section 1.3.

To support hypothesis H1, we can summarize the following contributions:

(H1). Deep neural networks in NLP leverage the advantage of existing relevant information from the Spanish biomedical textual data and the NER task, outperforming models that do not integrate this information properly.

- We have investigated and implemented different machine learning approaches as shown in Chapter 2. First, we have reviewed unsupervised models and then advanced to supervised models using traditional models such as CRF and deep neural networks.

- In our review of the state-of-the-art in deep learning, we have exposed what kind of architectures are used by the scientific community interested in NER (Section 3.2).

- We have proposed a model based on neural networks. Specifically, the architecture is composed of a BiLSTM network and a CRF layer (Section 4.3).

To support hypothesis H2, we provide the following contributions:

(H2). Combining different types of word embeddings by concatenating each embedding vector to form the final word vectors is an important part of the biomedical entity recognition task. The probability of recognizing a specific entity in a text should increase as optimal representations of that word are combined, because they are more comprehensively represented and integrate more knowledge.

- In our review of related literature we have found that word representations and, more specifically, word embeddings are the most commonly used methods (Section 3.3).

- We have selected different word embeddings to include in the neural network to address the NER problem in biomedicine (Section 4.2).

- We have presented a model based on a combination of word embeddings for a more exhaustive representation of the words, thus improving entity identification systems (Section 4.3).

The contributions that support hypothesis H3 can be summarized as follows:

(H3). Integrating domain-specific knowledge into the training corpus can be beneficial for improving the quality of word embeddings. Thus, this resource provides a more accurate representation of words in a particular context and domain.

- We have collected an unannotated corpus by extracting documents from different corpora and websites related to the biomedical domain, obtaining a vocabulary of 1,704,151 words (Section 4.2.1).

- We have generated new word embeddings specifically for the biomedical domain in Spanish (Section 4.2.1).

Finally, after describing the specific contributions for each hypothesis, the global contributions resulting from this study can be summarized as follows:

- We have developed a system based on machine learning that obtains substantial improvements in the NER task using NLP techniques applied to the biomedical domain in Spanish.

- We have implemented a model based on deep networks and a combination of word embedding in three biomedical scenarios: pharmacology (Section 5.1), oncology (Section 5.2), and knowledge discovery (Section 5.3).

- For each scenario, we have described the problem to be solved and the corpus to be used. We have presented the results achieved with each combination of embeddings, the execution time, and the size of the combination vector. Furthermore, we have been able to compare our system with state-of-the-art results and conducted an error analysis. Finally, we have exposed a discussion and thus compiled the main findings of that scenario.

- We have obtained performance improvements over the previous state-of-the-art using the proposed model.

6.2 Future work

For future work, we will study the performance of using more linguistic features as an input in the neural network because we believe it would add extra knowledge to the network by indicating how the word functions in meaning as well as grammatically within the sentence. A morphological and syntactic analysis will be performed on the biomedical reports to determine, on the one hand, the form, class, or grammatical category of each word and, on the other hand, the concordance and hierarchy relations that the words have with each other.

Considering the corpus with which the experiments are conducted, other annotation schemes will be used to solve the following problems:

1. Entities inside other entities, for instance in the sentence "*Se biopsian nuevamente informándose como linfoma cutáneo primario T tipo micosis fungoide*" (The biopsies are repeated and reported as primary cutaneous T-lymphoma mycosis fungoides) there are two annotated entities "*linfoma cutáneo primario T tipo micosis fungoide*" (primary cutaneous T-lymphoma, mycosis fungoides type) and "micosis fungoide" (mycosis fungoides), so there are words that are in two different entities.

2. Discontinuous entities in corpora, for example, in the sentence "*Los análisis de sangre y orina son la única manera de saber si usted tiene enfermedad renal*" (Blood and urine tests are the only way to know if you have kidney disease) the annotated entities are "*análisis de sangre*" (blood tests) (an entity with consecutive words) and "*análisis de orina*" (urine tests) (discontinuous entity).

Since our study has been conducted for a specific language, in future work, we plan to extend it to different languages such as English in order to see the usefulness of the different word embeddings trained for languages other than Spanish.

Finally, there are pre-trained models available for the biomedical domain such as BioBERT and ClinicalBERT that could be taken into consideration. Although all of them are in English, we will work on generating a new model for Spanish, taking into account the computational resources involved.

6.3 Publications

During the course of this thesis, work has been carried out in different workshops and tasks, some directly related to the thesis topic and others close to the task that have also served to add value to the research. As a result of this work, the following publications in high impact journals and conferences have been produced (in chronological order):

6.3.1 Journals

1. Andreu-Marin, A., Martínez-Santiago, F., Ureña-López, L. A., & **López-Úbeda, P.**. (2017). El lenguaje del pensamiento. Ciencia Cognitiva, 11:3, 50-52.

2. **López-Úbeda, P.**, Díaz-Galiano, M. C., Montejo-Ráez, A., Martínez-Santiago, F., Andreu-Marin, A., Martín-Valdivia, M. T & Ureña-López, L. A. (2018). Biomedical Semantic Information Retrieval. Procesamiento del Lenguaje Natural, (61), 189-192.

   Impact source: SCImago Journal Rankings (SJR): 0.21. Impact factor: Q2.

3. **López-Úbeda, P.**, Díaz-Galiano, M. C., Martín-Noguerol, T., Ureña-López, A., Martín-Valdivia, M. T., & Luna, A. (2020). Detection of unexpected findings in radiology reports: A comparative study of machine learning approaches. Expert Systems with Applications, 160, 113647.

   Impact source: WOS (JCR). Impact factor: Q1.

4. **López-Úbeda, P.**, Díaz-Galiano, M. C., Montejo-Ráez, A., Martín-Valdivia,

M. T., & Ureña-López, L. A. (2020). An Integrated Approach to Biomedical Term Identification Systems. Applied Sciences, 10(5), 1726.

Impact source: WOS (JCR). Impact factor: Q2.

5. **López-Úbeda, P.**, Díaz-Galiano, M. C., Martín-Noguerol, T., Luna, A., Ureña-López, L. A., & Martín-Valdivia, M. T. (2020). COVID-19 detection in radiological text reports integrating entity recognition. Computers in Biology and Medicine, 127, 104066.

   Impact source: WOS (JCR). Impact factor: Q2.

6. **López-Úbeda, P.**, Díaz-Galiano, M. C., Martín-Noguerol, T., Luna, A., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Automatic medical protocol classification using machine learning approaches. Computer Methods and Programs in Biomedicine, 105939.

   Impact source: WOS (JCR). Impact factor: Q1.

7. **López-Úbeda, P.**, Plaza del Arco, F. M. P., Díaz-Galiano, M. C., & Martín-Valdivia, M. T. (2021). How Successful is Transfer Learning for Detecting Anorexia on Social Media? Applied Sciences, 11, 1838.

   Impact source: WOS (JCR). Impact factor: Q2.

8. **López-Úbeda, P.**, Plaza del Arco, F. M. P., Díaz-Galiano, M. C., & Martín-Valdivia, M. T. (2021). NECOS: An annotated corpus to identify constructive news comments in Spanish. Procesamiento del Lenguaje Natural, (66).

Impact source: SCImago Journal Rankings (SJR): 0.21. Impact factor: Q2. Accepted for publication.

9. **López-Úbeda, P.**, Pomares-Quimbaya, A., Díaz-Galiano, M. C., & Schulz, S. (2021). Collecting specialty-related medical terms: Development and evaluation of a resource for Spanish. BMC Medical Informatics and Decision Making.

   Impact source: WOS (JCR). Impact factor: Q3. Accepted for publication.

10. **López-Úbeda, P.**, Díaz-Galiano, M. C., Ureña-López, L. A., & Martín-Valdivia, M. T., (2021). Combining word embeddings to extract chemical and drug entities in biomedical literature. BMC Bioinformatics.

    Impact source: WOS (JCR). Impact factor: Q1. Under review.

11. **López-Úbeda, P.**, Díaz-Galiano, M. C., Ureña-López, L. A., & Martín-Valdivia, M. T., (2021). Integrating hybrid word embeddings for deep learning in biomedical entity recognition. Expert Systems with Applications.

    Impact source: WOS (JCR). Impact factor: Q1. Under review.

6.3.2 Conferences

1. **López-Úbeda, P.**. (2018). Integración de Conocimiento para la Mejora de Sistemas de Recuperación de Información. In Proceedings of the Doctoral Symposium of the XXXIV International Conference of the SEPLN, pp. 31–36.

2. **López-Úbeda, P.**, Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Ureña-López, L. A. (2018). Machine Learning to Detect ICD10 Codes in Causes of Death. In CLEF (Working Notes).

3. **López-Úbeda, P.**, Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Ureña-López, L.A. (2018). Filtering and reranking using MetaMap named entities recognizer. In TREC.

4. **López-Úbeda, P.**, Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Ureña-López, L.A. (2018). Using clustering to filter results of an Information Retrieval system. In TREC.

5. **López-Úbeda, P.**, Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Ureña-López, L. A. (2018). Sinai en tass 2018 task 3. clasificando acciones y conceptos con umls en medline. Proceedings of TASS, 2172.

6. **López-Úbeda, P.**, Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Jiménez-Zafra, S. M. (2018). SINAI at DIANN-IberEval 2018. Annotating Disabilities in Multi-language Systems with UMLS. In IberEval@ SEPLN (pp. 37-43).

7. Díaz-Galiano, M. C., **López-Úbeda, P.**, Martín-Valdivia, M. T., & Ureña-López, L.A. (2018). SINAI at CLEF eHealth 2018 Task 3. Using cTAKES to Remove Noise from Expanding Queries with Google. In CLEF (Working Notes).

8. **López-Úbeda, P.**. (2019). Reconocimiento de Entidades Nombradas en español aplicado al dominio biomédico. In Proceedings of the Doctoral

Symposium of the XXXV International Conference of the SEPLN, pp. 20-25.

9. **López-Úbeda, P.**, Díaz-Galiano, M. C., Ureña-López, L.A., & Martín-Valdivia, M. T. (2019). Anonymization of Clinical Reports in Spanish: a Hybrid Method Based on Machine Learning and Rules. In IberLEF@ SEPLN (pp. 687-695).

10. **López-Úbeda, P.**, Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Ureña-López, L. A. (2019, August). Using machine learning and deep learning methods to find mentions of adverse drug reactions in social media. In Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task (pp. 102-106).

11. **López-Úbeda, P.**, Díaz-Galiano, M. C., Ureña-López, L. A., & Martín-Valdivia, M. T. (2019, November). Using Snomed to recognize and index chemical and drug mentions. In Proceedings of The 5th Workshop on BioNLP Open Shared Tasks (pp. 115-120).

12. **López-Úbeda, P.**, Vera-Ramos, J. A., & López-García, P. (2019). TREC 2019 Precision Medicine-Medical University of Graz. In TREC(Proceedings).

13. Plaza-del-Arco, F. M., **López-Úbeda, P.**, Díaz-Galiano, M. C., Ureña-López, L. A., & Martín-Valdivia, M. T. (2019). Integrating UMLS for Early Detection of Signs of Anorexia. In CLEF (Working Notes).

14. **López-Úbeda, P.**, Plaza del Arco, F. M. P., Díaz-Galiano, M. C., Ureña-López, L. A., & Martín-Valdivia, M. T. (2019, September). Detecting

Anorexia in Spanish Tweets. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019) (pp. 655-663).

15. Pomares-Quimbaya, A., **López-Úbeda, P.**, Oleynik, M., & Schulz, S. (2020). Leveraging PubMed to Create a Specialty-Based Sense Inventory for Spanish Acronym Resolution. Studies in health technology and informatics, 270, 292-296.

16. **López-Úbeda, P.**, Díaz-Galiano, M. C., Ureña-López, L. A., Martín-Valdivia, M. T., Martín-Noguerol, T., & Luna, A. (2020, May). Transfer learning applied to text classification in Spanish radiological reports. In Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultiligualBIO 2020) (pp. 29-32).

17. **López-Úbeda, P.**. (2020). Reconocimiento de entidades biomédicas para el español mediante la combinación de word embeddings. In Proceedings of the Doctoral Symposium on Natural Language Processing from the PLN.net network (PLNnet-DS-2020), pp. 51-57.

18. **López-Úbeda, P.**, Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Ureña-López, L. A. (2020). Extracting neoplasms morphology mentions in spanish clinical cases through word embeddings. Proceedings of IberLEF.

19. **López-Úbeda, P.**, Perea-Ortega, J. M., Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Ureña-López, L. A. (2020). SINAI at eHealth-KD

Challenge 2020: Combining Word Embeddings for Named Entity Recognition in Spanish Medical Records.

20. Pomares-Quimbaya, A., **López-Úbeda, P.**, & Schulz, S. (2021). Transfer learning for classifying Spanish and English text by clinical specialties. Studies in Health Technology and Informatics. Accepted for publication.

6.4 Research collaborations

6.4.1 Participation in projects

- **REDES project** (Reconocimiento de Entidades Digitales: Enriquecimiento y Seguimiento mediante Tecnologías del Lenguaje) with reference TIN2015-65136-C2, is funded by the Spanish Government through the program National Programme for Research Aimed at the Challenges of Society (Projects I+D+i 2015) of the Ministry of Economy, Industry and Competitiveness.

- **REC project** (Reconocimiento de Entidades Clínicas para mejorar el registro estandarizado de pacientes en el Servicio de Geriatría) with reference 2020/042, is partially funded by the Hospital Universitario San Ignacio (Bogotá, Colombia), Pontificia Universidad Javeriana (Bogotá, Colombia) and University of Jaén (Spain).

- **LIVING-LANG project** (Tecnologías del lenguaje humano para entidades digitales vivas) with reference RTI2018-094653-B-C21, is funded by the Spanish Government through the Knowledge Generation R&D

Projects and R&D&I Projects Research Challenges of the Ministry of Science, Innovation and Universities.

6.4.2 Organising committee

- Organising Committee of the 36th Conference of the Spanish Society for Natural Language Processing[2].
- Organising Committee of the 3rd edition of the PLN.net award[3] for the best new line of research in Natural Language Processing.

6.4.3 Research stays

- Research stay of three months at the University of Graz (Austria). Supervisor: Stefan Schulz. Position held: full professor at the Institute for Medical Informatics, Statistics and Documentation.
- One-week research stay at the University of Wolverhampton (United Kingdom). Supervisor: Ruslan Mitkov. Position held: Director of Research Institute of Information and Language Processing (RIILP).

6.5 Research awards and recognitions

1. **Buscador semántico biomédico.**

   Description: Finalist at "*II Hackathon de Tecnologías del Lenguaje*" in the modality "Biomedicine" within Four Years From Now (4YFN) of the Mobile World Congress (MWC).

[2]SEPLN 2020: `http://sepln2020.sepln.org/`
[3]PLN.net award: `https://gplsi.dlsi.ua.es/pln/node/60`

Granting agency: Red.es, in collaboration with *Secretaría de Estado para la Sociedad de la Información y la Agenda Digital* (SESIAD).

Date: February 26, 2018.

2. **Monitor de dispersión geográfica de enfermedades.**

   Description: Second prize at "*II Hackathon de Tecnologías del Lenguaje*" in the modality "General corpus" within Four Years From Now (4YFN) of the Mobile World Congress (MWC).

   Granting agency: Red.es, in collaboration with *Secretaría de Estado para la Sociedad de la Información y la Agenda Digital* (SESIAD).

   Date: February 26, 2018.

3. **Extracting Neoplasms Morphology Mentions in Spanish Clinical Cases through Word Embeddings.**

   Description: Third prize at "CANTEMIST: CANcer TExt Mining Shared Task" in the NER sub-task. IberLEF 2020 evaluation campaign at the SEPLN 2020.

   Granting agency: *Oficina Técnica de Sanidad* of the *Plan de Tecnologías del Lenguaje* (Plan TL).

   Date: September 22, 2020.

4. **SINAI at eHealth-KD Challenge 2020: Combining Word Embeddings for Named Entity Recognition in Spanish Medical Records.**

   Description: First prize at "eHealth-KD 2020: eHealth Knowledge Discovery" in the sub-task A (NER). IberLEF 2020 evaluation campaign at

the SEPLN 2020.

Granting agency: eHealth Knowledge Discovery 2020.

Date: September 22, 2020.

5. **Entity extraction in Spanish applied to the biomedical domain.**

   Description: Research award to the best work of research initiation at "*VI Premios Ada Lovelace de Tecnologías de la Información y la Comunicación*".

   Granting agency: *Centro de Estudios Avanzados en Tecnologías de la Información y la Comunicación* (CEATIC), Universidad de Jaén.

   Date: December 2, 2020.

6. **Innovation and research award.**

   Description: Award for innovation and research at the 7th HT médica corporate convention.

   Granting agency: HT médica.

   Date: March 6, 2021.

6.6 Transfer of research results

The transfer of research results is the process of promoting and transferring knowledge of all resources, methods and techniques obtained. In addition, technology transfer also aims to disseminate the studies carried out.

Moreover, the knowledge created in public research institutions is an important input for innovation in various sectors. Universities have become an important mechanism for generating technological innovations capable

of improving society's quality of life. For this reason, universities have come to play a more proactive role in innovation systems, seeking ways to interact with the productive sector to promote technological development that can be used in different areas.

Considering the benefits of research transfer, we have developed and implemented a number of open access applications as detailed below:

- **Detection of unexpected findings**

  As a result of the previous research study referenced in Section 6.3.1 item 3 [8], an API has been generated that uses a deep learning model based on CNN. This API is responsible for receiving anonymized radiological reports in Spanish for the detection of unexpected findings. Unexpected findings are the set of radiological signs identified in a given imaging modality examination that have two characteristics: they are apparently unrelated to the a priori expected results of the radiological examination and they imply an emergency or urgent clinical situation that must be promptly communicated to the prescribing physician or other medical specialist, as well as to the patient, in order to preserve life and/or prevent dangerous events.

  This API is available at `https://sinai.ujaen.es/crifis/`, and is currently being used by the *HT médica* clinic in Jaén as a decision-making support to save time in the detection of unexpected findings. So far, this API has received 67,932 requests from health specialists.

  Given the impact of the study and the applicability of the API, it is currently in the process of registering software with the aim of making

the research results generated more valuable.

- **Automatic assignment of procedure protocols**

  The assignment of procedure protocols in medical imaging requires extensive knowledge of the biomedical text. Protocol assignment is necessary prior to the acquisition of the radiological study, determining the procedure for each patient. The automation of this process is carried out by an API that we have developed. This API has been the result of previous work referenced in Section 6.3.1 item 6 [239] and in Section 6.3.2 16 [6]. This process of protocol assignment could improve the efficiency of patient diagnosis by performing it automatically, saving time for radiology specialists. The API has been developed in order to provide the expert decision support system for procedural protocols.

  The API is based on ML approaches to several radiological techniques: Magnetic Resonance Imaging (MRI), CT and ultrasound scans. It is also freely available and is currently being evaluated by the *HT médica* clinic in Jaén at `https://sinai.ujaen.es/protocolos/`. This system has been used 33,298 times by the clinic to obtain the appropriate protocol. Moreover, this development is currently in the process of software registration and under study for a possible U.S. patent.

- **COVID-19 detection**

  Another study derived from this thesis and implemented for knowledge transfer is a system for the detection of suspected positive cases of COVID-19. Given the current importance of the disease, we have

conducted a study referenced in Section 6.3.1 item 5 [16] in which we propose an automatic detection system for COVID-19 suspicions in radiological reports in Spanish using ML approaches. In addition, to improve the approach, a NER system was used to extract COVID-19 related disorders and include them as additional information to the algorithm. For this purpose, we use the SNOMED-CT vocabulary in its most current version which includes concepts related to SARS-CoV-2.

This API is currently being used, evaluated and validated by radiological experts from the *HT médica* clinic in Jaén. To date, it has been used 883 times by experts. Finally, this system is available at `https://sinai.ujaen.es/covid/`.

Bibliography

[1] Carol Friedman and Stephen B Johnson. "Natural language and text processing in biomedicine". In: *Biomedical Informatics*. Springer, 2006, pp. 312–343.

[2] Gobinda G Chowdhury. "Natural language processing". In: *Annual review of information science and technology* 37.1 (2003), pp. 51–89.

[3] Ralph Grishman and Beth M Sundheim. "Message understanding conference-6: A brief history". In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. 1996.

[4] Aaron M Cohen and William R Hersh. "A survey of current work in biomedical text mining". In: *Briefings in bioinformatics* 6.1 (2005), pp. 57–71.

[5] Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. "What can natural language processing do for clinical decision support?" In: *Journal of biomedical informatics* 42.5 (2009), pp. 760–772.

[6] Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, Teodoro Martín-Noguerol, Antonio Luna, L Alfonso Ureña-López, and María-Teresa Martín-Valdivia. "Automatic medical protocol classification using machine learning approaches". In: *Computer Methods and Programs in Biomedicine* (), p. 105939.

[7] Flor Miriam Plaza-del Arco, Pilar López-Úbeda, Manuel Carlos Dıaz-Galiano, L Alfonso Ureña-López, and María-Teresa Martın-Valdivia. "Integrating UMLS for Early Detection of Sings of Anorexia". In: (2019).

[8] Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, Teodoro Martín-Noguerol, Alfonso Ureña-López, María-Teresa Martín-Valdivia, and Antonio Luna. "Detection of unexpected findings in radiology reports: A comparative study of machine learning approaches". In: *Expert Systems with Applications* 160 (2020), p. 113647.

[9] 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*. 2019, pp. 1–10.

[10] Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. "CHEMDNER: The drugs and chemical names extraction challenge". In: *Journal of cheminformatics* 7.1 (2015), S1.

[11] A Miranda-Escalada, E Farré, and M Krallinger. "Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*. 2020.

[12] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. "2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records". In: *Journal of the American Medical Informatics Association* 27.1 (2020), pp. 3–12.

[13] Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez. "Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018". In: *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*. 2018, pp. 13–16.

[14] Alejandro Piad-Morffis, Yoan Gutiérrez, Hian Cañizares-Diaz, Suilan Estevez-Velarde, Rafael Muñoz, Andres Montoyo, Yudivian Almeida-Cruz, et al. "Overview of the ehealth knowledge discovery challenge at iberlef 2020". In: CEUR, 2020.

[15] Isabel Segura Bedmar, Paloma Martínez, and María Herrero Zazo. "Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)". In: Association for Computational Linguistics. 2013.

[16] Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, Teodoro Martín-Noguerol, Antonio Luna, L Alfonso Ureña-López, and M Teresa Martín-Valdivia. "COVID-19 detection in radiological text reports integrating entity recognition". In: *Computers in Biology and Medicine* 127 (2020), p. 104066.

[17] Pilar López-Ubeda, Manuel Carlos Díaz-Galiano, María Teresa Martín Valdivia, and Luis Alfonso Ureña López. "Filtering and reranking using MetaMap named entities recognizer." In: *TREC*. 2018.

[18] Manuel Carlos Díaz-Galiano, Pilar López-Úbeda, María-Teresa Martín-Valdivia, and L. Alfonso Ureña López. "SINAI at CLEF eHealth 2018 Task 3. Using cTAKES to Remove Noise from Expanding Queries with Google." In: *CLEF (Working Notes)*. 2018.

[19] John Lafferty, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In: (2001).

[20] Hanna M Wallach. "Conditional random fields: An introduction". In: *Technical Reports (CIS)* (2004), p. 22.

[21] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. Vol. 1. 2. MIT press Cambridge, 2016.

[22] John M Giorgi and Gary D Bader. "Towards reliable named entity recognition in the biomedical domain". In: *Bioinformatics* 36.1 (2020), pp. 280–286.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[24] 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, pp. 1532–1543.

[25] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching word vectors with subword information". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[27] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. "Unsupervised cross-lingual representation learning at scale". In: *arXiv preprint arXiv:1911.02116* (2019).

[28] John McCarthy. "What is artificial intelligence?" In: (1998).

[29] Amir Enshaei, CN Robson, and RJ Edmondson. "Artificial intelligence systems as prognostic and predictive tools in ovarian cancer". In: *Annals of surgical oncology* 22.12 (2015), pp. 3970–3975.

[30] A Miranda-Escalada, E Farré, and M Krallinger. "Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*. 2020.

[31] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. "Artificial intelligence in radiology". In: *Nature Reviews Cancer* 18.8 (2018), pp. 500–510.

[32] Eun-Jae Lee, Yong-Hwan Kim, Namkug Kim, and Dong-Wha Kang. "Deep into the brain: artificial intelligence in stroke imaging". In: *Journal of stroke* 19.3 (2017), p. 277.

[33] SA Senthilkumar, Bharatendara K Rai, Amruta A Meshram, Angappa Gunasekaran, and S Chandrakumarmangalam. "Big data in healthcare management: a review of literature". In: *American Journal of Theoretical and Applied Business* 4.2 (2018), pp. 57–69.

[34] Tom Michael Mitchell. *The discipline of machine learning*. Vol. 9. Carnegie Mellon University, School of Computer Science, Machine Learning . . ., 2006.

[35] Tom Mitchell. "Introduction to machine learning". In: *Machine Learning* 7 (1997), pp. 2–5.

[36] Alain Auger and Caroline Barrière. "Pattern-based approaches to semantic relation extraction: A state-of-the-art". In: *Terminology* 14.1 (2008), p. 1.

[37] Sergei Egorov, Anton Yuryev, and Nikolai Daraselia. "A simple and practical dictionary-based approach for identification of proteins in MEDLINE abstracts". In: *Journal of the American Medical Informatics Association* 11.3 (2004), pp. 174–178.

[38] Irina Rish et al. "An empirical study of the naive Bayes classifier". In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. 22. 2001, pp. 41–46.

[39] Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. "Tackling the poor assumptions of naive bayes text classifiers". In: *Proceedings of the 20th international conference on machine learning (ICML-03)*. 2003, pp. 616–623.

[40] Joseph M Hilbe. *Logistic regression models*. Chapman and hall/CRC, 2009.

[41] Steven C Bagley, Halbert White, and Beatrice A Golomb. "Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain". In: *Journal of clinical epidemiology* 54.10 (2001), pp. 979–985.

[42] Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. "Maximum Entropy Markov Models for Information Extraction and Segmentation." In: *Icml*. Vol. 17. 2000. 2000, pp. 591–598.

[43] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 144–152.

[44] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[45] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.

[46] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks". In: *Neural networks* 4.2 (1991), pp. 251–257.

[47] Russell Reed and Robert J MarksII. *Neural smithing: supervised learning in feedforward artificial neural networks*. Mit Press, 1999.

[48] Eric Jang, Shixiang Gu, and Ben Poole. "Categorical reparameterization with gumbel-softmax". In: *arXiv preprint arXiv:1611.01144* (2016).

[49] Sebastian Ruder. "An overview of gradient descent optimization algorithms". In: *arXiv preprint arXiv:1609.04747* (2016).

[50] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." In: *Journal of machine learning research* 12.7 (2011).

[51] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[52] Matthew D Zeiler. "Adadelta: an adaptive learning rate method". In: *arXiv preprint arXiv:1212.5701* (2012).

[53] Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. "Attention-over-attention neural networks for reading comprehension". In: *arXiv preprint arXiv:1607.04423* (2016).

[54] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

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

[56] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078* (2014).

[57] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[58] Yoon Kim. "Convolutional neural networks for sentence classification". In: *arXiv preprint arXiv:1408.5882* (2014).

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[60] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[61] Guillaume Lample and Alexis Conneau. *Cross-lingual Language Model Pretraining*. 2019. arXiv: `1901.07291 [cs.CL]`.

[62] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. "Pre-trained models for natural language processing: A survey". In: *arXiv preprint arXiv:2003.08271* (2020).

[63] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[64] Iz Beltagy, Kyle Lo, and Arman Cohan. "SciBERT: A pretrained language model for scientific text". In: *arXiv preprint arXiv:1903.10676* (2019).

[65] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. "Spanish Pre-Trained BERT Model and Evaluation Data". In: *to appear in PML4DC at ICLR 2020*. 2020.

[66] Andre Lamurias and Francisco M Couto. "Text mining for bioinformatics using biomedical literature". In: *Encyclopedia of bioinformatics and computational biology* 1 (2019), pp. 602–611.

[67] Andrew S Wu, Bao H Do, Jinsuh Kim, and Daniel L Rubin. "Evaluation of negation and uncertainty detection and its impact on precision and recall in search". In: *Journal of digital imaging* 24.2 (2011), pp. 234–242.

[68] Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. "The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes". In: *BMC bioinformatics* 9.11 (2008), pp. 1–9.

[69] Gayle A Sulik. "Managing biomedical uncertainty: the technoscientific illness identity". In: *Sociology of Health & Illness* 31.7 (2009), pp. 1059–1076.

[70] BE Jones, BR South, Y Shao, CC Lu, J Leng, BC Sauer, AV Gundlapalli, MH Samore, and Q Zeng. "Development and validation of a natural language processing tool to identify patients treated for pneumonia across VA emergency departments". In: *Applied clinical informatics* 9.1 (2018), p. 122.

[71] Byron C Wallace. "Automating biomedical evidence synthesis: Recent work and directions forward". In: *BIRNDL@ SIGIR*. 2018.

[72] Nathan Peiffer-Smadja, Timothy Miles Rawson, Raheelah Ahmad, Albert Buchard, Georgiou Pantelis, F-X Lescure, Gabriel Birgand, and Alison Helen Holmes. "Machine learning for clinical decision support in infectious diseases: a narrative review of current applications". In: *Clinical Microbiology and Infection* (2019).

[73] Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U Rajendra Acharya. "Automated detection of COVID-19 cases using deep neural networks with X-ray images". In: *Computers in biology and medicine* 121 (2020), p. 103792.

[74] Ioannis D Apostolopoulos and Tzani A Mpesiana. "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks". In: *Physical and Engineering Sciences in Medicine* 43.2 (2020), pp. 635–640.

[75] Yaseen M Arabi, Srinivas Murthy, and Steve Webb. "COVID-19: a novel coronavirus and a novel challenge for critical care". In: *Intensive care medicine* (2020), pp. 1–4.

[76] Miguel Miranda, Júlio Duarte, António Abelha, José Manuel Machado, and José Neves. "Interoperability and healthcare". In: (2009).

[77] Olaronke Iroju, Abimbola Soriyan, Ishaya Gambo, and Janet Olaleke. "Interoperability in healthcare: benefits, challenges and resolutions". In: *International Journal of Innovation and Applied Studies* 3.1 (2013), pp. 262–270.

[78] Frank Oemig and Bernd Blobel. "Natural language processing supporting interoperability in healthcare". In: *Text mining*. Springer, 2014, pp. 137–156.

[79] Fred Freitas, Stefan Schulz, and Eduardo Moraes. "Survey of current terminologies and ontologies in biology and medicine". In: *RECIIS-Electronic Journal in Communication, Information and Innovation in Health* 3.1 (2009), pp. 7–18.

[80] Hermenegildo Fabregat, Juan Martinez-Romo, and Lourdes Araujo. "Overview of the DIANN Task: Disability Annotation Task." In: *IberEval@ SEPLN*. 2018, pp. 1–14.

[81] Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. "Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*. 2020.

[82] Pilar López Úbeda, Manuel Carlos Díaz-Galiano, L Alfonso Ureña López, and M. Teresa Martín-Valdivia. "Using Snomed to recognize and index chemical and drug mentions." In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. 2019, pp. 115–120.

[83] Zhenfei Ju, Jian Wang, and Fei Zhu. "Named entity recognition from biomedical text using SVM". In: *2011 5th international conference on bioinformatics and biomedical engineering*. IEEE. 2011, pp. 1–4.

[84] Takaki Makino, Yoshihiro Ohta, Jun'ichi Tsujii, et al. "Tuning support vector machines for biomedical named entity recognition". In: *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*. 2002, pp. 1–8.

[85] Koichi Takeuchi and Nigel Collier. "Bio-medical entity extraction using support vector machines". In: *Artificial Intelligence in Medicine* 33.2 (2005), pp. 125–137.

[86] SL Li and YK Guo. "Biomedical Named Entity Recognition with CNN-BLSTM-CRF [J]". In: *Journal of chinese information processing* 32.1 (2018), pp. 116–122.

[87] Jerry Chun-Wei Lin, Yinan Shao, Youcef Djenouri, and Unil Yun. "ASRNN: a recurrent neural network with an attention model for sequence labeling". In: *Knowledge-Based Systems* 212 (2020), p. 106548.

[88] Jingchi Jiang, Huanzheng Wang, Jing Xie, Xitong Guo, Yi Guan, and Qiubin Yu. "Medical knowledge embedding based on recursive neural network for multi-disease diagnosis". In: *Artificial Intelligence in Medicine* 103 (2020), p. 101772.

[89] Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. "A survey of word embeddings for clinical text". In: *Journal of Biomedical Informatics: X* 4 (2019), p. 100057.

[90] Zhiheng Huang, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging". In: *arXiv preprint arXiv:1508.01991* (2015).

[91] SK Hong and Jae-Gil Lee. "DTranNER: biomedical named entity recognition with deep learning-based label-label transition model". In: *BMC bioinformatics* 21.1 (2020), p. 53.

[92] Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. "Overview of BioCreative II gene mention recognition". In: *Genome biology* 9.2 (2008), pp. 1–19.

[93] Jiao Li, Yueping Sun, R Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. "Annotating chemicals, diseases, and their interactions in biomedical literature". In: *Proceedings of the fifth BioCreative challenge evaluation workshop*. The Fifth BioCreative Organizing Committee. 2015, pp. 173–182.

[94] Katikapalli Subramanyam Kalyan and Sivanesan Sangeetha. "BertMCN: Mapping colloquial phrases to standard medical concepts using BERT and Highway Network". In: *Artificial Intelligence in Medicine* (2020), p. 102008.

[95] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. "Publicly available clinical BERT embeddings". In: *arXiv preprint arXiv:1904.03323* (2019).

[96] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. "Domain-specific language model pretraining for biomedical natural language processing". In: *arXiv preprint arXiv:2007.15779* (2020).

[97] Veysel Kocaman and David Talby. "Biomedical Named Entity Recognition at Scale". In: *arXiv preprint arXiv:2011.06315* (2020).

[98] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. *SpanBERT: Improving Pre-training by Representing and Predicting Spans*. 2020. arXiv: `1907.10529 [cs.CL]`.

[99] Karen Kukich. "Techniques for automatically correcting words in text". In: *Acm Computing Surveys (CSUR)* 24.4 (1992), pp. 377–439.

[100] Antonio Jimeno, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga, and Dietrich Rebholz-Schuhmann. "Assessment of disease named entity recognition on a corpus of annotated sentences". In: *BMC bioinformatics*. Vol. 9. 3. BioMed Central. 2008, pp. 1–10.

[101] Simone Magnolini, Valerio Piccioni, Vevake Balaraman, Marco Guerini, and Bernardo Magnini. "How to use gazetteers for entity recognition with neural models". In: *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*. 2019, pp. 40–49.

[102] Chan Hee Song, Dawn Lawrie, Tim Finin, and James Mayfield. "Improving neural named entity recognition with gazetteers". In: *arXiv preprint arXiv:2003.03072* (2020).

[103] Yoav Goldberg. "Neural network methods for natural language processing". In: *Synthesis Lectures on Human Language Technologies* 10.1 (2017), pp. 1–309.

[104] Onur Kuru, Ozan Arkan Can, and Deniz Yuret. "Charner: Character-level named entity recognition". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, pp. 911–921.

[105] Peng-Hsuan Li, Ruo-Ping Dong, Yu-Siang Wang, Ju-Chieh Chou, and Wei-Yun Ma. "Leveraging linguistic structures for named entity recognition with bidirectional recursive neural networks". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 2664–2669.

[106] Gerard Salton and Christopher Buckley. "Term-weighting approaches in automatic text retrieval". In: *Information processing & management* 24.5 (1988), pp. 513–523.

[107] Shaodian Zhang and Noémie Elhadad. "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts". In: *Journal of biomedical informatics* 46.6 (2013), pp. 1088–1098.

[108] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[109] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. "What do you learn from context? probing for sentence structure in contextualized word representations". In: *arXiv preprint arXiv:1905.06316* (2019).

[110] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep contextualized word representations". In: *arXiv preprint arXiv:1802.05365* (2018).

[111] Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger, and Jordi Armengol-Estapé. "Medical word embeddings for Spanish: Development and evaluation". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 2019, pp. 124–133.

[112] *Scientific Electronic Library Online*. `http://scielo.isciii.es/`. Accessed: 20 Feb. 2021.

[113] Sara Santiso, Arantza Casillas, Alicia Pérez, and Maite Oronoz. "Word embeddings for negation detection in health records written in Spanish". In: *Soft Computing* 23.21 (2019), pp. 10969–10975.

[114] Yang Liu. "Fine-tune BERT for extractive summarization". In: *arXiv preprint arXiv:1903.10318* (2019).

[115] Peng Shi and Jimmy Lin. "Simple bert models for relation extraction and semantic role labeling". In: *arXiv preprint arXiv:1904.05255* (2019).

[116] Kathleen C. Fraser, Isar Nejadgholi, Berry De Bruijn, Muqun Li, Astha LaPlante, and Khaldoun Zine El Abidine. *Extracting UMLS Concepts from Medical Text Using General and Domain-Specific Deep Learning Models*. 2019. arXiv: 1910.01274 [cs.CL].

[117] Liliya Akhtyamova and John Cardiff. "LM-based Word Embeddings Improve Biomedical Named Entity Recognition: a Detailed Analysis". In: *International Work-Conference on Bioinformatics and Biomedical Engineering*. Springer. 2020, pp. 624–635.

[118] Liliya Akhtyamova, Paloma Martínez, Karin Verspoor, and John Cardiff. "Testing Contextualized Word Embeddings to Improve NER in Spanish Clinical Case Narratives". In: (2020).

[119] Pilar López-Úbeda, MC Dıaz-Galiano, M Teresa Martın-Valdivia, and L. Alfonso Ureña-López. "Extracting neoplasms morphology mentions in spanish clinical cases through word embeddings". In: *Proceedings of IberLEF* (2020).

[120] Ludwig Wittgenstein. *Philosophical investigations*. John Wiley & Sons, 2009.

[121] Michael Krauthammer and Goran Nenadic. "Term identification in the biomedical literature". In: *Journal of biomedical informatics* 37.6 (2004), pp. 512–526.

[122] Nicola Guarino, Daniel Oberle, and Steffen Staab. "What is an ontology?" In: *Handbook on ontologies*. Springer, 2009, pp. 1–17.

[123] Mike Uschold, Martin King, Stuart Moralee, Yannis Zorgios, et al. "The enterprise ontology". In: *The knowledge engineering review* 13.1 (1998), pp. 31–89.

[124] Thomas R Gruber. "A translation approach to portable ontology specifications". In: *Knowledge acquisition* 5.2 (1993), pp. 199–220.

[125] Robert Stevens, Carole A Goble, and Sean Bechhofer. "Ontology-based knowledge representation for bioinformatics". In: *Briefings in bioinformatics* 1.4 (2000), pp. 398–414.

[126] Olivier Bodenreider. "The unified medical language system (UMLS): integrating biomedical terminology". In: *Nucleic acids research* 32.suppl_1 (2004), pp. D267–D270.

[127] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. "Gene ontology: tool for the unification of biology". In: *Nature genetics* 25.1 (2000), pp. 25–29.

[128] Ingrid M. Keseler, Amanda Mackie, Alberto Santos-Zavaleta, Richard Billington, César Bonavides-Martínez, Ron Caspi, Carol Fulcher, Socorro Gama-Castro, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Luis Muñiz-Rascado, Quang Ong, Suzanne Paley, Martin Peralta-Gil, Pallavi Subhraveti, David A. Velázquez-Ramírez, Daniel Weaver, Julio Collado-Vides, Ian Paulsen, and Peter D. Karp. "The EcoCyc database: reflecting new knowledge about Escherichia coli K-12". In: *Nucleic Acids Research* 45.D1 (Nov. 2016), pp. D543–D550. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1003. eprint: https://academic.oup.com/nar/article-pdf/45/D1/D543/8846544/gkw1003.pdf. URL: https://doi.org/10.1093/nar/gkw1003.

[129] University of Manchester. *The TAMBIS Project*. http://www.cs.man.ac.uk/~stevensr/tambis/details.html. Accessed: 20 Feb. 2021. 1998.

[130] Elliot G Brown, Louise Wood, and Sue Wood. "The medical dictionary for regulatory activities (MedDRA)". In: *Drug safety* 20.2 (1999), pp. 109–117.

[131] Kent A Spackman, Keith E Campbell, and Roger A Côté. "SNOMED RT: a reference terminology for health care." In: *Proceedings of the AMIA annual fall symposium*. American Medical Informatics Association. 1997, p. 640.

[132] Carolyn E Lipscomb. "Medical subject headings (MeSH)". In: *Bulletin of the Medical Library Association* 88.3 (2000), p. 265.

[133] Sunil Mohan and Donghui Li. "Medmentions: a large biomedical corpus annotated with UMLS concepts". In: *arXiv preprint arXiv:1902.09476* (2019).

[134] 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". In: *Scientific data* 3.1 (2016), pp. 1–9.

[135] Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M Van Mulligen, and Dietrich Rebholz-Schuhmann. "A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC". In: *Journal of the American Medical Informatics Association* 22.5 (2015), pp. 948–956.

[136] Luca Soldaini and Nazli Goharian. "Quickumls: a fast, unsupervised approach for medical concept extraction". In: *MedIR workshop, sigir*. 2016, pp. 1–4.

[137] Min Song, Hwanjo Yu, and Wook-Shin Han. "Developing a hybrid dictionary-based bio-entity recognition technique". In: *BMC medical informatics and decision making* 15.1 (2015), pp. 1–8.

[138] Fernando Suarez Saiz, Corey Sanders, Rick Stevens, Robert Nielsen, Michael Britt, Leemor Yuravlivker, Anita M Preininger, and Gretchen P Jackson. "Artificial Intelligence Clinical Evidence Engine for Automatic Identification, Prioritization, and Extraction of Relevant Clinical Oncology Research". In: *JCO Clinical Cancer Informatics* 5 (2021), pp. 102–111.

[139] Shikhar Vashishth, Rishabh Joshi, Ritam Dutt, Denis Newman-Griffis, and Carolyn Rose. "MedType: Improving Medical Entity Linking with Semantic Type Prediction". In: *arXiv preprint arXiv:2005.00460* (2020).

[140] Jitendra Jonnagaddalaa, C Siaw-teng Liaw, A Pradeep Rayb, and Manish Kumarc. "TMUNSW: identification of disorders and normalization to SNOMED-CT terminology in unstructured clinical notes". In: (2015).

[141] Amir M Tahmasebi, Henghui Zhu, Gabriel Mankovich, Peter Prinsen, Prescott Klassen, Sam Pilato, Rob van Ommering, Pritesh Patel, Martin L Gunn, and Paul Chang. "Automatic normalization of anatomical phrases in radiology reports using unsupervised learning". In: *Journal of digital imaging* 32.1 (2019), pp. 6–18.

[142] I Martinez Soriano and J Castro. "DNER Clinical (named entity recognition) from free clinical text to Snomed-CT concept". In: *WSEAS Transactions on Computers* 16 (2017), pp. 83–91.

[143] Hao Liu, Yehoshua Perl, and James Geller. "Transfer learning from BERT to support insertion of new concepts into SNOMED CT". In: *AMIA Annual Symposium Proceedings*. Vol. 2019. American Medical Informatics Association. 2019, p. 1129.

[144] Hao Liu, James Geller, Michael Halper, and Yehoshua Perl. "Using convolutional neural networks to support insertion of new concepts into SNOMED CT". In: *AMIA Annual Symposium Proceedings*. Vol. 2018. American Medical Informatics Association. 2018, p. 750.

[145] Aurélie Névéol, Aude Robert, Robert Anderson, Kevin Bretonnel Cohen, Cyril Grouin, Thomas Lavergne, Grégoire Rey, Claire Rondet, and Pierre Zweigenbaum. "CLEF eHealth 2017 Multilingual Information Extraction task Overview: ICD10 Coding of Death Certificates in English and French." In: *CLEF (Working Notes)*. 2017.

[146] Aurélie Névéol, Aude Robert, Francesco Grippo, Claire Morgand, Chiara Orsi, Laszlo Pelikan, Lionel Ramadier, Grégoire Rey, and Pierre Zweigenbaum. "CLEF eHealth 2018 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian." In: *CLEF (Working Notes)*. 2018.

[147] Curtis P Langlotz. *RadLex: a new method for indexing online educational materials*. 2006.

[148] Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana. *Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks*. 2016. arXiv: 1609.08409 [cs.CL].

[149] Surabhi Datta, Jordan Godfrey-Stovall, and Kirk Roberts. "RadLex Normalization in Radiology Reports". In: *arXiv preprint arXiv:2009.05128* (2020).

[150] Henry J Lowe and G Octo Barnett. "MicroMeSH: a microcomputer system for searching and exploring the National Library of Medicine's Medical Subject Headings (MeSH) Vocabulary". In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association. 1987, p. 717.

[151] Randolph A Miller, Filip M Gieszczykiewicz, John K Vries, and Gregory F Cooper. "CHARTLINE: providing bibliographic references relevant to patient charts using the UMLS Metathesaurus Knowledge Sources." In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association. 1992, p. 86.

[152] Thomas C Rindflesch, Lorraine Tanabe, John N Weinstein, and Lawrence Hunter. "EDGAR: extraction of drugs, genes and relations from the biomedical literature". In: *Biocomputing 2000*. World Scientific, 1999, pp. 517–528.

[153] 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". In: *Journal of the American Medical Informatics Association* 17.5 (2010), pp. 507–513.

[154] Guergana K Savova, Jin Fan, Zi Ye, Sean P Murphy, Jiaping Zheng, Christopher G Chute, and Iftikhar J Kullo. "Discovering peripheral arterial disease cases from radiology notes using natural language processing". In: *AMIA Annual Symposium Proceedings*. Vol. 2010. American Medical Informatics Association. 2010, p. 722.

[155] Jacqueline Peng, Mengge Zhao, James Havrilla, Cong Liu, Chunhua Weng, Whitney Guthrie, Robert Schultz, Kai Wang, and Yunyun Zhou. "Natural language processing (NLP) tools in extracting biomedical concepts from research articles: a case study on autism spectrum disorder". In: *BMC Medical Informatics and Decision Making* 20.11 (2020), pp. 1–9.

[156] Yunqing Xia, Xiaoshi Zhong, Peng Liu, Cheng Tan, Sen Na, Qinan Hu, and Yaohai Huang. "Combining MetaMap and cTAKES in Disorder Recognition: THCIB at CLEF eHealth Lab 2013 Task 1." In: *CLEF (Working Notes)*. 2013.

[157] Ruth Reátegui and Sylvie Ratté. "Comparison of MetaMap and cTAKES for entity extraction in clinical notes". In: *BMC medical informatics and decision making* 18.3 (2018), pp. 13–19.

[158] Alan R Aronson. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." In: *Proceedings of the AMIA Symposium*. American Medical Informatics Association. 2001, p. 17.

[159] Ying He and Mehmet Kayaalp. "Biological entity recognition with conditional random fields". In: *AMIA Annual Symposium Proceedings*. Vol. 2008. American Medical Informatics Association. 2008, p. 293.

[160] Alejandro Rodríguez-González, Roberto Costumero, Marcos Martinez-Romero, Mark D Wilkinson, and Ernestina Menasalvas-Ruiz. "Extracting diagnostic knowledge from MedLine Plus: a comparison between MetaMap and cTAKES Approaches". In: *Current Bioinformatics* 13.6 (2018), pp. 573–582.

[161] Francisco Carrero, José Carlos Cortizo, and José María Gómez. "Building a Spanish MMTx by using automatic translation and biomedical ontologies". In: *International Conference on Intelligent Data Engineering and Automated Learning*. Springer. 2008, pp. 346–353.

[162] Elena Castro, Ana Iglesias, Paloma Martínez, and Leonardo Castano. "Automatic identification of biomedical concepts in spanish-language unstructured clinical texts". In: *Proceedings of the 1st ACM International Health Informatics Symposium*. 2010, pp. 751–757.

[163] Naiara Perez, Montse Cuadros, and German Rigau. "Biomedical term normalization of EHRs with UMLS". In: *arXiv preprint arXiv:1802.02870* (2018).

[164] Rodrigo Agerri, Josu Bermudez, and German Rigau. "IXA pipeline: Efficient and Ready to Use Multilingual NLP tools." In: *LREC*. Vol. 2014. 2014, pp. 3823–3828.

[165] Eneko Agirre and Aitor Soroa. "Personalizing pagerank for word sense disambiguation". In: *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. 2009, pp. 33–41.

[166] Maite Oronoz, Arantza Casillas, Koldo Gojenola, and Alicia Perez. "Automatic annotation of medical records in Spanish with disease, drug and substance names". In: *Iberoamerican Congress on Pattern Recognition*. Springer. 2013, pp. 536–543.

[167] Xavier Carreras, Isaac Chao, Lluis Padró, and Muntsa Padró. "FreeLing: An Open-Source Suite of Language Analyzers." In: *LREC*. 2004, pp. 239–242.

[168] Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, Arturo Montejo-Ráez, María-Teresa Martín-Valdivia, and L Alfonso Ureña-López. "An Integrated Approach to Biomedical Term Identification Systems". In: *Applied Sciences* 10.5 (2020), p. 1726.

[169] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. "The Stanford CoreNLP natural language processing toolkit". In: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014, pp. 55–60.

[170] Lluís Padró and Evgeny Stanilovsky. "Freeling 3.0: Towards wider multilinguality". In: *LREC2012*. 2012.

[171] Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. "How to train good word embeddings for biomedical NLP". In: *Proceedings of the 15th workshop on biomedical natural language processing*. 2016, pp. 166–174.

[172] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[173] Mathias Etcheverry and Dina Wonsever. "Spanish word vectors from wikipedia". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 3681–3685.

[174] *Índice Bibliográfico Español en Ciencias de la Salud*. `https://ibecs.isciii.es/`. Accessed: 20 Feb. 2021.

[175] *PubMed*. `https://pubmed.ncbi.nlm.nih.gov/`. Accessed: 20 Feb. 2021.

[176] *MedlinePlus*. `https://medlineplus.gov/xml.html`. Accessed: 20 Feb. 2021.

[177] Jörg Tiedemann. "News from OPUS-A collection of multilingual parallel corpora with tools and interfaces". In: *Recent advances in natural language processing*. Vol. 5. 2009, pp. 237–248.

[178] *Wikipedia health*. `https://bit.ly/3bqkXip`. Accessed: 20 Feb. 2021.

[179] *Webconsultas: revista de salud y bienestar*. `https://www.webconsultas.com`. Accessed: 20 Feb. 2021.

[180] *WebMD Health News Center - The latest Spanish news*. `https://www.webmd.com/news/spanish`. Accessed: 20 Feb. 2021.

[181] *Organización Mundial de la Salud*. `https://www.who.int/es`. Accessed: 20 Feb. 2021.

[182] *Mujer y Salud. Mejor salud, vida y bienestar para la mujer de hoy en día*. `http://www.mujerysalud.es`. Accessed: 20 Feb. 2021.

[183] *Mejor con Salud - Revista sobre buenos hábitos y cuidados para tu salud*. `https://mejorconsalud.com`. Accessed: 20 Feb. 2021.

[184] *Mayo clinic*. `https://www.mayoclinic.org/es-es`. Accessed: 20 Feb. 2021.

[185] *Diario médico*. `https://www.diariomedico.com/`. Accessed: 20 Feb. 2021.

[186] *Efe Salud - Noticias sobre salud por la Agencia Efe*. https://www.efesalud.com/espana. Accessed: 20 Feb. 2021.

[187] Alan Akbik, Duncan Blythe, and Roland Vollgraf. "Contextual string embeddings for sequence labeling". In: *Proceedings of the 27th International Conference on Computational Linguistics*. 2018, pp. 1638–1649.

[188] Alan Akbik, Tanja Bergmann, and Roland Vollgraf. "Pooled Contextualized Embeddings for Named Entity Recognition". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 724–728. DOI: 10.18653/v1/N19-1078. URL: https://www.aclweb.org/anthology/N19-1078.

[189] Jörg Tiedemann. "Parallel Data, Tools and Interfaces in OPUS." In: *Lrec*. Vol. 2012. 2012, pp. 2214–2218.

[190] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units". In: *arXiv preprint arXiv:1508.07909* (2015).

[191] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144* (2016).

[192] Taku Kudo. "Subword regularization: Improving neural network translation models with multiple subword candidates". In: *arXiv preprint arXiv:1804.10959* (2018).

[193] Taku Kudo and John Richardson. "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing". In: *arXiv preprint arXiv:1808.06226* (2018).

[194] Stefan Kombrink, Tomáš Mikolov, Martin Karafiát, and Lukáš Burget. "Recurrent neural network based language modeling in meeting recognition". In: *Twelfth annual conference of the international speech communication association*. 2011.

[195] Tomáš Mikolov, Anoop Deoras, Daniel Povey, Lukáš Burget, and Jan Černockỳ. "Strategies for training large scale neural network language models". In: *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE. 2011, pp. 196–201.

[196] Chen Lyu, Bo Chen, Yafeng Ren, and Donghong Ji. "Long short-term memory RNN for biomedical named entity recognition". In: *BMC bioinformatics* 18.1 (2017), p. 462.

[197] Jeffrey L Elman. "Finding structure in time". In: *Cognitive science* 14.2 (1990), pp. 179–211.

[198] Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A Smith. "Discriminative lexical semantic segmentation with gaps: running the MWE gamut". In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 193–206.

[199] Lev Ratinov and Dan Roth. "Design challenges and misconceptions in named entity recognition". In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*. 2009, pp. 147–155.

[200] Mike Schuster and Kuldip K Paliwal. "Bidirectional recurrent neural networks". In: *IEEE transactions on Signal Processing* 45.11 (1997), pp. 2673–2681.

[201] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. "Neural architectures for named entity recognition". In: *arXiv preprint arXiv:1603.01360* (2016).

[202] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 54–59. DOI: 10.18653/v1/N19-4010. URL: https://www.aclweb.org/anthology/N19-4010.

[203] Rich Caruana, Steve Lawrence, and C Lee Giles. "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping". In: *Advances in neural information processing systems*. 2001, pp. 402–408.

[204] Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. "Drug name recognition: approaches and resources". In: *Information* 6.4 (2015), pp. 790–810.

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

[206] Isabel Segura-Bedmar, Paloma Martínez, and María Segura-Bedmar. "Drug name recognition and classification in biomedical texts: a case study outlining approaches underpinning automated systems". In: *Drug discovery today* 13.17-18 (2008), pp. 816–823.

[207] Pernille Warrer, Ebba Holme Hansen, Lars Juhl-Jensen, and Lise Aagaard. "Using text-mining techniques in electronic patient records to identify ADRs from medicine use". In: *British journal of clinical pharmacology* 73.5 (2012), pp. 674–684.

[208] Jon Patrick, Yefeng Wang, and Peter Budd. "An automated system for conversion of clinical notes into SNOMED clinical terminology". In: *Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68*. Australian Computer Society, Inc. 2007, pp. 219–226.

[209] 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*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1–10. DOI: 10.18653/v1/D19-5701. URL: https://www.aclweb.org/anthology/D19-5701.

[210] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. "BRAT: a web-based tool for NLP-assisted text annotation". In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012, pp. 102–107.

[211] Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štepánek, et al. "The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages". In: (2009).

[212] Ying Xiong, Yedan Shen, Yuanhang Huang, Shuai Chen, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Jun Yan, and Yi Zhou. "A Deep Learning-Based System for PharmaCoNER". In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. 2019, pp. 33–37.

[213] Fernando Sónchez León and Ana González Ledesma. *Annotating and normalizing biomedical NEs with limited knowledge*. 2019. arXiv: 1912.09152 [cs.CL].

[214] Benjamin Heinzerling and Michael Strube. *BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages*. 2017. arXiv: 1710.02187 [cs.CL].

[215] Fernando Sánchez-León. "Arborex: Abbreviation resolution based on regular expressions for barr2". In: *IberEval@ SEPLN* (2018), pp. 302–315.

[216] Leon Derczynski. "Complementarity, F-score, and NLP Evaluation". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 261–266.

[217] World Health Organization et al. "Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018". In: *International Agency for Research on Cancer. Geneva: World Health Organization* (2018). URL: https://www.iarc.who.int/wp-content/uploads/2020/12/pr292_E.pdf.

[218] Sociedad Española de Oncología Médica (SEOM). *Las cifras del cáncer en España 2020*. https://seom.org/seomcms/images/stories/recursos/Cifras_del_cancer_2020.pdf. Accessed: 20 Feb. 2021.

[219] Irena Spasić, Jacqueline Livsey, John A Keane, and Goran Nenadić. "Text mining of cancer-related information: review of current status and future directions". In: *International journal of medical informatics* 83.9 (2014), pp. 605–623.

[220] Hua Xu, Kristin Anderson, Victor R Grann, and Carol Friedman. "Facilitating cancer research using natural language processing of pathology reports." In: *Studies in health technology and informatics* 107.Pt 1 (2004), p. 565.

[221] Yuqi Si and Kirk Roberts. "A frame-based NLP system for cancer-related information extraction". In: *AMIA Annual Symposium Proceedings*. Vol. 2018. American Medical Informatics Association. 2018, p. 1524.

[222] Joshua C Denny, Neesha N Choma, Josh F Peterson, Randolph A Miller, Lisa Bastarache, Ming Li, and Neeraja B Peterson. "Natural language processing improves identification of colorectal cancer testing in the electronic medical record". In: *Medical Decision Making* 32.1 (2012), pp. 188–197.

[223] Xiaohui Zhang, Yaoyun Zhang, Qin Zhang, Yuankai Ren, Tinglin Qiu, Jianhui Ma, and Qiang Sun. "Extracting comprehensive clinical information for breast cancer using deep learning methods". In: *International Journal of Medical Informatics* 132 (2019), p. 103985.

[224] Ying Xiong, Yuanhang Huang, Qingcai Chen, Xiaolong Wang, Yuan Nic, and Buzhou Tang. "A Joint Model for Medical Named Entity Recognition and Normalization". In: *Proceedings http://ceur-ws. org ISSN* 1613 (2020), p. 0073.

[225] Aitor García-Pablos, Naiara Perez, and Montse Cuadros. "Vicomtech at CANTEMIST 2020". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*. 2020.

[226] Lukas Lange, Xiang Dai, Heike Adel, and Jannik Strötgen. *NLNDE at CANTEMIST: Neural Sequence Labeling and Parsing Approaches for Clinical Concept Extraction*. 2020. arXiv: 2010.12322 [cs.CL].

[227] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. "A unified mrc framework for named entity recognition". In: *arXiv preprint arXiv:1910.11476* (2019).

[228] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. "Zero-shot relation extraction via reading comprehension". In: *arXiv preprint arXiv:1706.04115* (2017).

[229] Juntao Yu, Bernd Bohnet, and Massimo Poesio. "Named Entity Recognition as Dependency Parsing". In: *arXiv preprint arXiv:2005.07150* (2020).

[230] Alejandro Piad-Morffis, Yoan Gutiérrez, Suilan Estévez-Velarde, Yudivián Almeida-Cruz, Andrés Montoyo, and Rafael Muñoz. "Analysis of eHealth knowledge discovery systems in the TASS 2018 Workshop". In: *Procesamiento de Lenguaje Natural* (2019). ISSN: 19897553. DOI: 10.26342/2019-62-1.

[231] 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: *CEUR Workshop Proceedings*. 2019.

[232] Alejandro Piad-Morffis, Yoan Gutiérrez, Suilan Estevez-Velarde, Yudivián Almeida-Cruz, Rafael Muñoz, and Andrés Montoyo. "Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2020". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*. 2020.

[233] Naomi Miller, Eve-Marie Lacroix, and Joyce EB Backus. "MEDLINEplus: building and maintaining the National Library of Medicine's consumer health Web service". In: *Bulletin of the Medical Library Association* 88.1 (2000), p. 11.

[234] Alejandro Rodríguez-González, Marcos Martínez-Romero, Roberto Costumero, Mark D Wilkinson, and Ernestina Menasalvas-Ruiz. "Diagnostic knowledge extraction from medlineplus: an application for infectious diseases". In: *9th International Conference on Practical Applications of Computational Biology and Bioinformatics*. Springer. 2015, pp. 79–87.

[235] Nancy Chinchor and Beth Sundheim. "MUC-5 evaluation metrics". In: *Proceedings of the 5th conference on Message understanding*. Association for Computational Linguistics. 1993, pp. 69–78.

[236] Aitor García-Pablos, Naiara Perez, Montse Cuadros, and Elena Zotova. "Vicomtech at eHealth-KD Challenge 2020: Deep End-to-End Model for Entity and Relation Extraction in Medical Text". In: *Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@ SEPLN*. Vol. 2020. 2020.

[237] S Medina and J Turmo. "TALP at eHealth-KD Challenge 2020: Multi-Level Recurrent and Convolutional Neural Networks for Joint Classification of Key-Phrases and Relations". In: *Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@ SEPLN*. Vol. 2020. 2020.

[238] Alejandro Rodríguez-Pérez, Ernesto Quevedo-Caballero, Jorge Mederos-Alvarado, Rocío Cruz-Linares, and Juan Pablo Consuegra-Ayala. "UH-MAJA-KD at eHealth-KD Challenge 2020: Deep Learning Models for Knowledge Discovery in Spanish eHealth Documents". In: ().

[239] Pilar López Úbeda, Manuel Carlos Díaz-Galiano, L Alfonso Ureña López, M Teresa Martín-Valdivia, Teodoro Martín-Noguerol, and Antonio Luna. "Transfer learning applied to text classification in Spanish radiological reports". In: *Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultiligualBIO 2020)*. 2020, pp. 29–32.

Appendix A

Comparative results

Word embeddings	**PharmaCoNER**			**Cantemist**			**eHeatlh-kd**		
	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
Medical word embeddings for Spanish [111]	83.76	81.29	82.51	78.1	82.4	80.2	83.46	78.96	81.15
Spanish Unannotated Corpora [25]	84.78	81.97	83.35	80.1	82	80.6	83.02	79.59	81.27
SBWC[1] [25]	84.32	84.23	84.27	79.7	79.6	79.6	82.39	78.69	80.5
SME + Pooled	91.85	90.13	90.98	85.6	85.2	85.4	84.18	80.4	82.24
SME + BETO	86.82	85.15	85.98	83.3	77.5	80.3	84.25	79.86	81.99
SME + BETO + XLMRoberta	85.51	84.46	84.98	81.9	76.4	79	84.62	80.67	82.62
SME + BETO + mBERT	84.64	84.55	84.59	82.8	76.9	79.8	85.59	80.67	83.06
Pooled + BETO + mBERT	81.24	82.07	81.66	79.1	68.5	73.4	84.11	80.94	82.49
Glove + SME + Pooled	92.71	90.83	91.76	85.7	85.2	85.5	83.12	80.58	81.83
Glove + SME + Pooled + BETO	85.58	85.13	85.35	82.7	76.7	79.6	83.77	79.86	81.77
Glove + SME + Pooled + XLMRoberta	90.98	89.38	90.17	86	83.7	84.8	83.3	79.86	81.54
Glove + SME + BETO + XLMRoberta	85.75	84.56	85.15	82.3	76.7	79.4	83.77	79.86	81.77
Glove + SME + BETO	87.47	85.79	86.62	84	78.6	81.2	83.03	80.94	81.97
Glove + SME + XLMRoberta	90.67	89.12	89.89	85.8	84.1	85	83.18	80.04	81.58
Glove + SME + mBERT	89.5	87.15	88.31	84.2	79	81.5	83.68	79.77	81.68
BETO+XLMRoberta+mBERT	82.04	81.12	81.58	80.3	70.9	75.3	84.65	80.85	82.7

Table A.1: Additional results of the NER task performance in the different scenarios proposed.