UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÁTICA

Named Entity Recognition and Linking in a Multilingual Biomedical Setting

Vítor Daniel Torres Andrade

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:

Professor Doutor Francisco José Moreira Couto

2021

Agradecimentos

Esta dissertação é dedicada à memória da minha mãe. Apesar de ter me apoiado no início desta dissertação e durante todo o meu percurso académico, ela não pôde ver me concluir esta etapa. Sem os sacrifícios para me proporcionar a melhores condições de vida, os ensinamentos, as brincadeiras, o carinho e amor incondicional não seria possível ter chegado aqui. Obrigado mãe.

Agradeço ao Professor Francisco Couto por me ter dado a oportunidade de trabalhar no seu grupo e por toda a orientação que me deu ao longo deste trabalho. Gostaria também de agradecer ao Pedro Ruas pela disponibilidade para me ajudar sempre que necessitei e por todas as sugestões e críticas que foram essenciais para a realização desta dissertação.

Agradeço à FCT por ter apoiado financeiramente esta dissertação através do projeto DeST: Deep SemanticTagger project, ref. PTDC/CCI-BIO/28685/2017 e à unidade de investigação LASIGE, ref. UIDB/00408/2020.

Por último, mas não mesmo importante, um obrigado à minha família, em especial ao meu pai e a minha irmã, que durante a fase mais difícil das nossas vidas continuaram a apoiar-me para concluir este trabalho.

Abstract

Information analysis is an essential process for all researchers and physicians. However, the amount of biomedical literature that we currently have available and the format in which it is found make this process difficult. Therefore, it is essential to apply text mining tools to automatically obtain information from these documents. The problem is that most of these tools are not designed to deal with non-English languages, which is critical in the biomedical literature, since many of these documents are written in the authors' native language.

Although there have been organized several shared tasks where text mining tools were developed for the Spanish language, the same does not happen for the Portuguese language. However, due to the lexical similarity between the two languages, it is possible to hypothesize that the tools for the two languages may be similar and that there is an annotation transfer between Portuguese and Spanish.

To contribute to the development of text mining tools for Portuguese and Spanish, this dissertation presents the ICERL (Iberian Cancer-related Entity Recognition and Linking) system, a NERL (Named Entity Recognition and Linking) system that uses deep learning and it is composed of two similar pipelines for each language, and the parallel corpus ICR (Iberian Cancer-related) corpus. Both these tools are focused on the oncology domain. The application of the ICERL system on the ICR corpus resulted in 3,999 annotations in Spanish and 3,287 in Portuguese. The similarities between the annotations of the two languages and the F1-score of 0.858 that resulted from the comparison of the Portuguese annotations with the Spanish ones confirm the hypothesis initially presented.

Keywords: Biomedical Literature, Named Entity Recognition, Named Entity Linking, Deep Learning, Iberian Setting.

Resumo

A divulgação de descobertas realizadas pelos investigadores e médicos é feita através de vários documentos como livros, artigos, patentes e outros tipos de publicações. Para que investigadores estejam atualizados sobre a sua área de interesse, é essencial que realizem uma análise rápida e eficaz destes documentos. Isto porque, quanto mais eficiente for esta fase, melhores serão os resultados que serão obtidos e, quanto mais rápida for, mais tempo poderão dedicar a outras componentes dos seus trabalhos. No entanto, a velocidade com que estes documentos são publicados e o facto de o texto presente nos mesmos ser expresso em linguagem natural dificulta esta tarefa. Por isso, torna-se essencial a aplicação de ferramentas de prospeção de texto para a extração de informação.

As ferramentas de prospeção de texto são compostas por diversas etapas, como por exemplo, Reconhecimento de Entidades Nomeadas (em inglês *Named Entity Recognition* ou NER) e Mapeamento de Entidades Nomeadas (em inglês *Named Entity Linking* ou NEL). A etapa NER corresponde à identificação de uma entidade no texto. NEL consiste na ligação de entidades a uma base de conhecimento. Os sistemas estado-de-arte para a NER são métodos de aprendizagem profunda e normalmente utilizam a arquitetura BiLSTM-CRF. Por outro lado, os sistemas estado-de-arte NEL usam não só métodos de aprendizagem profunda, mas também métodos baseados em grafos.

A maioria dos sistemas de prospeção de texto que atualmente temos disponíveis está desenhada apenas para a língua inglesa, o que é problemático, pois muitas das vezes a literatura biomédica encontra-se descrita na língua nativa dos autores. Para resolver este problema têm surgido competições para desenvolver sistemas de prospeção de texto para outras línguas que não o inglês. Uma das línguas que têm sido um dos principais focos destas competições é a língua espanhola. O espanhol é a segunda língua com o maior número de falantes nativos no mundo e com um elevado número de publicações biomédicas disponível. Um dos exemplos de competições para a língua espanhola é o CANTEMIST. O objetivo do CANTEMIST passa pela identificação de entidades do domínio oncológico e a ligação das mesmas à base de dados *Clasificación Internacional de Enfermedades para Oncología* (CIE-O). Por outro lado, o português não têm sido alvo de grande interesse por parte destas competições.

Devido ao facto de que o português e o espanhol derivarem do latim, existe uma semelhança lexical elevada entre as duas línguas (89%). Portanto, é possível assumir que as soluções encontradas para espanhol possam ser adaptadas ou utilizadas para o português, e que exista transferências de anotações entre as duas línguas. Por isso, o objetivo deste trabalho passa por criar ferramentas que validem esta hipótese: o sistema ICERL (*Iberian Cancer-related Entity Recognition and Linking*) e o corpus ICR

(*Iberian Cancer-related*). O sistema ICERL é um sistema NERL (*Named Entity Recognition and Linking*) bilíngue português-espanhol, enquanto que o ICR é um corpus paralelo para as mesmas línguas. Ambas as ferramentas estão desenhadas para o domínio oncológico.

A primeira etapa no desenvolvimento do sistema ICERL passou pela criação de uma *pipeline* NERL para a língua espanhola específica para o domínio oncológico. Esta *pipeline* foi baseada no trabalho desenvolvido pela equipa LasigeBioTM na competição CANTEMIST. A abordagem apresentada pelo LasigeBioTM no CANTEMIST consiste na utilização da *framework* Flair para a tarefa NER e do algoritmo *Personalized PageRank* (PPR) para a tarefa NEL. O Flair é uma ferramenta que permite a combinação de diferentes *embeddings* (representações vetoriais para palavras) de diferentes modelos num só para a tarefa NER. O PPR é uma variação do algoritmo *PageRank* que é utilizado para classificar importância de páginas *web*. O algoritmo *PageRank* é aplicado sobre um grafo. Originalmente, cada nó do grafo representava uma página *web* e as ligações entre nós representavam hiperligações entre páginas. O algoritmo estima a coerência de cada nó no grafo, isto é, a sua relevância. No contexto da tarefa NEL, o grafo é composto por candidatos para as entidades de interesse. O Flair foi utilizado pela equipa LasigeBioTM para o treino de *embeddings* que foram obtidos em documentos em espanhol do PubMed. Estes *embeddings* foram integrados num modelo para NER que foi treinado nos conjuntos de treino e desenvolvimento do corpus do CANTEMIST. O modelo treinado foi depois utilizado no conjunto de teste do corpus do CANTEMIST para a obtenção de ficheiros de anotação com as entidades reconhecidas. Foi depois feita uma procura pelos candidatos para a tarefa de NEL das entidades reconhecidas em três bases de dados: o CIE-O, o *Health Sciences Descriptors* (DeCS) e o *International Classification of Diseases* (ICD). A partir destes candidatos foi construído um grafo e através do algoritmo PPR os candidatos foram classificados e foi escolhido o melhor candidato para ligar cada entidade. Esta *pipeline* foi aperfeiçoada através da adição de novos *embeddings*, um prolongamento do treino no modelo NER e uma correção de erros no código do sistema para a tarefa NEL. Apesar destas alterações contribuírem para um aumento significativo na performance da tarefa NEL (medida-F de 0.0061 para 0.665), o mesmo não aconteceu para a tarefa NER (medida-F de 0.741 para 0.754). A versão final do sistema ICERL é composta por uma *pipeline* para a língua portuguesa e pela *pipeline* que foi testada no corpus do CANTEMIST, com uma ligeira diferença na tarefa NEL: em vez de ser escolhido apenas um candidato para cada entidade, é escolhida uma lista de candidatos do CIE-O e o DeCS. Já na *pipeline* portuguesa são escolhidos candidatos do DeCS e da Classificação Internacional de Doenças (CID). Esta diferença na tarefa NEL deve-se ao método que foi utilizado para avaliar a performance do sistema ICERL e para não restringir o sistema a apenas um candidato e a um vocabulário. Para a construção da *pipeline* portuguesa, três modelos para a tarefa NER foram testados e concluiu-se que a melhor abordagem passaria pela combinação de um modelo semelhante ao modelo utilizado na *pipeline* espanhola e o modelo BioBERTpt. Devido à elevada semelhança lexical entre as duas línguas, foi testada a hipótese de utilização da mesma *pipeline* para as duas línguas. No entanto, através do software NLPStatTest foi possível concluir que a utilização de uma *pipeline* específica para cada língua traduz-se numa melhoria de 58 por cento na medida-F para os textos em português.

O corpus ICR é composto por 1555 documentos para cada língua que foram retirados do SciELO. Uma vez que a *pipeline* espanhola foi treinada com ficheiros do CANTEMIST corpus, foi também necessário

retirar documentos do SciELO e do PubMed para treinar a *pipeline* portuguesa.

O sistema ICERL foi aplicado ao corpus ICR e o método de avaliação passou pela comparação dos resultados das anotações portuguesas com as anotações em espanhol. Isto porque foi possível avaliar a performance da *pipeline* espanhol no corpus do CANTEMIST, e os resultados obtidos foram próximos do estado-de-arte. A aplicação do sistema ICERL no corpus ICR resultou em 3999 anotações em espanhol sendo que 216 dessas anotações são únicas e 3287 em português sendo que 171 dessas anotações são únicas. Para além disso, a entidade *câncer* é a entidade mais frequente para as duas línguas. Para além destas semelhanças nas anotações, o facto de ter sido obtido 0.858 em medida-F no método de avaliação permite concluir que existe transferências de anotações entre as duas línguas e que é possível utilizar ferramentas de prospeção de texto semelhantes para ambas.

Palavras Chave: Literatura Biomédica, Reconhecimento de Entidade, Mapeamento de Entidade, Aprendizagem Profunda, Contexto Ibérico.

Contents

List of Figures

List of Tables

Acronyms

#SMM4H Social Media Mining for Health Applications.
ANNs Artificial Neural Networks.
BiLSTM Bidirectional Long Short Term Memory.
BIREME Latin American and Caribbean Center on Health Sciences Information.
ChEBI Chemical Entities of Biological Interest.
CID Classificação Internacional de Doenças.
CID-O Classificação Internacional de Doenças para Oncologia.
CIE Clasificación Internacional de Enfermedades.
CIE-O Clasificación Internacional de Enfermedades para Oncología.
CNN Convolutional Neural Network.
CodiEsp Clinical Case Coding in Spanish Shared Task.
CRF Conditional Random Field.
DeCS Descritores em Ciências da Saúde.
DO Disease Ontology.
FN False Negative.
FP False Positive.
GO Gene Ontology.
HMM Hidden Markov Models.
HPO Human Phenotype Ontology.
IBECS Índice Bibliográfico Español en Ciencias de la Salud.
IC Information Content.
ICD International Classification of Diseases.
ICERL Iberian Cancer-related Entity Recognition and Linking.
ICR Iberian Cancer-related.
IE Information Extraction.
KB Knowledge Base.
LILACS Literatura Latino-Americana e do Caribe em Ciências da Saúde.
LSTM Long Short Term Memory.
MeSH Medical Subject Headings.
MRRAD Multilingual Radiology Research Articles Dataset.
NEL Named Entity Linking.

NER Named Entity Recognition.
NERL Named Entity Recognition and Linking.
NLM National Library of Medicine.
NLP Natural Language Processing.
OBO Open Biomedical Ontology.
OWL Web Ontology Language.
POS Part-of-speech.
PPR Personalized PageRank.
REEC Registro Español de Estudios Clínicos.
RNN Recurrent Neural Network.
RO Relations Ontology.
SSM Semantic Similarity Measurement.
SVM Support Vector Machine.
TP True Positive.
UMLS Unified Medical Language System.
WHO World Health Organization.

Chapter 1

Introduction

Biomedical literature is the primary dissemination method used by researchers and physicians to share their findings. It includes articles, patents, and other written reports, so it is an essential source of knowledge [Hearst, 1999]. The biomedical experts studying a subject must have access to cutting-edge information about it. However, the large quantity of literature being published in recent years makes this task difficult [Lamurias and Couto, 2019]. Too much time is wasted in this process, and it is unfeasible to analyze all the data of interest. This could affect the veracity and quality of the information because the data could be outdated or incomplete and negatively affect an entire project. In addition, the significant amount of time spent by the researchers in this process could be helpful for the remaining phases of their research. Furthermore, biomedical literature comprises an extensive collection of text expressed in natural language, which computers usually do not understand. These two reasons motivate the application of text mining tools to extract information from those documents automatically [Sousa, 2019].

Text mining is the process of extracting interesting and non-trivial patterns or knowledge from unstructured text [Tan et al., 1999]. Text mining pipeline includes, among others, **Named Entity Recognition (NER)** and **Named Entity Linking (NEL)**. NER corresponds to the recognition of entities mentioned in the text. These entities can be described as anything with a proper name: a person, a location, or an organization [Jurafsky and Martin, 2009]. In the biomedical field, these entities can be a gene or a disease. NEL corresponds to the mapping of the recognized entities to entries in a given knowledge base (KB). These tasks provide researchers and physicians a more effective way to obtain, integrate and interpret data from different sources [Zhu et al., 2013] and to reduce the required time to process information [Wei et al., 2013; Simon et al., 2019]. They can be carried out by different methods such as rule-based, machine learning, and deep learning. Deep learning methods are the state-of-the-art for the NER task, but rule-based methods such as graph models can also be included in the state-of-the-art for the NEL task.

One of the problems associated with biomedical text mining is the non-English text. Even though there is a considerable amount of these texts in repositories such as PubMed and SciELO [Neves et al., 2016], there is also a lack of resources to deal with them. For example, in 2017, less than 10% of the publications about ”Natural Language Processing” (NLP) were focused on non-English languages [Névéol

et al., 2018]. This number is problematic since only about 5% of the world population has English as their native language[1], which means that the majority of current solutions is not designed for 95% of the world population. The lack of these tools is critical for biomedical texts since usually, these are expressed in the author's native language. The Spanish language is a prime example of the necessity of these resources. Spanish is the second most spoken native language in the world with more than 460,000,000 native speakers[1] and with more than 374,000 entries in PubMed. Although there is a lack of text mining tools for the Spanish language, this shortage is even more significant for the Portuguese language. Even though the Portuguese language has a smaller population of native speakers than Spanish, 220,000,000 [1], and a smaller number of texts available, more than 111,000 entries in PubMed, these numbers are also significant and require text mining tools designed to extract information.

There are two possible text mining approaches for non-English text. The first is translating the text in English and then applying text mining tools in the translated text, like, for example, the work of [Campos et al., 2017] which developed a system for translations of radiology articles in Portuguese. The second is applying the text mining techniques directly in text and use terminologies in that language. Even though the first approach have shown to be successful, it was demonstrated that it does not always brings benefits in the results [Rosales-Méndez et al., 2018], since a translation could result in a loss of relevant information, such as the name of an entity, especially if done by non-experts. Furthermore, the first approach does not take advantage or contributes to the development of the multilingual tools like for example, the multilingual version of a KB like "Clasificación Internacional de Enfermedades" (CIE)[2], "Classificação Internacional de Doenças" (CID) and Health Sciences Descriptors (DeCS) [3], or NER systems in the biomedical field designed for non-English languages such as BioBERTpt. Therefore, the second approach will be the focus of this work.

1.1 Objectives

Given the large amount of biomedical literature available, the application of text mining tools is essential for performing information retrieval by researchers and physicians. Considering the lack of these tools for languages other than English, a series of shared tasks to create state-of-the-art NER and NEL solutions for clinical text has emerged in recent years. One example of these shared tasks is the CANTEMIST[4] whose goal is to recognize entity mentions of tumor morphology in Spanish health documents and assign them to their respective "Clasificación Internacional de Enfermedades para Oncología" (CIE-O) codes. Due to the significant amount of biomedical literature published in Spanish and the number of Spanish speakers globally, this language has been addressed in several shared tasks. However, the number of shared tasks and text mining tools for Portuguese text is lower when compared to Spanish. Since there

[1] https://web.archive.org/web/20190312060544/https://www.ethnologue.com/statistics/size
[2] https://icd.who.int/es
[3] https://decs.bvsalud.org/
[4] https://temu.bsc.es/cantemist/

is a high lexical similarity between Portuguese and Spanish [Castro et al., 2018, as cited in Ulsh, 1971], due to the fact that they both derived from Latin, this work hypothesizes that it is possible to use similar text mining tools for Portuguese and Spanish and to transfer annotations between the two languages.

1.2 Methodology

The methodology used to validate the hypothesis presented in this work is composed of four steps:

1. Development of a state-of-the-art Named Entity Recognition and Linking (NERL) pipeline for the Spanish language;

2. Creation of a parallel corpus for Portuguese and Spanish;

3. Development of a bilingual Portuguese and Spanish deep learning NERL system;

4. Application of the bilingual NERL system to the parallel corpus and compare the results between Portuguese and Spanish.

The developed Spanish pipeline was based on the LasigeBioTM team's work at the CANTEMIST [Ruas et al., 2020] shared task. Our approach consisted in using the Flair framework for the NER task and the Personalized PageRank (PPR) for the NEL task. Improvements were made to that pipeline to achieve the state-of-the-art results that other approaches obtained in CANTEMIST. These improvements refer to the addition of Character and Bytepair embeddings, the extension of the training process for the NER model, and the correction of code errors in the system developed for the NEL task. This pipeline and the other two tools developed in this work: the ICR (Iberian Cancer-related) and the ICERL (Iberian Cancer-related Entity Recognition and Linking), are focused on the oncological domain.

To build the ICR corpus, parallel abstracts in Portuguese and Spanish were retrieved from SciELO. The ICERL system is composed of the Spanish pipeline resulting from the first phase of this dissertation and a similar pipeline for the Portuguese language combined with the BioBERTpt model. The Spanish pipeline was trained using the training and development sets from the CANTEMIST corpus and abstracts from PubMed. Since a correspondence in Portuguese of these files is not available, files from SciELO and Pubmed in Portuguese were also retrieved for the training of the Portuguese pipeline of the ICERL system.

The ICERL system was applied to the ICR corpus to generate the annotations and compare the ICERL system's performance in the two languages. Since it was possible to assess the performance of the Spanish pipeline on the CANTEMIST corpus, the evaluation method consisted in comparing the results of the Portuguese pipeline with the results of the Spanish pipeline. This evaluation method can be done since KBs with versions of each language were selected for the NEL task: the vocabulary DeCS in Portuguese and Spanish, the CIE, and the CID.

To further evaluate the ICERL system, this work resulted in the participation in shared tasks. However, during this dissertation, there were none available for the Portuguese language or the oncological domain; it was only possible to test the ICERL system at ProfNER and MESINESP2: shared tasks for the Spanish language not cancer-related.

1.3 Contributions

The two main contributions of this dissertation are the ICERL system and the ICR corpus. These contributions are of the utmost importance because, as far as I know, there are no NERL systems in the oncological domain that are designed to handle both Portuguese and Spanish. Furthermore, although there are biomedical corpora in Spanish and Portuguese for NERL systems, such as the corpora provided by the text mining shared tasks, to the best of my knowledge, there are no Portuguese-Spanish parallel corpora for the oncological domain.

The development of the ICERL system and the ICR corpus resulted in the submission of a paper:

- Andrade, V. D., Ruas, P., and Couto, F. M. (2021). Named Entity Recognition and Linking: a Portuguese and Spanish Oncological Parallel Corpus. doi: 10.1101/2021.09.16.460605.

This work also contributed to the participation of LasigeBioTM at ProfNER and MESINESP2. At ProfNER, the developed NER system and the rule-based module achieved the second-best performance in the classification task. At MESINESP2, I created a NEL feature for the extreme multi-label classification system used in the shared task. These two participations are described in two workshop papers:

- Ruas, P., Andrade, V. D., and Couto, F. M. (2021). Lasige-BioTM at Profner: BILSTM-CRF and contextual Spanish embeddings for Named Entity Recognition and Tweet Binary Classification. In Proceedings of the Sixth Social Media Mining for Health (SMM4H) Workshop and Shared Task;

- Ruas, P., Andrade, V. D., and Couto, F. M. (2021) Lasige-BioTM at MESINESP2: entity linking with semantic similarity and extreme multi-label classification on Spanish biomedical documents. In Proceedings of the Working Notes of Conference and Labs of the Evaluation Forum (CLEF).

In addition to ProfNER and MESINESP2, I was also involved in the participation of LasigeBioTM at CANTEMIST. In this shared task, my role was to test the solutions produced by the team and find ways to improve them. This participation is described in the following workshop paper:

- Ruas, P., Neves, A., Andrade, V. D., and Couto, F. M (2020). LasigeBioTM at CANTEMIST: Named Entity Recognition and Normalization of Tumour Morphology Entities and Clinical Coding of Spanish Health-related documents. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020).

1.4 Document structure

In addition to the first chapter, this document is structured in four chapters as follows:

- Chapter 2 (Related work): introduces the necessary concepts to understand the work done in this dissertation, which includes a description of NER and NEL tasks and the approaches available to perform them, biomedical knowledge organization systems, text mining shared tasks, and the evaluation methodology used in this work.
- Chapter 3 (CANTEMIST pipeline): describes the LasigeBioTM's pipeline at the CANTEMIST and the improvements that I made to it.
- Chapter 4 (Iberian corpus (ICR) and NERL system (ICERL)): describes the development of the ICR corpus and the ICERL system. The first section of this chapter explains the work done for the ICR corpus and the other documents used to train the ICERL system. The second part refers to the steps taken to develop the final version of the ICERL system from the pipeline described in the Chapter 3.
- Chapter 5 (Conclusions): presents the main conclusions of this dissertation and some suggestions for future work.

Chapter 2

Related Work

2.1 Information Extraction Approaches

Information Extraction (IE) corresponds to the process of turning unstructured data into structured data [Jurafsky and Martin, 2009]. This process includes approaches such as rule-based, machine learning, and deep learning. The IE is based on rules that include terms, regular expressions, and sentence constructions defined by experts [Lamurias and Couto, 2019]. These rules allow the system not to depend on training corpus as with machine learning and deep learning approaches.

The majority of rule-based approaches tend to focus on pattern matching. For example, dictionary approaches, that given a text and a lexicon with the terms of interest, perform string matching between text and terms [Couto and Lamurias, 2018].

2.1.1 Machine learning

Machine learning (ML) is the process used by systems to learn tasks automatically. This learning is acquired through training and validation data to make predictions in test data. In the text mining domain, these data can correspond to annotated corpora [Lamurias and Couto, 2019]. There are two learning paradigms for ML models: supervised and unsupervised. In supervised learning, the labels of each instance of the training data are known, and in unsupervised learning, the training data is not labeled. Models such as Support Vector Machine (SVM) and Conditional Random Field (CRF) use supervised learning, while the Hidden Markov Models (HMM) use both supervised and unsupervised learning.

SVM is used in classification and regression problems. SVM aims to find the boundary separating different classes and has a maximal distance from any point on the training data, allowing SVM to have better classifications than other models [Manning et al., 2008; Li et al., 2009]. For example, in SVM and other machine learning and deep learning methods that deal with text, these points are vectors represen-

tations of words called word embeddings. Hence, SVM enables text classification, used in text mining tasks such as NER, NEL, and POS (Part-of-speech) tagging.

HMM computes a probability distribution for a sequence of variables given an observed sequence. The idea of this model is that the variables of the same sequence are independent of each other [Beal et al., 2001]. HMM is used in POS tagging and NER.

CRF is a probabilistic method for labeling and segmenting sequential data [Wallach, 2004]. CRF moderates strong independence assumptions between variables made by other probabilistic models like HMM, [Lafferty et al., 2001] which makes it appropriate for the NER task since the meaning of words is heavily dependent on the context.

2.1.2 Deep learning

Deep learning is a form of machine learning that enables computational models to learn data representations with multiple levels of abstraction [LeCun et al., 2015]. These methods are based on artificial neural networks (ANNs) with multiple hidden layers and can be used in speech and audio recognition, object detection, bioinformatics, NLP, and other domains. ANNs are computational networks inspired by biological neural networks and are composed of processing units called nodes comparable to neurons in the biological neural networks. These nodes can be organized into three different types of layers: the input layers, the hidden layers, and the output layers. ANNs architectures include:

- **Recurrent Neural Network (RNN)**. RNN is an ANN where it occurs a recurrent connection in the nodes during a time period. This neural network uses its memory to process sequences of inputs which makes it applicable to tasks with sequential data such as text mining. Long short-term memory (LSTM) is a Recurrent Neural Network capable of handling long-term dependencies. When two LSTM layers are processing a sequence in different directions, the resulting architecture is called Bidirectional LSTM (BiLSTM). A BiLSTM is a combination of forward and backward language models. The forward language model predicts the next token given the current one, and the backward language model does the inverse and predicts the previous token given the current one. Hence, BiLSTM allows the use of future and past input features in a given time frame and has shown better results when compared to a simple LSTM in text mining [Huang et al., 2015].

- **Convolutional Neural Network (CNN or ConvNET)**. CNN is a multilayer node usually used to process images. This neural network has a light pre-processing step due to its ability to learn from filters, while in other neural networks, the filters are hand-engineered. Even though this neural network was originally designed to deal with images, CNN has been used in text applications in recent years [Hu et al., 2014]. CNN can be used for linking biomedical entities [Deng et al., 2019] and ranking them according to their semantic information [Li et al., 2017].

2.2 Named Entity Recognition

NER corresponds to the identification of entities in text. In the biomedical domain, these entities can represent diseases, genes, proteins, among others. Therefore, finding and classifying entities is one of the most fundamental tasks in text mining, not only because it is the first to be executed but also because its output will condition the performance of subsequent tasks.

Usually, NER starts by splitting the text into smaller pieces called tokens. This process is called tokenization. The main challenge of tokenization is to determine the target of the splits. The simplest solution is to delimit the tokens by spaces and punctuation. For example, in the expression "Attention deficit hyperactivity disorder" the tokenization process will separate the words "attention", "deficit", "hyperactivity," and "disorder" which in this case were meant to be a single entity and refer to the condition known as ADHD. This problem is compounded in a multilingual setting where each language has its own lexical characteristics, but several of the recent tokenizers are designed to resolve these issues [Cruz Díaz and Maña López, 2015]. According to its definition and context, the words resulting from the tokenization are assigned to a POS label. The state-of-the-art NER approaches in the biomedical field use deep learning and the BiLSTM-CRF architecture [Huang et al., 2015] and pre-trained language models. The BiLSTM-CRF architecture (Figure 2.1) is defined by the use of past and future input features by the BiLSTM and the use of sentence-level tag information due to CRF layer [Hu et al., 2014].

It is essential to have manually annotated entity datasets called gold standards to train and evaluate NER systems. However, since it is difficult to generate manual annotations, the datasets are not available in some biomedical domains, and the available ones can be limited in size [Crichton et al., 2017].

2.2.1 Pre-trained language models

Pre-trained language models are trained over extensive corpora using a given training task. The resulting word representations are then used on different tasks and corpora, a process designated by transfer learning [Edunov et al., 2019]. This is useful since it allows the reuse of representation without the need for training from scratch. Pre-trained language models can be divided into two generations depending on the type of word embedding. The first generation of pre-trained language models includes Word2vec and GloVe, which learn non-contextual embeddings, and the second generation includes CoVe, ELMo, and BERT that learn from contextual embeddings [Qiu et al., 2020]. For instance, in the sentences "John is in a prison cell" and "Organism composed by a single cell", the meaning of the word cell is different in the two cases. A second-generation model understands that the context in two sentences is different and does not use the same embedding for the word "cell" in these two cases.

2.2.1.1 Non-Contextual Embeddings Models

Word2Vec [Mikolov et al., 2013b] is one of the simplest and most used pre-trained language methods. It predicts words using two different methods: CBOW and Skip-gram. The CBOW model predicts the cur-

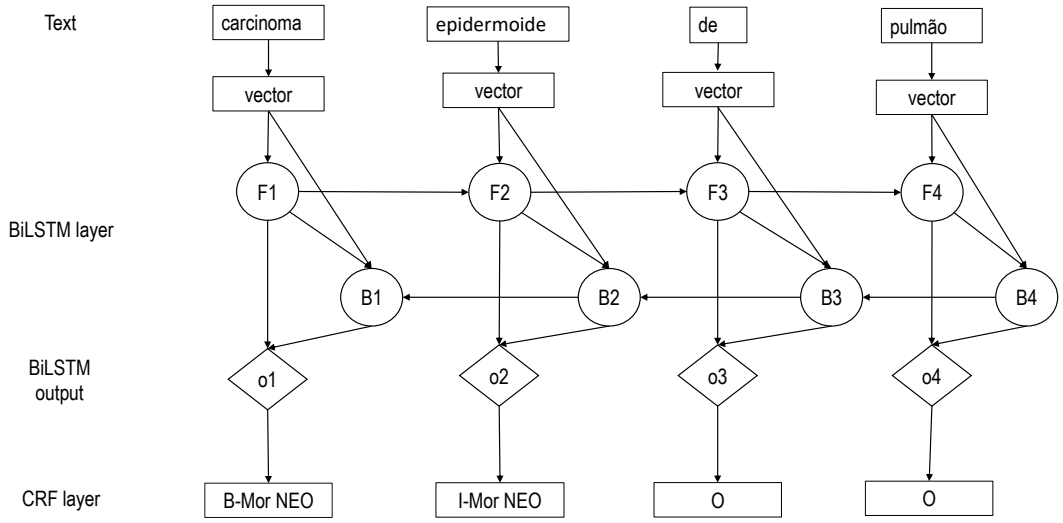

Figure 2.1: BiLSTM architecture. The BiLSTM layer comprises F nodes corresponding to the forward LSTM and the B nodes, the backward LSTM. CRF layer is used for POS tagging. B stands for the beginning of the entity, I for the inside of the entity, and O for out of the entity. In this case, the entity is tagged with the label "Mor NEO" (*Morfología de neoplasias)*. This image was adapted from [Ji et al., 2019].

rent word based on the surrounding words. The Skip-gram makes the opposite, it predicts the surrounding words given the current word [Mikolov et al., 2013a]. Word2vec has an extension named FastText with a variation of this approach. In FastText, each word is represented as a set of character n-grams. The embedding is done in each character n-grams, and the words are represented as the sum of these embeddings [Bojanowski et al., 2017].

GloVe developed by [Pennington et al., 2014] is trained on global word-word co-occurrence on a corpus. The resulting model produces linear substructures of the word vector space. This model outperforms Word2vec in NER tasks.

2.2.1.2 Contextual Embeddings Models

Unlike first-generation pre-trained language methods, ELMo uses the entire input sentence function and computes it on top of a BiLSTM (Figure 2.2) [Peters et al., 2018]. As a result, ELMo can be added to existing models to improve text mining tasks like NER and question answering.

BERT [Devlin et al., 2018] has a multilayer biredirectional transformer encoder architecture (Figure 2.3). The Transformer architecture is based on attention mechanisms to draw dependencies between input and output [Vaswani et al., 2017]. As described by [Devlin et al., 2018], BERT's implementation

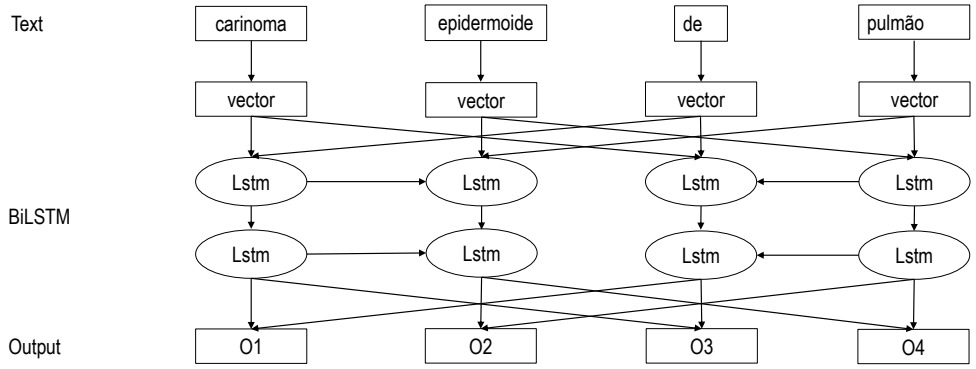

Figure 2.2: Pre-training of ELMo. ELMo uses a BiLSTM in pre-training and a feature-based approach for prediction. This image was adapted from [Devlin et al., 2018]

is composed of two essential steps: pre-training and fine-tuning. First, BERT is pre-trained on unlabeled data using two procedures:

- **MASK LM**: it masks some percentage of the input tokens at random and predicts those tokens. This procedure is made with the intent of creating a deep bidirectional model in training.
- **Next Sequence Prediction**: responsible for making the model identify relationships between sentences which is fundamental in tasks like question answering.

At first, BERT used BooksCorpus and Wikipedia for pre-training, but variations trained on biomedical and scientific corpora emerged, like SciBERT, BioBERT, and ClinicalBERT. These models have shown significant improvements when compared to traditional BERT in text mining tasks in the scientific domain [Lee et al., 2020; Alsentzer et al., 2019]. In addition to domain-specific models, multilingual models have also emerged, such as BETO, trained on a Spanish corpus, and BioBERTpt, trained on Portuguese corpus related to the biomedical field.

In the fine-tuning step, the model starts with pre-trained parameters that will be fine-tuned for a specific task like NER. This is the main difference between BERT and ELMo [Devlin et al., 2018]. After pre-training the model, ELMo uses an architecture-specific for each task [Peters et al., 2018].

2.3 Named Entity Linking

NEL, also called normalization or disambiguation, corresponds to the text mining task of linking mentions in a text to entries in a KB. This task is important because an entity can be mentioned in many different

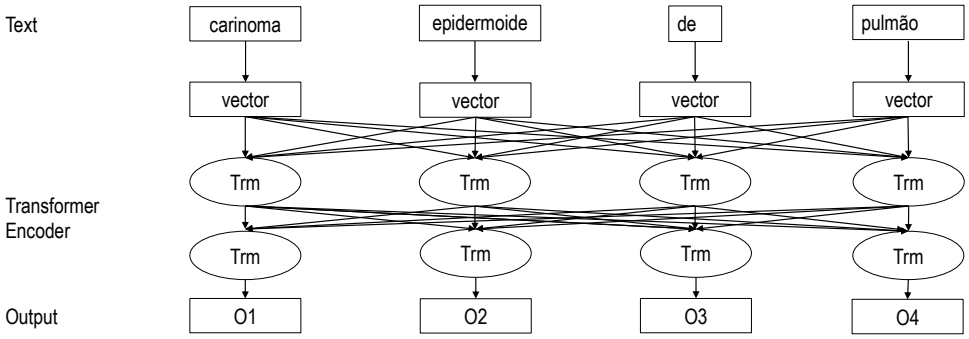

Figure 2.3: Pre-training of BERT. BERT uses the Transformer Encoder in pre-training. This image was adapted from [Devlin et al., 2018]

ways, and the same name can be used to describe different entities [Shen et al., 2014]. Thus NEL allows the identification of name variations of a certain entity and distinction in cases of ambiguity.

A KB contains the definition, the description, and other relevant information for each containing entity [Zheng et al., 2010]. DBpedia, for example, is a multilingual KB that was built with information extracted from Wikipedia [Lehmann et al., 2015] and it is one of the most important KBs available since it covers a variety of domains such as companies, geographic information, people and scientific publications [Bizer et al., 2009]. However, there are also KBs designed especially for the biomedical domain. As will be explained in detail in Section 2.4, these KBs can be in the form of a thesaurus, vocabularies, and ontologies.

2.3.1 Graph Models

As previously mentioned, NEL's state-of-the-art approaches include graph-based models. A graph model is composed of a set of random variables and a graph. Each node of a graph is associated with one of the random variables, and the edges express the dependence between the random variables [Scutari and Strimmer, 2010].

PageRank was the algorithm originally used by Google to rank webpages in their search engine [Page et al., 1999]. In this algorithm, the web is considered a graph, where each graph's node is a different webpage. In addition, each webpage has links for other pages called forward links and links from other pages called backlinks. The output of this algorithm is the probability distribution of getting to a webpage after several iterations.

The PPR, a variation of the PageRank algorithm, is being used in NEL [Lamurias et al., 2019]. The

difference between this algorithm and the original PageRank is that the crossing of the graph is not random. In PPR, it goes from a node to a chosen forward link.

2.4 Biomedical Knowledge Bases

KBs gathers all variations of the same term, eliminating ambiguity cases and controlling synonyms that facilitate the indexing of entity mentions and the retrieval of information through browsing and searching [Gudivada et al., 2018; Zeng, 2008]. For text mining, it is important that KBs also can support semantic relations between concepts because these relations can be important for text mining tasks. Examples of KBs are:

- **Thesaurus**: a thesaurus usually includes the definition of concepts and the relationships between them. It coordinates processes of indexing and document retrieval [Frakes and Baeza-Yates, 1992].
- **Subject Headings**: a subject heading is a list of words or phrases used to describe a topic of texts in books, articles, and other documents and link them to other texts with similar subjects [Harpring, 2010]
- **Taxonomies**: a taxonomy is a hierarchical classification for a specific topic. In taxonomy, a term has one or more parent/child relationships to other terms, which is a simpler structure when compared to a thesaurus [Harpring, 2010]. Taxonomy is usually used to classify organisms, but it can also be used in the health domain.
- **Ontologies**: As described by [Borst, 1999], an ontology is a formal, explicit specification of a shared conceptualization. That is, an ontology concept is defined explicitly and formally in the same way by a group of people. Ontologies are defined to be understood by humans and by computers. The main language to define ontologies is the Web Ontology Language (OWL). In the biomedical field, a group of developers designated as the Open Biomedical Ontology (OBO)[1] foundry that came together to develop and define a set of principles to be followed to describe ontologies. The main principles are:

    1. The ontologies must be open to being used, and its origin have to be identified and not altered;
    2. All the ontologies are defined in a common shared syntax;
    3. Each property in an ontology such as a relation must have a unique URI identifier;
    4. The ontology developer has procedures to identify different versions;
    5. The ontology has a distinctly scope and content;
    6. The ontology has textual definitions for the majority of its terms;

[1] http://www.obofoundry.org/

7. The ontology uses relations defined by the Relations Ontology (RO);

OBO Foundry's ontologies include the Gene Ontology (GO) to annotate genes, genes, products, and sequences with concepts related to their functions [Consortium, 2008], the Chemical Entities of Biological Interest (ChEBI) to annotate small chemical compounds involved in biological process [Degtyarenko et al., 2007] and the Disease Ontology (DO) to provide definitions about diseases, genotypes, phenotypes, proteomes, epitopes and drugs [Kibbe et al., 2015].

2.4.1 English KBs

The Unified Medical Language System (UMLS) is an example of developing a thesaurus in the biomedical field. UMLS was developed by the National Library of Medicine (NLM) and consists of biomedical vocabularies designed to help health professionals and researchers retrieve and integrate biomedical information such as electronic health records from different sources [Bodenreider, 2004]. UMLS comprises three different knowledge sources: Metathesaurus, Semantic Network, and SPECIALIST Lexicon and Lexical Tools. The biggest component of UMLS is the Metathesaurus. The Metathesaurus[2] is a large multilingual biomedical thesaurus composed of nearly two hundred different vocabularies from several sources. One of the main goals of Metathesaurus is to understand the meaning of each word in a vocabulary and link all the words that have the same meaning. Therefore the Metathesaurus is organized by concepts. Metathesaurus assigns each concept and all the related words unique identifiers and points out relations between concepts.

Created by the National Library of Medicine (NLM) in 1960, the Medical Subject Headings (MeSH) was designed to index, catalog, and search health-related information in MEDLINE/PubMed, books, journals, and NLM's catalog [Nelson et al., 2001].

2.4.2 Multilingual KBs

SNOMED CT[3] is a multilingual clinical healthcare terminology. SNOMED CT enables the consistent representation of clinical content in electronic health records providing an automatic way to interpret them and better patient care.

DeCS[4] was developed from MeSH. DeCS is a multilingual vocabulary created by the Latin American and Caribbean Center on Health Sciences Information (BIREME) to allow common terminology for searching in three languages: English, Portuguese and Spanish. In addition to the health-related terms provided by MeSH, DeCS has terms from other four topics: Public Health, Homeopathy, Science and Health, and Health Surveillance.

[2] https://www.ncbi.nlm.nih.gov/books/NBK9684
[3] snomed.org
[4] https://decs.bvsalud.org/en/about-decs/

Created by the World Health Organization (WHO), the International Classification of Diseases (ICD)[5] classifies diseases, disorders, injuries, medical procedures, and other related conditions for clinical and research purposes [Dodd et al., 2018]. ICD is revised periodically and has domain-specific extensions. For example, ICD-O-3 corresponds to the third edition of the International Classification of Diseases for Oncology. This taxonomy allows easy storage, retrieval analysis, and sharing of health information. ICD also has versions in forty-two languages, such as the CIE in Spanish and CID in Portuguese.

The Human Phenotype Ontology (HPO)[6] provides phenotype vocabulary and disease-phenotype annotations. HPO is the standard system of phenotypic information by various groups such as international rare disease organizations, registries, clinical labs, biomedical resources, and clinical software tools [Köhler et al., 2017]. Currently, HPO has more than 13,000 terms that are connected by is-a relationships. HPO is an ongoing project that aims to translate all of these terms, synonyms, and textual definitions into various languages: Chinese, Dutch, French, German, Italian, Japanese, Portuguese, Russian, Spanish, and Turkish.

2.5 Shared tasks

In recent years a series of shared tasks to evaluate state-of-the-art text mining solutions for clinical text in Spanish has emerged. This dissertation's work contributed to the participation in three of these shared tasks: ProfNER, MESINESP2, and CANTEMIST.

The ProfNER shared task, whose goal is the identification of professions and occupations in Health-related tweets in Spanish, is organized by the "Social Media Mining for Health Applications (#SMM4H) Shared Task 2021". It includes two sub-tasks:

- **Track A – Tweet binary classification:** to determine if a tweet has a mention of occupation or not;
- **Track B - NER offset detection and classification:** to recognize the span of mentions of occupations and to classify them in the respective category ("PROFESION" or "SITUACION LABORAL").

The MESINESP[7] task organized by the BioASQ that consists in the indexing of documents from the "Índice Bibliográfico Español en Ciencias de la Salud" (IBECS) and "Literatura Latino-Americana e do Caribe em Ciências da Saúde" (LILACS) with DeCS terms. The second edition, MESINESP2, includes the following sub-tasks:

- **MESINESP-L – Scientific Literature:** indexing with DeCS terms of Spanish abstracts from two databases, IBECS, and LILACS;

[5] https://www.who.int/classifications/icd/en/
[6] https://hpo.jax.org
[7] https://temu.bsc.es/mesinesp/

- **MESINESP-T - Clinical Trials:** indexing with DeCS terms of Spanish clinical trials from REEC (Registro Español de Estudios Clínicos);
- **MESINESP-P – Patents:** indexing with DeCS terms Spanish patents extracted from Google Patents.

The CANTEMIST is the only task related to tumor morphology. CANTEMIST includes three subtasks:

- **CANTEMIST-NER:** identification of tumor entities;
- **CANTEMIST-NORM:** indexing the identified entity to the corresponded CIE-O term;
- **CANTEMIST-Coding:** provide for each document a ranked list of its corresponding CIE-O codes. In addition to the developed tools, the clinical texts in Spanish provided by these tasks can be used to train future applications.

In addition to these shared tasks, the Clinical Case Coding in Spanish Shared Task (CodiEsp)[8] where the participants are asked to index clinical documents to CIE vocabulary and the PharmaCoNER[9] task that seeks to recognize chemical and proteins entities [Sun and Yang, 2019] are other examples of text mining shared tasks for the Spanish biomedical domain.

2.6 Multilingual Corpora

As mentioned throughout this work, the number of text mining resources for non-English languages, such as corpora, is reduced. Still, it is possible to find some cases of English parallel corpora with another language. Examples of these corpora are the BVS corpus [Soares and Krallinger, 2019] and the Multilingual Radiology Research Articles Dataset (MRRAD) [Campos et al., 2017]. The BVS corpus comprises abstracts from the BVS database, a database with biomedical information about Latin America and Carib created by BIREME in Portuguese, Spanish, and English. MRRAD is a parallel corpus of Portuguese research articles related to radiology and alternative translations in English.

For the corpora exclusively in Spanish, some of the corpora available come from text mining competitions like those described in the Section 2.6. In addition to those corpora, there are also instances like the DrugSemantics [Moreno et al., 2017], a corpus for NER related to the pharmacotherapeutic domain.

The amount of corpora exclusively in Portuguese is scarcer than those available in Spanish, and most of them are not available to be tested. One of the few examples is SemClinBr, an annotated corpus for clinical text mining tasks in Brazilian Portuguese with 1,000 clinical notes, labeled with 65,117 entities and 11,263 relations [Peters et al., 2020]. Another example is the HSL dataset, which contains information

[8] https://temu.bsc.es/codiesp/
[9] https://temu.bsc.es/pharmaconer/

about the patients from the Syrian-Lebanese Hospital (HSL), also in Brazilian Portuguese [Reys et al., 2020]. These corpora are not open-access which makes it difficult to replicate the results obtained.

2.7 NER and NEL tools

MER stands out for its simplicity and efficiency among the specifically developed systems to work in biomedical text. MER is a user-friendly NER and NEL tool that only needs an ontology or a list of terms representing the entities, their identifiers, and a Unix shell. This simple tool takes advantage of the grep and awk commands for text processing [Couto and Lamurias, 2018].

In addition to MER, other systems are designed for the NER task, such as the default NER model provided by the Flair framework and the BioBERTpt model. The NER model provided by Flair is an example of a system that uses deep learning methods for the NER task. Flair allows the use of several word embeddings combinations to design a single model without additional engineering effort [Akbik et al., 2019]. Besides using word embeddings from other pre-trained models such as FastText, ELMo, or BERT, Flair has its own contextual word embeddings. The difference between these contextual embeddings and others is that they are trained without any explicit notion of words, and words are treated just as sequences of characters [Akbik et al., 2018]. The NER model provided by Flair has been applied to Spanish biomedical text [Akhtyamova et al., 2020]. On the other hand, BioBERTpt uses the pre-trained language model BERT trained on Portuguese clinical and biomedical corpora for NER in clinical and biomedical Portuguese text [Schneider et al., 2020].

There are also systems specifically for the NEL task, like the PPR-SSM [Lamurias et al., 2019]. PPR-SSM is based on the PPR algorithm and uses the semantic similarity between the candidates for each entity and their information content (IC) to improve the results of the NEL task. In the PPR-SSM, candidates with high semantic similarity with the other nodes and a high IC have a better ranking in the disambiguation graph. The IC is the frequency of the presence of an entity in a corpus. The semantic similarity measurement (SSM) corresponds to the similarity value between entities using the relations defined in KBs. The Resnik is one of the metrics used to measure the semantic similarity between entities [Resnik, 1995]. This metric is defined as:

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

In which, $IC_{shared(e1,e2)}$ is the IC of most informative common ancestor of the entities $e_1$and e_2.

For the NER and NEL systems designed for languages other than English, shared tasks have been shown to play a key role. The systems derived from these shared tasks have obtained promising results NER and NEL when applied to biomedical entities. For example, in CANTEMIST, the systems with the best results have used pre-trained models, more specifically BERT and some of its variations [García-Pablos et al., 2020; Xionga et al., 2020].

2.8 Evaluation

The evaluation of NERL systems is made by comparing the outputted predictions against the manual annotations done by experts, the gold standard test. The metrics normally used in the evaluation are **precision (P)**, **recall (R)**, and **F1-score**. These metrics are calculated by the following instances:

- **True positives (TP):** number of correctly identified predictions. The model identifies an entity that is present in the gold standard.
- **False positives (FP):** number of incorrect predictions classified as positive. The model identifies an entity that is not present in the gold standard or incorrectly identifies an entity that is present in the gold standard.
- **False negatives (FN):** number of incorrect predictions classified as negative. The model does not identify an entity that is present in the gold standard.

The precision is the percentage of instances that the system detected that are in fact positive. The recall is the percentage of instances that were correctly identified [Jurafsky and Martin, 2009]. For example, in the NER task, if a system identifies 100 entities in the text and from those 100, only 80 are in the gold standard, that model has 80% precision. On the other hand, if a system identifies 80 correct entities of the 100 present in the gold standard, the model has 80% recall [Campos, 2017]. F1-score is the harmonic mean of recall and precision. This metric is used to compare the efficiency of different systems.

There are two methods to evaluate a system's performance: the micro-average and the macro-average. The micro-average corresponds to the sum of all error types of all documents to make the average of each metric. The macro-average consists of calculating the Precision, Recall, and F1-score for each document and then make the average for all documents. Since micro-average weighs each instance separately, it will capture the imbalance between the documents. Therefore the ICERL system uses it as the evaluation method. On the other hand, if all documents are equally important, the macro-average is more suitable [Jurafsky and Martin, 2009].

$$Micro - average \quad precision = \frac{TP1 + TP2 + ... + TPn}{TP1 + TP2 + ... + TPn + FP1 + FP2... + FPn}$$

$$Micro - average \quad recall = \frac{TP1 + TP2 + ... + TPn}{TP1 + TP2 + ... + TPn + FN1 + FN2 + ... + FNn}$$

$$Macro - average \quad precision = \frac{P1 + P2 + ... + Pn}{n}$$

$$Macro - average \quad recall = \frac{R1 + R2 + ... + Rn}{n}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Chapter 3

CANTEMIST pipeline

As previously mentioned, the first phase of this dissertation consisted of developing a NERL pipeline for the Spanish language. It was intended for this pipeline to have a performance similar to the state-of-the-art achieved at the CANTEMIST shared task: an F1-score equal or higher than 0.87 and 0.83 for the NER and NEL tasks, respectively. The performance of this pipeline and the other approaches developed for this competition was evaluated by comparing the annotations generated with the annotations manually generated by experts [Miranda-Escalada et al., 2020].

The starting point to develop this pipeline was the approach developed by the LasigeBioTM team for the CANTEMIST shared task (Figure 3.1) [Ruas et al., 2020]. LasigeBioTM used the Flair framework for the training of embeddings in Spanish PubMed abstracts. The resulting embeddings were integrated into a NER model, which was then trained on the training and development sets of the CANTEMIST corpus. Next, the trained model was applied to the CANTEMIST corpus test set to obtain the annotation files with the recognized entities. The next step was to perform NEL, more concretely, by searching candidates for the recognized entities. This search was made in CIE-O-3 and also in two other vocabularies, DeCS and ICD-10. Finally, a disambiguation graph was constructed with these candidates. Then, the candidates were ranked through the PPR-SSM algorithm, and the best candidate was linked to the entity mentioned in the text.

3.1 NER methods

LasigeBioTM used the Flair framework to developed four NER models: ”base”, ”large”, ”medium” and ”pubmed” [Ruas et al., 2020]. Based on these four models, a fifth model was developed, ”medium 2.0”, corresponding to an improved version of the model with the best results. Since the best approaches developed for the CANTEMIST-NER sub-task used BERT [Miranda-Escalada et al., 2020], other systems based on this pre-trained language model were also developed.

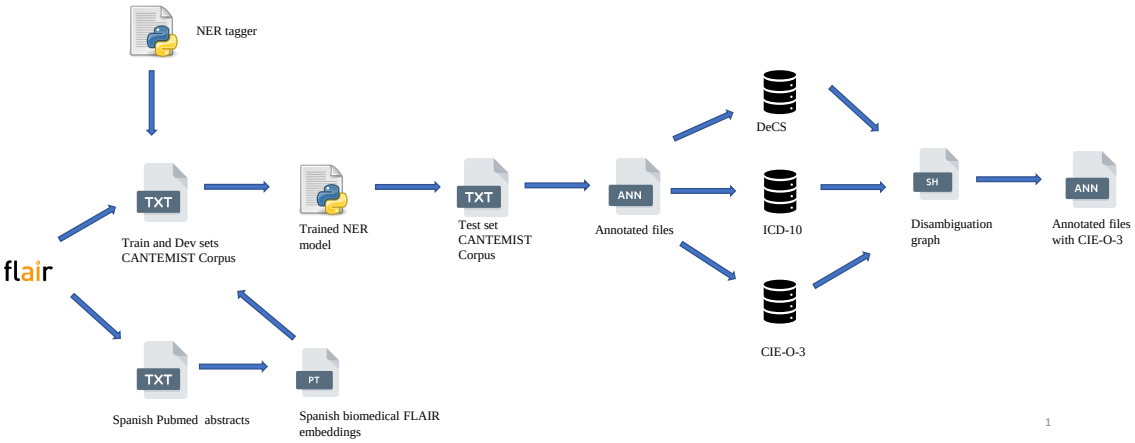

Figure 3.1: LasigeBioTM procedure for the CANTEMIST shared task.

3.1.1 NER models

- **"base":** it includes Flair embeddings (es-forward and es-backward) trained on Spanish Wikipedia and Spanish FastText embeddings.
- **"large":** it includes Flair embeddings (es-forward and es-backward) trained on Spanish Wikipedia, Spanish FastText embeddings, and PubMed Flair embeddings trained on two PubMed splits.
- **"medium":** it includes Flair embeddings (es-forward and es-backward) trained on Spanish Wikipedia, Spanish FastText embeddings, and PubMed Flair embeddings trained on one PubMed split.
- **"pubmed":** it includes PubMed Flair embeddings trained on four PubMed splits.
- **"medium 2.0":** it includes Flair embeddings (es-forward and es-backward) trained on Spanish Wikipedia, PubMed Flair embeddings trained on one PubMed split, Spanish FastText embeddings, Spanish Bytepair embeddings, and Character embeddings.
- **"medium+bert":** it includes Flair embeddings (es-forward and es-backward) trained on Spanish Wikipedia, Spanish FastText embeddings, PubMed Flair embeddings trained on one PubMed split, and BERT embeddings trained on Wikipedia.
- **"bert":** it includes Spanish FastText embeddings and BERT embeddings trained on Wikipedia.
- **"scibert":** it includes Spanish FastText embeddings and BERT embeddings trained on documents from Semantic Scholar.

- **"beto":** it includes BERT embeddings trained on Spanish Corpora composed by several sources such as Wikipedia.
- **"scibert+beto":** it includes BERT embeddings trained on documents from Semantic Scholar and other sources.

3.1.2 Training setup

LasigeBioTM trained the Flair embeddings in translated abstracts of PubMed articles [Ruas et al., 2020]. These 130,000 articles with 155,366,645 tokens were divided in four splits to optimize the training process, which one with 80%/10%/10% of the articles in the train, validation and test files, respectively:

1. 32,500 articles, 40,987,614 tokens
2. 32,500 articles, 35,352,727 tokens
3. 32,500 articles, 39,021,229 tokens
4. 32,500 articles, 40,005,075 tokens

The embeddings were trained forward and backward, and the parameters used to create them are described in Table 3.1. For the NER tagger, all the models use Flair's default architecture, BiLSTM with a CRF decoding layer, and the training parameters are described in Table 4.5. The NER phase training was done using one NVIDIA Tesla P4 and one NVIDIA Tesla M10 GPU.

Table 3.1: Flair embeddings training parameters of the CANTEMIST pipeline.

Models	**Hidden size**	**Nlayers**	**Dropout**	**Sequence length**	**Mini batch size**	**Max epochs**	**Patience**
LasigeBioTM	1024	1	0.1	250	32	100	25
medium 2.0	1024	1	0.1	250	32	200	25

Table 3.2: Training parameters for NER models of the CANTEMIST pipeline.

Models	**Hidden size**	**Learning rate**	**Mini batch size**	**Max epochs**	**Patience**
LasigeBioTM	256	0.1	32	55/150	3
medium 2.0	256	0.1	32	150	3
BERT models	256	0.1	8	55/150	3

3.2 NEL methods

After generating the output files of the NER step, the ten best candidates from CIE-O-3, the five best candidates from ICD-10, and the five best candidates from DeCS were retrieved by string matching. These candidates were used to build a disambiguation graph in which the PPR-SSM is applied. As previously mentioned, the PPR-SSM uses the IC and the semantic similarity to rank each candidate for each entity. In this case, the IC corresponds to the frequency of each candidate in the training and development of the CANTEMIST corpus, and the semantic similarity is related to the CIE-O-3 hierarchy. This hierarchy is used to make the edges between candidates in the disambiguation graph; two candidates are considered linked in the graph if they are also linked in the CIE-O-3.

Only the best candidate of CIE-O-3 is selected in the NEL task. The role of the non-CIE-O-3 candidates is to create a bigger disambiguation graph with more semantic information, which was expected to improve the precision of the disambiguation [Ruas et al., 2020]. Thus, LasigeBioTM developed two models for NEL:

- **"single ont:"** only uses candidates from CIE-O-3
- **"multi ont:"** uses candidates from CIE-O-3, ICD-10 and DeCS

Both models had flaws in the string matching steps, such as uppercase and lowercase letters and space between words which were corrected. These corrections resulted in models in which the search for the best candidate is first done through a dictionary method and only through PPR-SSM if this approach does not work. In other words, if an exact match candidate is found in the CIE-O-3 list by string matching, the model will link the entity to that candidate. On the other hand, if there is no exact match, the model applies the PPR-SSM algorithm to select the best candidate. This method was inspired by SINAI's approach, one of the teams that participated in the CANTEMIST shared task [López-Úbedaa et al., 2020]. Like the SINAI's approach, the search for candidates is done in the list of CIE-O provided by the CANTEMIST authors and the complete list of CIE-O-3 and the train and development set CANTEMIST corpus.

3.3 Results and discussion

This work started by replicating the "medium" model developed by LasigeBioTM, and similar results were obtained (F1-score of 0.741 and 0.061 for the NER and NEL task, respectively). Due to a lack of time, LasigeBioTM only used this model at the CANTEMIST shared task, so I tested the other three after replicating this model. Since none had a better performance (Table 3.3), the improvements were only made in the "medium" model. These improvements refer to the usage of BytePair and Character embeddings, the training of the NER tagger up to 150 epochs, and training of the PubMed embeddings up to 200 epochs. The introduction of BytePair and Character embeddings has proved to be efficient for recognizing biomedical entities in Spanish [Akhtyamova et al., 2020]. As can be seen from the Table

3.3, these embeddings and the re-training of PubMed's embeddings increased the performance of the "medium" model. Since the improvement was small, models that use the pre-trained language model BERT were also evaluated. However, as shown in Table 3.3, the performance of these models was worse than the "medium" and the state-of-the-art achieved at the CANTEMIST [Miranda-Escalada et al., 2020]. There are several possible reasons behind this decrease. The first one is the architecture used. All the models I developed for NER were built on Flair, which means they all feature a BiLSTM-CRF architecture. Unlike the models developed by the other teams that used other frameworks and other architectures [García-Pablos et al., 2020; Xionga et al., 2020]. In addition to that, due to incompatibilities between BERT and the Flair framework, it was necessary to reduce the NER taggers' batch size. Since BERT has a token length limitation of 512, I cropped the sentences in each document up to that limit which may have resulted in a loss of information and consequently a decrease in performance.

Table 3.3: Performance of tested models for CANTEMIST-NER.

Models	**NER tagger trained up to 55 epochs (F1-Score)**	**NER tagger trained up to 150 epochs (F1-Score)**
medium	0.741	0.753
base	0.710	—
pubmed	0.736	—
medium + bert	0.642	—
bert	0.68	—
scibert	0.598	—
beto	0.60	—
scibert + beto	0.642	—
medium 2.0	**0.743**	**0.754**

The modifications on the NEL task were effective since, as is described in Table 3.4, the model "single ont" and "multi ont" achieved an F1-Score of 0.664 and 0.665. However, these results are still far from expected, which may be related to the low performance of the NER step. To test this hypothesis, these models were applied to the training and development sets from the NER step and used NEL's training and development sets to assess the performance. As can be seen from the Table 3.5, this hypothesis is confirmed because the evaluation metrics were close to 100%. Therefore, it is possible to consider that the results obtained are similar to those obtained by the state-of-the-art of the CANTEMIST shared task.

Table 3.4: Performance of improved models for CANTEMIST-NORM.

Models	**Precision**	**Recall**	**F1- score**
single ont	0.690	0.639	0.664
multi ont	**0.691**	**0.641**	**0.665**

Table 3.5: Performance of improved ”multi ont” model on training and development sets.

	Precision	**Recall**	**F1- score**
train set	0.974	0.974	0.974
dev set	0.971	0.971	0.971

Chapter 4

Iberian corpus (ICR) and NERL system (ICERL)

After improving the CANTEMIST pipeline, the next phases of this dissertation consisted of creating the ICR corpus and the ICERL system. The first section of this chapter presents the methodology used to retrieve the files that composed the ICR corpus and a description of their annotation. In addition, this section also describes the methodology used to retrieve the files used for the training of the Portuguese pipeline. The second part of this chapter describes the adaptation of improved pipeline to the final version of the ICERL system, the methods used to evaluate the system, the results and discussion of the application of the ICERL system on the ICR corpus, and how the solutions found for the ICERL system were used at ProfNER and MESINESP2 shared tasks.

4.1 ICR corpus

As shown in Table 4.1, the ICR corpus is composed of 1,555 abstracts for each language. The average and maximum length of these abstracts and the total number of annotations are similar between the two languages. Furthermore, the ICR corpus presents the same most frequent annotation in both languages. The main difference between the texts in Portuguese and Spanish refer to the number of occurrences of the most frequent annotation. These annotations were done by applying the final version of the ICERL system on the ICR corpus. The reason for the annotations differences between the two languages is going to be explained in detail in Section 4.2.

To train the Portuguese NER model, 974 documents from SciELO and 41 from PubMed were retrieved (Table 4.2). The query used to retrieve the SciELO files was the same that was used for the ICR corpus; therefore, a filter was applied to the training files to ensure that none of the files was also in the ICR corpus. Since the number of training files is low when compared to the ICR corpus, nlpaug[1], a Python library for data augmentation was used to increase the number of training files. This tool replaced

[1] https://github.com/makcedward/nlpaug

Table 4.1: Description of the ICR corpus.

	Spanish	**Portuguese**
Number of documents	1,555	1,555
Average text length	1,227	1,172
Max text length	2,744	2,850
Total annotations	3,399	3,287
Unique annotations	216	171
Most frequent annotation	cáncer	câncer
Number of occurrences of the most frequent annotation	2,553	1,880

the words from the training files with the respective synonyms from WordNet. For example, in some files, the word ”câncer” was replaced with ”câncro”. For each file, two other files were created using nlpaug, thus making 3,045 documents for training. These files were then annotated using the Spanish pipeline of the ICERL system. As the CANTEMIST corpus, the annotations on these documents follow the IOB format and use the tag ”MORFOLOGIA_NEOPLASIA” (”MOR_NEO”). Therefore, each token from the training files was tagged with the label “B-MOR_NEO” if it is the beginning of annotation, the label “I-MOR_NEO” if it is the inside of an annotation, and the label “O” if it is the outside of an annotation (Table 4.3). These annotations were subjected to manual corrections regarding the tokenization and labeling process, for example, in some cases the entity ”tumor” was split into two tokens (”tum” and ”r”) and entities such as ”cáncer de boca” were labeled:

- ”cáncer” -> “B-MOR_NEO”
- ”de” -> ”O”
- ”boca” -> ”O”

when it should be labeled:

- ”cáncer” -> “B-MOR_NEO”
- ”de” -> ”I-MOR_NEO”
- ”boca” -> ”I-MOR_NEO”

In addition to the ICR corpus and the training files for the Portuguese NER model, articles were retrieved from SciELO to train the Portuguese embeddings. In total, 65,903 articles were retrieved, but only 500 were used due to a lack of time. Furthermore, unlike the ICR corpus and the NER models

Table 4.2: Querys used to retrieve the ICR corpus, the Portuguese files for the NER model training, and Portuguese Files for the embedding training.

	Query		Date
	SciELO	Pubmed	
ICR corpus	*((*) AND(oncology)) OR (cancer))*	-	11/02/2021
Portuguese Files for NER model training	*((*) AND(oncology)) OR (cancer))*	*Case Reports[Publication Type] AND POR[LA] AND Cancer[Filter]*	31/03/2021
Portuguese Files for embedding training	(*)	-	02/04/2021

Table 4.3: Number of tokens of Portuguese training files.

number of ”B-MOR_NEO” tokens	9,814
number of ”I-MOR_NEO” tokens	7,305
number of ”O” tokens	7,755,603
total number of tokens	7,772,722

training files, no filter was used to retrieve these files. This is because it was intended that these files were as similar as possible to the PubMed files used in training Spanish embeddings. Therefore the training files for Portuguese embeddings encompass several biomedical domains.

4.2 ICERL system

The ICERL system comprises two pipelines; one is designed to deal with the Spanish text and the other with the Portuguese. For the NER task, the Spanish pipeline uses the NER model ”medium 2.0”, which was described in Chapter 3, and for the Portuguese pipeline, three models were developed:

- **”cantemistpt”:** it includes Flair embeddings (pt-forward and pt-backward) trained on Portuguese Wikipedia, Portuguese FastText embeddings, and SciELO Flair embeddings, Portuguese Bytepair embeddings, and Character embeddings.
- **”cantemistpt + biobertpt”:** it includes Flair embeddings (pt-forward and pt-backward) trained on Portuguese Wikipedia, Portuguese FastText embeddings, and SciELO Flair embeddings, Portuguese Bytepair embeddings, Character embeddings, and BioBERTpt embeddings.
- **”biobertpt”:** it includes Portuguese FastText embeddings and BioBERTpt embeddings.

The NEL task is the main difference between the ICERL system and the procedure used in the CANTEMIST corpus. Instead of returning the best candidate from CIE-O, the ICERL system returns a list of CIE-O and DeCS for the Spanish texts and CID-O and DeCS for Portuguese texts. This approach was made in order to not restrict the ICERL system to just one candidate and one vocabulary. For example, the entity ”Neoplasia benigna” is in the DeCS vocabulary with the code ”D009369” and in CID with the code ”8000/0”. An ID was also created for all entities found, which identifies the entities by the line and order in which they are found in the line. This ID is used by the evaluation method (Subsection 4.2.2) to compare entities from the two languages.

The Figure 4.1 illustrates the ICERL system. As can be seen, in addition to making the Spanish annotations, the Spanish pipeline is also applied to the Portuguese training files. The resulting annotations and the Portuguese embeddings are used to train a NER model that, together with a NEL model, are applied to the Portuguese texts.

To improve the performance of the ICERL system in Portuguese, the entities found in Portuguese text were expanded. The expansion refers to the replacement of the Portuguese entities by their synonyms in DeCS. To do so, MER was applied in the Portuguese text to retrieve the entities in the text. Only the entities composed by terms present in the CID vocabulary were chosen to filter the entities of the oncological domain. The resulting entities were then replaced by their Portuguese synonyms. For example, in the sentence ”Os bloqueios neurolíticos, para o controle da dor em paciente com tumores cuja possibilidade terapêutica é difícil.” the DeCS synonym of the entity ”tumores”, ”tumores malignos” was added to the original sentence : ”Os bloqueios neurolíticos, para o controle da dor em paciente com tumores / tumores malignos cuja possibilidade terapêutica é difícil.”

4.2.1 Training setup

The training setup of the Spanish pipeline was described in the Subsection 3.1.2. As with the CANTEMIST models, the Portuguese models used a BiLSTM with a CRF decoding layer, and they were trained using one NVIDIA Tesla P4 and one NVIDIA Tesla M10 GPU. The parameters for training the Portuguese embeddings and the models are described in Tables 4.4 and 4.5.

Table 4.4: Portuguese Flair embeddings training parameters.

Hidden size	**Nlayers**	**Dropout**	**Sequence length**	**Mini batch size**	**Max epochs**	**Patience**
1024	1	0.1	250	32	110 foward 118 backward	25

Table 4.5: Training parameters of the Portuguese models.

Models	Hidden size	Learning rate	Mini batch size	Max epochs	Patience
cantemistpt	256	0.1	32	150	3
biobertpt	256	0.1	32	150	3
cantemistpt + biobertpt	256	0.1	16	150	3

4.2.2 Evaluation method

Since the performance of the Spanish pipeline on the CANTEMIST corpus is available, and it is similar to the state-of-the-art, the evaluation method consists of the comparison between the results of the Portuguese and Spanish pipelines. Therefore, the evaluation method considers a TP, if at least one of the candidates of the Portuguese entity is on the candidates of the Spanish entity; an FP, if the Portuguese pipeline does not find an entity which was found by the Spanish pipeline or none of the candidates of the Portuguese entity is on the candidates of the Spanish entity; an FN, if the Spanish pipeline does not find an entity and the Portuguese pipeline does (Figure 4.2). Thus, the difference between using the expansion for the evaluation method is that the list of Portuguese candidates is composed of several entities instead of just one.

4.2.3 Results and discussion

The Table 4.6 presents the Portuguese NER models and the baseline results. The baseline corresponds to the application of the Spanish pipeline on the Portuguese texts. The baseline's precision can be explained by the high semantic similarity between the two languages. This result demonstrates that the application of the Spanish pipeline in texts in Portuguese is reasonable. However, the recall reveals that some of the entities that the system found in Spanish were not found in Portuguese, which means that it was possible to develop an approach with a better performance. To do so, the three Portuguese NER models were created.

The results of the three models show that the BioBERTpt or a similar pipeline to the one used for Spanish are both valid solutions for the Portuguese texts. However, the performance achieved by these models is related to the manual correction of the annotations of the training files in Portuguese. For example, the recall and F1-score of the ”cantemistpt” model decrease to 0.155 and 0.264, respectively, without the manual correction of the Portuguese training files. Since the three models had the same F1-score, the recall was used to decide which model had the best performance. This metric was chosen at the expense of precision as it expresses the number of entities found in Spanish that were not found in Portuguese. The higher the recall, the lower the number of these entities. Therefore, I considered the ”cantemistpt + biobertpt” model the best performance and the final version of the ICERL system used it

for the Portuguese pipeline.

The NLPStatTest, a toolkit to compare the performance of two NLP systems, was used for the baseline and the final version of the ICERL system. The F1-score in each document was the score used to compare the two systems, and as it is described in Table 4.7 and Figure 4.3, the ICERL system's F1-score is 58 percentage points higher than the baseline. The same toolkit was used to confirm this hypothesis by the student's t-test.

Table 4.6: Performance of the baseline and the Portuguese NER models.

Models	**Precision**	**Recall**	**F1-score**
baseline	0.887	0.171	0.287
cantemistpt	0.874	0.842	**0.858**
cantemistpt + biobertpt	0.873	0.844	**0.858**
biobertpt	**0.898**	0.821	**0.858**

Table 4.7: Statistics of the baseline and ICERL system.

Score	**Mean**	**Median**	**Std. Dev.**	**Minimum**	**Maximum**
ICERL system	**0.708**	**0.703**	**0.121**	**0.369**	**0.954**
baseline	0.121	0.089	0.118	0.000	0.578

In addition to the Portuguese NER models, an expansion of the Portuguese entities was carried out. This approach was made under the hypothesis that the ICERL system did not recognize some Portuguese entities, but their synonyms could be. However, the expansion had the opposite effect as expected; the performance decreased slightly, making the expansion of the Portuguese entities unnecessary (Table 4.8). MER did not find new entities and, in some cases, did not find entities that the two pipelines had found. Therefore, the final version of ICERL does not use this approach.

Table 4.8: Results of the expansion of the Portuguese entities.

Models	**Precision**	**Recall**	**F1-score**
cantemistpt	0.869 (-0.005)	0.842 (0.000)	0.855 (-0.003)
cantemistpt + biobertpt	0.867 (-0.006)	**0.843** (-0.001)	0.855 (-0.003)
biobertpt	**0.892** (+0.006)	0.823 (+0.002)	**0.856** (-0.002)

After the evaluation of the ICERL system, an error analysis was conducted in one hundred and eighty-three annotations, and the thirty-three errors were found. As expected, given the results of the models, the majority of the errors corresponds to entities that are not being recognized by the Portuguese (60.6%) and Spanish pipeline (30.3%). However, 9.1 % of the errors are related to the comparisons between dif-

ferent entities, which was not expected. For example, in the document ”S0034-70942013000200006”, the Spanish annotation is ”tumores epidermoides”, and the Portuguese is ”tumores”. None of the candidates of the Portuguese annotation is in the list of candidates of the Spanish annotation, which is why the system considers this example as FP. On the other hand, there are also some cases in which the system does not consider an error when comparing different entities. For example, in the document ”S0034-70942003000500011” the Portuguese annotation ”carcinoma na mama”, the Spanish annotation is ”carcinoma” and the system considers this case as a TP. In this case, the system found similarities between the two entities, and therefore they shared some of the same candidates. These annotation differences and the other two types of errors are assumed to be because different files were used to train the two pipelines. In addition, the Portuguese training files had to be corrected manually, which may have further differentiated the training annotations of the two languages.

The results obtained in Table 4.6 show that the ICERL system has a similar performance in both languages. Furthermore, the fact that there are no errors in which the same entities for Portuguese and Spanish do not have the same candidates reinforces this claim.

4.2.4 Applications

This subsection describes the work I did in the participation of LasigeBioTM at the ProfNER and MESINESP2 shared tasks [Ruas et al., 2021b,a]. Although these shared tasks are not specifically related to oncology, the participation in them served to apply the solutions found for the ICERL system.

My role in LasigeBioTM’s participation at the ProfNER shared task consisted in developing three NER models. These models were used to predict the entities in sub-track B (NER), and the resulting predictions were used in sub-track A (tweet binary classification). If the model recognized at least one entity in a tweet, the label ”1” was assigned to that tweet. If the model did not recognized a tweet, the label ”0” was assigned. The three models developed were:

- **“base”:** it includes Flair embeddings (es-forward and es-backward) trained on Spanish Wikipedia and Spanish FastText embeddings.
- **“twitter”:** it includes FastText Spanish COVID-19 Twitter Embeddings.
- **“medium”:** it includes FastText Spanish COVID-19 Twitter Embeddings, Flair embeddings (es-forward and es-backward) trained on Spanish Wikipedia and Spanish FastText embeddings.

The training parameters used in these models were the following: hidden size=256, minimum batch size=32, maximum epochs=55 and patience=3.

Since the ”base” model obtained better performance in the application on the validation set, this model was selected to be applied on the test set. On the test set, the ”base” model achieved an F1-score of 0.727 in sub-track B and 0.971 in sub-track A, which led to the second-best performance in sub-track A in the entire competition (Table 4.9).

Table 4.9: Results of the ”base” model and the median of all participants in sub-tracks A and B. P, R, and F1 refer to precision, recall, and F1-score, respectively.

	Sub-track A			**Sub-track B**		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**
base	**0.951**	**0.886**	**0.917**	**0.814**	**0.657**	**0.727**
median	0.919	0.855	0.886	0.842	0.727	0.761

For the MESINESP2 shared task, I developed a NEL module that uses the entities provided by the competition and links them to the DeCS. The entities are then given to the X-Transformer, a text classifier model, to perform the extreme multi-label classification, which is the competition's goal.

As the methodology used by the ICERL system, this NEL model uses string matching and the PPR-SSM algorithm to select the best candidate for each entity. After selecting the best candidate for all entities, the number of entities given to the extreme multi-label classification model is filtered by semantic similarity. That is, the Resnik's metric will select the entities that are similar to other entities recognized in the same document. Two models were created to assess the effect of selecting only the most relevant entities as the classifier model's input. The first selected all entities; the second only selected the top 25% according to their average semantic similarity with all entities.

The best results achieved by the classifier model corresponds to an F1-Score of 0.2007, 0.0686, and 0.0314 for sub-tracks MESINESP-L, MESINESP-T, and MESINESP-P, respectively. In the sub-tracks MESINESP-L and MESINESP-P, the results were achieved by using all the entities to the classifier model, while on sub-track MESINESP-T, only 25% of the entities were used. These results are low compared to the systems presented by the other participants, and the reason is related to the short training time of the classifier model.

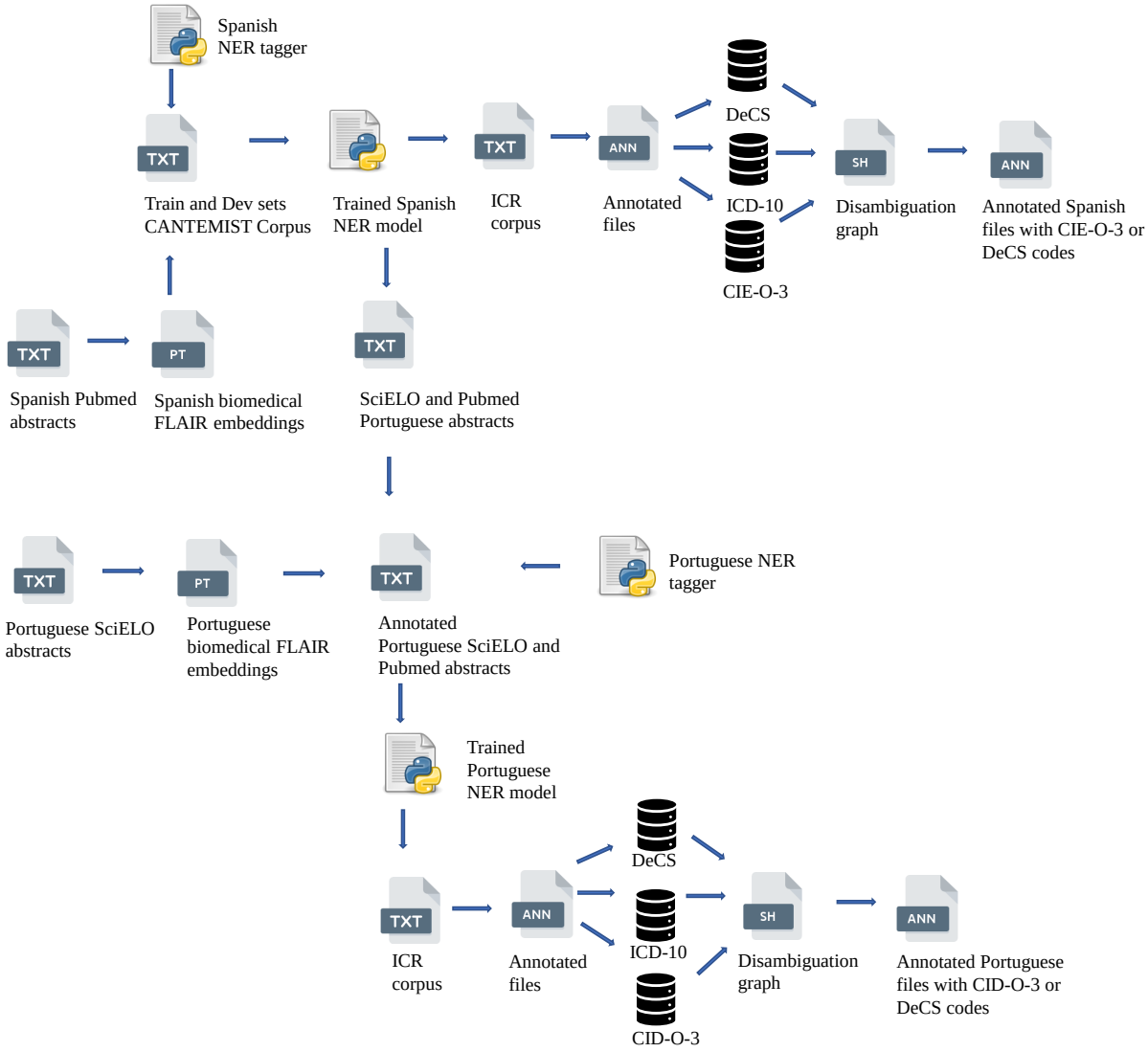

Figure 4.1: Description of the ICERL system application on the ICR corpus.

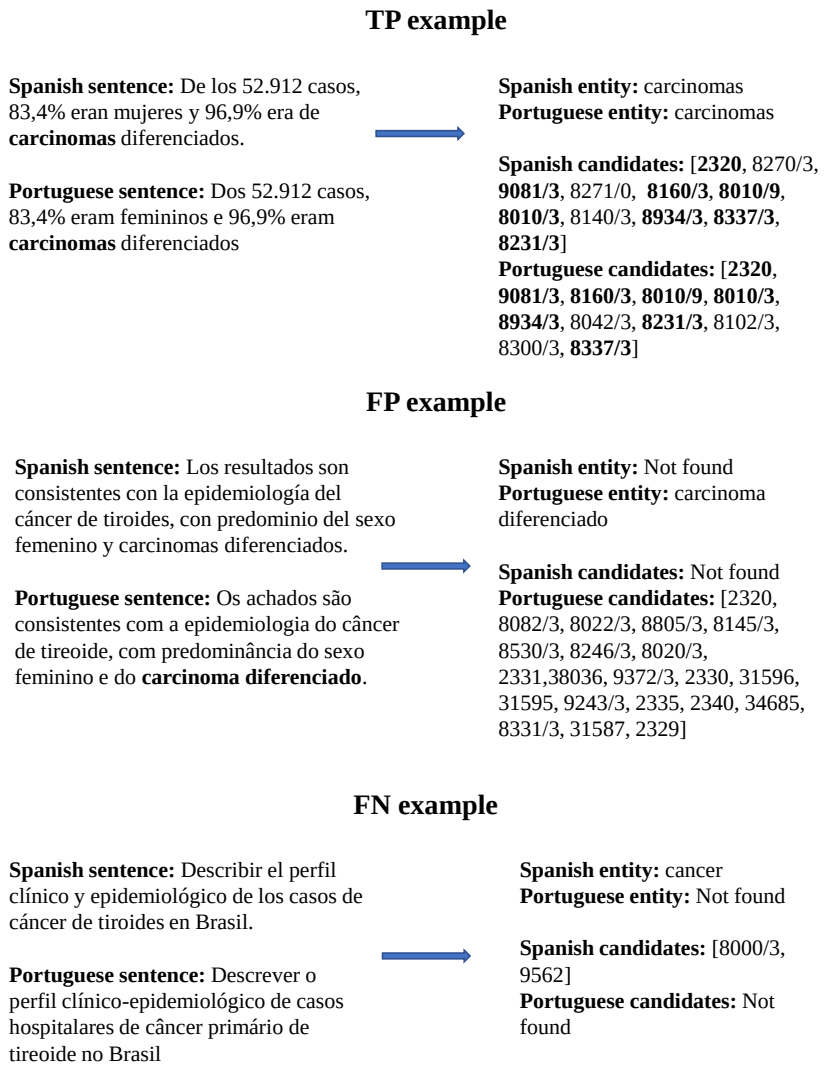

Figure 4.2: Evaluation examples of the ICERL system.

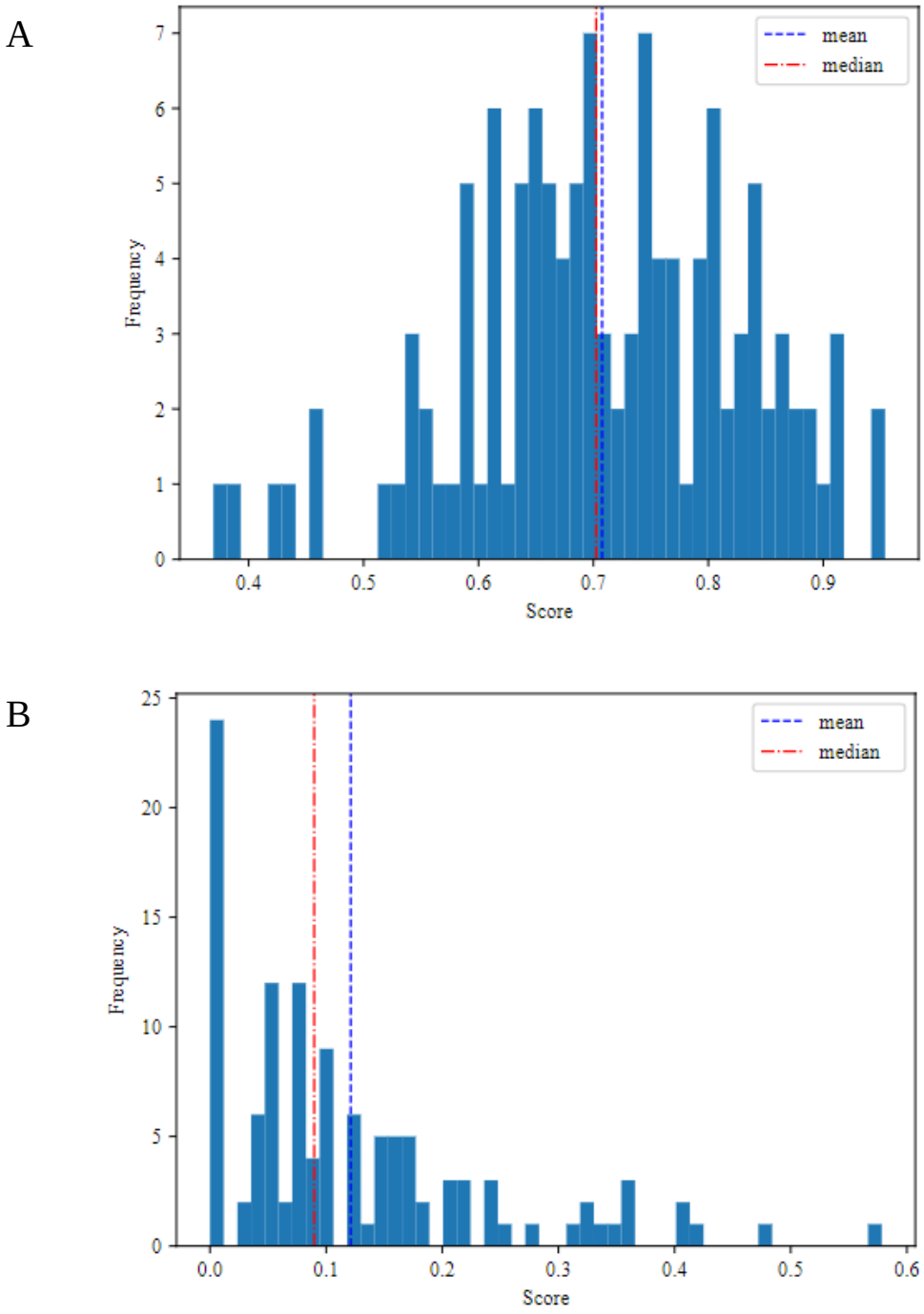

Figure 4.3: Score frequencies of the ICERL system (A) and the baseline (B) on the ICR corpus.

Chapter 5

Conclusions

The emergence of text mining tools for languages other than English is important in biomedical literature since it is often written in the author's native language. Thus, the ICERL system and the ICR corpus are two important contributions to text mining tools in the biomedical context, especially in the oncological domain for the Portuguese and Spanish languages, where similar tools do not exist. The ICERL system and the ICR corpus are available at https://github.com/lasigeBioTM/ICERL_system-ICR_Corpus.

The modifications carried out in the pipeline developed by LasigeBioTM constituted the first step in developing the ICERL system. The resulting pipeline from these modifications corresponds to the pipeline used by the ICERL system in texts written in Spanish. The modifications improved the F1-score from 0.741 to 0.754 for the NER task and from 0.061 to 0.665 for the NEL task.

The second step was to create a pipeline for the Portuguese language. For this, three solutions for the NER task were tested. The solution with the best results was a model composed of the BioBERTpt system and a pipeline equivalent to the Spanish pipeline. Although the lexical similarity between the two languages allows the application of the same pipeline for both languages, it was concluded that the specific pipeline for each language results in an F1-score 58 percentage points higher. In addition to the three models, an expansion of the Portuguese entities was tested to increase the performance of the pipeline. Still, the results obtained came to refute this approach.

The ICR corpus corresponds to 1,555 documents in Portuguese and Spanish taken from SciELO and PubMed. The application of the ICERL system in the ICR corpus resulted in 3,399 annotations for the Spanish language, of which 216 correspond to unique annotations and 3,287 in Portuguese, with 171 being unique annotations. The entity "cancer" was the most frequent annotation for both languages.

The similarity between the annotations statistics of the two languages and a 0.858 F1-score achieved by the evaluation method confirms the hypothesis proposed at the beginning of this dissertation; it is possible to use similar text mining tools for Portuguese and Spanish and to transfer annotations between the two languages maintaining comparable performance.

The solutions found for the ICERL system played an important role in the participation of Lasige-BioTM at the MESINESP2 and ProfNER shared tasks. At ProfNER, this contribution is highlighted by the second place achieved in the sub-track A. On the other hand, at MESINESP2, it is impossible to determine the impact of the NEL module created for the classifier model, as the low achieved results were due to the classifier model's training time. However, the participation in these shared tasks demonstrates that the work done for the ICERL system can be adapted to other domains.

5.1 Future work

Although the modifications made in the Spanish pipeline have greatly increased the performance in the NEL model, the same did not happen in the NER model. Therefore, the first suggestion for future work will be the improvement of this model. To do so, I recommend using a corpus from a shared task, such as the Codiesp, to generate new Flair embeddings. Another suggestion to achieve state-of-the-art results is using BERT models in their original framework. As already mentioned, the results achieved were lower than expected in the Flair framework. These approaches would increase the results in the NER task and the NEL task since, as shown in Chapter 3, the results of the NEL task are conditioned to the NER task.

One of the problems of this dissertation is related to the time spent training the NER models and the Flair embeddings. This was a limitation to test new approaches since some of these models took about four weeks to be ready. Therefore, it would be important to estimate the number of epochs that optimize the time spent training the models and the embeddings.

As mentioned in Chapter 4, the most frequent annotation corresponds to 75% of the total annotations in Spanish and 66% in Portuguese. The ICR could be composed by different annotations through the use of other queries related to the oncology domain, such as *Case Reports[Publication Type] AND POR[LA] AND oncology[MeSH]* and *Case Reports[Publication Type] AND POR[LA] AND Neoplasm[MeSH]*. Furthermore, as mentioned in Chapter 5, annotation differences between the two languages for the same entity were also observed. A solution to this problem could be using a parallel corpus for the training of the ICERL system and the same annotation criteria for both languages. This is because, despite manually correcting training annotations for the Portuguese pipeline, the amount of annotations does not allow all errors to be corrected. Moreover, the corrected training annotations for the Portuguese pipeline do not always follow the same annotation criteria used in the training files of the Spanish pipeline. Since the Spanish pipeline training files came from the CANTEMIST shared task and were designed only for entities present in the CIE-O vocabulary and not for DeCS. For example, the entity "neoplasia de la mama" is not present in the CIE-O, but it can be found in the DeCS.

The participation at the ProfNER and MESINESP2 shared tasks demonstrated that the hypothesis behind the development of the ICERL system could be adapted to other domains. The ICERL system and the ICR corpus could also be extended to other languages such as Catalan, Galician, Italian and French since there is a high lexical similarity between them and Portuguese and Spanish. In this case, the ICR corpus would have to be constituted by parallel corpora that included English in addition to these

languages. The evaluation method of the ICERL system would be the comparison of the performance on these languages with a performance of a state-of-the-art tool on the texts written in English. Since there has been a greater focus on text mining tools for the English language, the performance obtained on texts in English will be higher than for other languages. Therefore, the results obtained on the texts in English will be considered as the gold standard.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). 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)*, pages 54–59. 17

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649. 17

Akhtyamova, L., Martínez, P., Verspoor, K., and Cardiff, J. (2020). Testing contextualized word embeddings to improve ner in spanish clinical case narratives. 17, 24

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*. 11

Beal, M., Ghahramani, Z., and Rasmussen, C. (2001). The infinite hidden markov model. *Advances in neural information processing systems*, 14:577–584. 8

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. *Journal of web semantics*, 7(3):154–165. 12

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270. 14

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. 10

Borst, W. N. (1999). Construction of engineering ontologies for knowledge sharing and reuse. 13

Campos, L., Pedro, V., and Couto, F. (2017). Impact of translation on named-entity recognition in radiology texts. *Database*, 2017. 2, 16

Campos, L. F. L. (2017). Semantic annotation of electronic health records in a multilingual environment. 18

Castro, S., Bonanata, J., and Rosá, A. (2018). A high coverage method for automatic false friends detection for spanish and portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 29–36. 3

Consortium, G. O. (2008). The gene ontology project in 2008. *Nucleic acids research*, 36(suppl_1):D440–D444. 14

Couto, F. M. and Lamurias, A. (2018). Mer: a shell script and annotation server for minimal named entity recognition and linking. *Journal of Cheminformatics*, 10(1):58. 7, 17

Crichton, G., Pyysalo, S., Chiu, B., and Korhonen, A. (2017). A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):368. 9

Cruz Díaz, N. P. and Maña López, M. (2015). An analysis of biomedical tokenization: Problems and strategies. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 40–49, Lisbon, Portugal. Association for Computational Linguistics. 9

Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M. (2007). Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl_1):D344–D350. 14

Deng, P., Chen, H., Huang, M., Ruan, X., and Xu, L. (2019). An ensemble cnn method for biomedical entity normalization. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 143–149. 8

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 10, 11, 12

Dodd, S., Clarke, M., Becker, L., Mavergames, C., Fish, R., and Williamson, P. R. (2018). A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. *Journal of clinical epidemiology*, 96:84–92. 15

Edunov, S., Baevski, A., and Auli, M. (2019). Pre-trained language model representations for language generation. *arXiv preprint arXiv:1903.09722*. 9

Frakes, W. B. and Baeza-Yates, R., editors (1992). *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Inc., USA. 13

García-Pablos, A., Perez, N., and Cuadros, M. (2020). Vicomtech at cantemist 2020. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*. 17, 25

Gudivada, V. N., Rao, D. L., and Gudivada, A. R. (2018). Information retrieval: Concepts, models, and systems. In *Handbook of statistics*, volume 38, pages 331–401. Elsevier. 13

Harpring, P. (2010). *Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works*. Getty Publications. 13

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

Hu, B., Lu, Z., Li, H., and Chen, Q. (2014). Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pages 2042–2050. 8, 9

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*. 8, 9

Ji, B., Liu, R., Li, S., Yu, J., Wu, Q., Tan, Y., and Wu, J. (2019). A hybrid approach for named entity recognition in chinese electronic medical record. *BMC medical informatics and decision making*, 19(2):149–158. 10

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc. 1, 7, 18

Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C. J., Binder, J. X., Malone, J., Vasant, D., et al. (2015). Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research*, 43(D1):D1071–D1078. 14

Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S. M., Boerkoel, C. F., Boycott, K. M., et al. (2017). The human phenotype ontology in 2017. *Nucleic acids research*, 45(D1):D865–D876. 15

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 8

Lamurias, A. and Couto, F. (2019). *Text Mining for Bioinformatics Using Biomedical Literature*, page 602–611. 1, 7

Lamurias, A., Ruas, P., and Couto, F. M. (2019). Ppr-ssm: personalized pagerank and semantic similarity measures for entity linking. *BMC bioinformatics*, 20(1):1–12. 12, 17

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444. 8

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. 11

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195. 12

Li, H., Chen, Q., Tang, B., Wang, X., Xu, H., Wang, B., and Huang, D. (2017). Cnn-based ranking for biomedical entity normalization. *BMC bioinformatics*, 18(11):79–86. 8

Li, Y., Bontcheva, K., and Cunningham, H. (2009). Adapting svm for natural language learning: A case study involving information extraction. *Natural Language Engineering*, 15(2):241–271. 7

López-Úbedaa, P., Dıaz-Galianoa, M., Martın-Valdiviaa, M., and Ureña-Lópeza, L. A. (2020). Extracting neoplasms morphology mentions in spanish clinical cases through word embeddings. *Proceedings of IberLEF*. 24

Manning, C. D., Schütze, H., and Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge university press. 7

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. 10

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119. 9

Miranda-Escalada, A., Farré, E., and Krallinger, M. (2020). Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. 21, 25

Moreno, I., Boldrini, E., Moreda, P., and Romá-Ferri, M. T. (2017). Drugsemantics: a corpus for named entity recognition in spanish summaries of product characteristics. *Journal of biomedical informatics*, 72:8–22. 16

Nelson, S. J., Johnston, W. D., and Humphreys, B. L. (2001). Relationships in medical subject headings (mesh). In *Relationships in the Organization of Knowledge*, pages 171–184. Springer. 14

Névéol, A., Dalianis, H., Velupillai, S., Savova, G., and Zweigenbaum, P. (2018). Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12. 1

Neves, M., Yepes, A. J., and Névéol, A. (2016). The scielo corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, Portorož, Slovenia. European Language Resources Association (ELRA). 1

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab. 12

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543. 10

Peters, A. C., da Silva, A. M. P., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Hasan, S. A., Moro, C. M. C., et al. (2020). Semclinbr–a multi institutional and multi specialty semantically annotated corpus for portuguese clinical nlp tasks. *arXiv preprint arXiv:2001.10071*. 16

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*. 10, 11

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. (2020). Pre-trained models for natural language processing: A survey. *arXiv preprint arXiv:2003.08271*. 9

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*. 17

Reys, A. D., Silva, D., Severo, D., Pedro, S., e Sá, M. M. d. S., and Salgado, G. A. (2020). Predicting multiple icd-10 codes from brazilian-portuguese clinical notes. In *Brazilian Conference on Intelligent Systems*, pages 566–580. Springer. 17

Rosales-Méndez, H., Hogan, A., and Poblete, B. (2018). Machine translation vs. multilingual approaches for entity linking. 2

Ruas, P., Andrade, V. D., and Couto, F. M. (2021a). Lasige-biotm at mesinesp2: entity linking with semantic similarity and extreme multi-label classification on spanish biomedical documents. *Proceedings of the Working Notes of Conference and Labs of the Evaluation Forum (CLEF)*, pages 324–334. 33

Ruas, P., Andrade, V. D., and Couto, F. M. (2021b). Lasige-biotm at profner: Bilstm-crf and contextual spanish embeddings for named entity recognition and tweet binary classification. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 108–111. 33

Ruas, P., Neves, A., Andrade, V., and Couto, F. (2020). Lasigebiotm at cantemist: Named entity recognition and normalization of tumour morphology entities and clinical coding of spanish health-related documents. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*. 3, 21, 23, 24

Schneider, E. T. R., de Souza, J. V. A., Knafou, J., e Oliveira, L. E. S., Copara, J., Gumiel, Y. B., de Oliveira, L. F. A., Paraiso, E. C., Teodoro, D., and Barra, C. M. C. M. (2020). Biobertpt-a portuguese

neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72. 17

Scutari, M. and Strimmer, K. (2010). Introduction to graphical modelling. *arXiv preprint arXiv:1005.1036*. 12

Shen, W., Wang, J., and Han, J. (2014). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460. 12

Simon, C., Davidsen, K., Hansen, C., Seymour, E., Barnkob, M. B., and Olsen, L. R. (2019). Bioreader: a text mining tool for performing classification of biomedical literature. *Bmc Bioinformatics*, 19(13):57. 1

Soares, F. and Krallinger, M. (2019). Bvs corpus: A multilingual parallel corpus of biomedical scientific texts and translation experiments. 16

Sousa, D. F. d. (2019). *Extracting phenotype-gene relations from biomedical literature using distant supervision and deep learning*. PhD thesis. 1

Sun, C. and Yang, Z. (2019). Transfer learning in biomedical named entity recognition: An evaluation of bert in the pharmaconer task. pages 100–104. 16

Tan, A.-H. et al. (1999). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, volume 8, pages 65–70. sn. 1

Ulsh, J. L. (1971). *From spanish to portuguese*. Foreign Service Institute, Department of State. 3

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008. 10

Wallach, H. M. (2004). Conditional random fields: An introduction. *Technical Reports (CIS)*, page 22. 8

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013). Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(W1):W518–W522. 1

Xionga, Y., Huanga, Y., Chena, Q., Wanga, X., Nic, Y., and Tanga, B. (2020). A joint model for medical named entity recognition and normalization. *Proceedings http://ceur-ws. org ISSN*, 1613:0073. 17, 25

Zeng, M. L. (2008). Knowledge organization systems (kos). *KO KNOWLEDGE ORGANIZATION*, 35(2-3):160–182. 13

Zheng, Z., Li, F., Huang, M., and Zhu, X. (2010). Learning to link entities with knowledge base. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 483–491. 12

Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W., and Shen, B. (2013). Biomedical text mining and its applications in cancer research. *Journal of biomedical informatics*, 46(2):200–211. 1