uc3m | Universidad Carlos III de Madrid

Master in Big Data Analytics

2019-2020

Master thesis

"Augmenting knowledge about rare diseases by Deep Learning"

Ainara Apezteguia Garcia
Tutor
Isabel Segura Bedmar
Madrid, July 2020

ACKNOWLEDGEMENTS

This work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain, (DeepEMR project TIN2017-87548-C2-1-R) and the Interdisciplinary Projects Program for Young Researchers at Universidad Carlos III of Madrid founded by the Community of Madrid (NLP4Rare-CM-UC3M).

I would like to thank the research staff at Universidad Carlos III de Madrid for the opportunity given to develop this Project and let to widen my academic knowledge. Special thanks to Isabel Seura Bedmar for her professional guidance and valuable critiques of this work

ABSTRACT

Due to the low incidence of rare diseases, representation and research in this field are limited resulting in a poor understanding of the natural history of the epidemiology and a lack of knowledge resources for both patients and professionals. It was not until 1966 that an inventory of rare disorders was created (by the Online Mendelian Inheritance in Man (OMIM)). Still, the efforts made so far to improve the attention to these patients are not enough. For this reason, the European Commission is engaged in initiatives that contribute to the expansion of knowledge in this area**.**

In this study, different Natural Language Processing (NLP) and Deep Learning (DL) techniques are explored to extract knowledge about rare diseases from the disperse literature via the task of named entity recognition (NER). In particular, Bidirectional Long Short-Term Memory (LSTM) models with word and character embedding, as well as a Bidirectional Encoder Representations from Transformers (BERT) are compared to baseline approaches, such as the classifier Conditional Random Field (CRF) and the NLP tool, *spaCy*. To do so, a silver corpus has been created automatically due to the lack of availability of corpora annotated with rare diseases. Experiments on this corpus concluded that CRF achieves the highest macro F1-score (0.69), probably because deep learning models are not able to overcome classical machine learning algorithms when the training data is limited. It is also concluded that Bi-LSTM with character embedding provides a better F1-score than the model with word embedding. Although BERT has been successfully applied to many NLP tasks, in this study, it does not overcome the other studied techniques. The code developed to carry out the experiments is freely available at *https://github.com/ainara6/NLP4RARE.*

Keywords: Named Entity Recognition, Rare diseases, Machine Learning, Deep Learning

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Rare diseases or orphan diseases affect a small number of people in comparison with the general population; in Europe, this ratio is 1 person per 2000 (Schieppati, Henter, Daina, & Aperia, 2008). Nearly all genetic diseases are rare diseases while not all rare diseases are genetic diseases (Vincent, 2012). They are characterized by a large number and wide diversity of symptoms, which vary not only from disease to disease but also within the same pathology and, in many cases, signs may be observed at birth or in childhood. In a high percentage, they are chronic and degenerative (65% of these pathologies are severe and invalidating). The mortality of these patients is very high, the five-year survival rate from the diagnosis is 50%. Due to the low incidence, research is limited and the pharmaceutic industry usually does not invest in this kind of disease. This leads to an insufficient knowledge for the diagnosis and treatment of these diseases, which sometimes results in the lack of treatment of more than 94% of rare diseases, or if they do have a treatment it is not suitable (Austin et al., 2018).

One of the principal objectives of the European Union is to encourage research to improve the attention to these patients and increase the knowledge about rare diseases. Natural Language Processing (NLP) refers to how computers understand language (e.g., speech, text) in terms of language translation, semantic understanding, and summarization (Graham et al., 2020). It builds computational algorithms to automatically analyze and represent human language. There are several underlying tasks and machine learning models powering NLP applications. Specifically, deep learning (DL) approaches have obtained very high performance across many different NLP tasks recently (Julia Hirschberg, 2015). The goal of this study is to explore different NLP and DL techniques to extract the knowledge about rare diseases, that are described and, many times disperse in the medical literature and databases. The processing and analysis of the texts enable the transformation of the information in a non-structured format to a structured format, providing a formal representation of the knowledge in the domain of rare diseases. This formal representation could lead to benefits in research and clinical practice. First of all, the existence of a structured resource will speed up notably the access to the relevant information for a specific disease and could provide the identification of the complex relations present in rare diseases. Secondly, the structured information can be processed by automatic algorithms capable of inferring new relations and patterns between the diseases. As a final result, this could contribute significantly to the acceleration of the diagnosis of these diseases, as well as provide more information about possible treatments. The automatic extraction of relevant information about rare diseases alleviates the workload of experts, saving time and money.

In particular, this project aims to develop a Named-entity recognition (NER) system for detecting rare diseases and their symptoms from texts. Most of the research in NER for the biomedical domain has been focused on the recognition of gene and protein names (Pérez-Pérez et al., 2017). More recent efforts have been aimed at identifying chemical and disease names (Krallinger, Leitner, & Rabal, 2013); (Doğan, Leaman, & Lu, 2014). However, research dedicated to extracting the information in the rare disease sector has been much smaller.
To train these machine learning models for any NLP task, a collection of annotated texts (also named as corpora) is required. The trained model is highly dependent on the amount and quality

of the annotations of these corpora. There are two types of annotated corpora depending on the source of the annotations: i) Gold Standard Corpora (GSC) where annotations are performed manually by expert annotators, following fixed guidelines, and ii) Silver Standard Corpora (SSC) where annotations are automatically generated by computerized systems (Sakurai, 2012). An SSC with annotations of rare diseases and symptoms has been created due to the almost complete absence of corpora for rare diseases (Fabregat, Araujo, & Martinez-Romo, 2018).

NER system has tools that classify different types of named entities applying different methodologies in different domains and languages. Among the most used tools for NER, there are *spaCy*, *NLTK*, and *Stanford CoreNLP*. The *spaCy* library (Honnibal, M., & Montani, 2017) (a free, open-source library for advanced NLP in *Python* (Van Rossum, G., & Drake Jr, 1995)) is used in this study for building a baseline model along with the Conditional Random Field (CRF) model. DL approaches Bidirectional Long Short-Term Memory (Bi-LSTM) and Bidirectional Encoder Representations from Transformers (BERT) are also used as they have proven state of the art performance. LSTM has proven to be very good in comparison with other methods such as convolutional neural networks (CNN) (Xu, Shi, Zhao, & Zheng, 2018). It is proved (The Anh Le, Mikhail Y. Arkhipov, 2017) that the extension of Bi-LSTM model with CRF significantly increased the quality of predictions for other fields such as Russian NER task but it also achieves satisfactory performance for NER in biomedical texts (Rivera Zavala, Martínez, & Segura-Bedmar, 2018). It is shown (Labusch, Neudecker, & Zellh, 2019) that an appropriately pre-trained BERT model delivers decent recognition performance in a variety of settings and even provides state of the art performance (Bi-LSTM-CRF) in many cases without extensive fine-tuning and optimization requirements. It is demonstrated (Akhtyamova, 2020) how the NER model detects pharmacological substances, compounds, and proteins in the dataset obtained from the Spanish Clinical Case Corpus (SPACCC) with a BERT language representation model trained from scratch. Although these DL models have been applied for NER for clinical entities (Qin & Zeng, 2018) and disease name recognition (Sahu & Anand, 2016), there is hardly any research on their application for NER for rare diseases, (Fabregat et al., 2018) makes use of an LSTM network to relate disabilities and rare diseases.
This study gathers relevant projects related to NER and rare diseases to know the state of the art of the challenge, learn different tools for text preprocessing, study the principal techniques of DL that are currently used for NER and create a corpus. Moreover, it implements DL techniques and adapts them to the problem to train a model for the recognition of rare diseases and symptoms, and it performs the evaluations of such models as well as the error analysis, that enables the discussion and comparison of the different models.

Therefore, this document is structured as follows: this introduction constitutes the first chapter, followed by the state of the art. Next, the creation of the corpora is described in the third chapter. Besides, the fourth incorporates the methods of Machine Learning (ML) applied to the recognition of rare diseases. Moreover, the fifth section states the evaluation and discussion of the results and finally, chapter 6 includes conclusions and future work.

2. STATE OF THE ART

This chapter describes the main approaches and techniques that have been used to develop NER systems with a special focus in the biomedical domain. It also makes a review of the main annotated corpora with diseases.

In the Big Data Era, the NLP has become one of the most relevant research areas among those that enable the processing and analysis of huge volumes of non-structured information available in any field of knowledge. The application of NLP techniques in the processing of the information contained in the texts from the biomedical domain enables the extraction of the relevant information (expressed in concepts, relations, and events) and its transformation to a structured format, which facilitates the access and analysis of its information.

A clear example is the health domain, as there is a huge amount of information in electronic formats such as databases, ontologies, scientific literature, electronic historical clinics of patients, clinical trials, information about medicines, briefing notes and about security, health agencies bulletins, or even social networks and specific forums about health, etc.

NER is an NLP task aimed at identifying references to specific entities in a natural language text and labeling them with their location and type. Those entities can be of different types such as genes, proteins, organizations, places, dates, quantities, monetary values, or percentages.

NER is typically modeled as a label sequence problem, which may be defined formally as follows according to Settles (Settles, 2004): given a sequence of input tokens $x = (x_1 \ldots x_n)$, and a set of labels L, determine a sequence of labels $y = (y_1, \ldots, y_n)$ such that $y_i \in L$ for $1 \leq i \leq n$. Token position can be modeled in different formats, but in this study, the IOB format has been used. It specifies whether each token is at the beginning of an entity (B), inside an entity (I), or outside (O), is capable of distinguishing between consecutive entities (See Figure 1). IOB was defined in the study of Carreras et al. (Carreras & Màrquez, 2005) and became popular due to its practicality and good performance.

Figure 1: example of IOB tagging.

The main approaches for NER can be categorized as being based on rules, dictionary matching, or ML. Each approach meets different demands, depending on the linguistic characteristics of the entities to be identified.

Rule-based approaches make use of regular expressions which combine information from terminological resources and characteristics of the entities of interest. The principal drawback of these techniques is that the rules have to be built manually, which requires a lot of time and is domain-dependent (Farmakiotou et al., 2000); (Petasis et al., 2001).

Dictionary-based techniques consist of matching the text phrases with concept synonyms that exist in the terminological resources (dictionaries) (Eftimov, Seljak, & Korošec, 2017). These methods are useful with closely defined vocabulary names such as diseases and species. However, it only recognizes the entity mentions that exist in the resources, but the terminological resources are regularly updated with new concepts and synonyms (Miller, Gieszczykiewicz, Vries, & Cooper, 1992); (X. Zhou, Zhang, & Hu, 2006). These methods achieve high precision, but the recall can be low if dictionaries are not complete (Hai Hu, Richard J. Mural, 2008).

ML-based approaches work well with the presence of strong variability and highly dynamic vocabulary of names (e.g. genes and proteins). ML-based solutions usually provide the best performance results. The ML models can be classified as supervised or semi-supervised models, depending on unannotated data being used or not. Supervised learning only uses annotated data and has received the most research interest. As a consequence, different supervised models have been used on NER systems, namely Conditional Random Fields (CRF) (Lafferty & Mccallum, 2001), Support Vector Machines (SVMs) (Steinwart, I., & Christmann, 2008), Hidden Markov Models (HMMs) (Baum & Petrie, 1966) and Maximum Entropy Markov Models (MEMMs) (McCallum, Freitag, & Pereira, 2000) and more recently, deep learning models such as Bi-LSTM (P. Zhou et al., 2016).

We now present the main NER systems for the biomedical domain, with a special focus on those based on deep learning. NER in the biomedical field is overall regarded as more difficult than in other domains, some of the reasons are stated (Leaman & Gonzalez, 2008). First, millions of entity names are in use and there is a constant addition of new ones, meaning that neither dictionaries nor training data will be sufficiently comprehensive. Second, very similar or identical names and acronyms are used for different concepts and the fast evolution of the biomedical field hampers the consensus on the name to be used for a given entity or the exact concept it defines. As a consequence, significant ambiguities can be found in this domain. Although there are naming conventions, these are not followed by authors frequently as they prefer to introduce their abbreviations and use it throughout the scientific article. Finally, as entity names in texts of this field are longer on average than names from other domains, it is more difficult to detect the boundaries of the entity name than determining if the entity is present.

In the last decade, various competitions such as BioCreative (Hirschman, Yeh, Blaschke, & Valencia, 2005), i2b2 (Ö. Uzuner, South, Shen, & DuVall, 2011) BioNLP shared tasks (Kim et al., 2012) and DDIExtraction (Segura-Bedmar, Martínez, & Herrero-Zazo, 2013) have largely contributed to advancing research on NLP techniques applied to the fields of biology and biomedicine. As a result, many systems and tools have been developed for entity recognition and the extraction of relations in this domain. However, research

devoted to obtaining information about the rare disease domain is still scarce (Chen & Altman, 2017); (Métivier, Serrano, Charnois, Cuissart, & Widlöcher, 2015); (Laburu, Perez, Casillas, Goenaga, & Oronoz, 2019).

Hyejin Cho and Hyunju Lee (Cho & Lee, 2019) propose a NER system for biomedical entities, named as Contextual LSTM (CLSTM), by using n-grams (sequences of words or characters) in a Bi-LSTM neural network with a CRF classifier in the last layer. The CRF layer tags the input tokens sequence according to the IOB format described above.

In addition to the CLSTM model, the authors also evaluate several deep learning models: i) Bi-LSTM ii) Bi-LSTM-CRF without using n-grams, iii) GRAM-CNN, a CNN exploiting n-grams and word embeddings (word vectors) as input, and iv) BERT which is described in the fourth section. The authors use three corpora: the disease corpus of the National Center for Biotechnology Information (NCBI) (Doğan et al., 2014), the BioCreative II Gene Mention corpus (GM) (Smith, Larry; Tanabe, 2008), and the BioCreative V Chemical Disease Relation corpus (CDR) (Li et al., 2016). The BERT model provides the best F1 score in the GM dataset (81.65%) while in the NCBI and CDR datasets the CLSTM model with word and character levels provides the best F1 score (85.68% and 86.44% respectively).

The study conducted by Liu et al. (Liu et al., 2017) investigates the performance of a recurrent neural network (RNN) on clinical entity recognition and protected health information (PHI) recognition. They propose an extension of the IOB format, "BIOES" (B-beginning of an entity, I-insider an entity, O-outsider an entity, E-end of an entity, S-a single-token entity) to represent entities. Moreover, they prove that LSTM outperforms CRF, by introducing two types of character-level word representations (sentence converted to a vector in the character level, which captures morphological and shape information instead of syntactic and semantic information that provides the word level embedding) (Aaltonen et al., 2015) into the input layer of LSTM. Experiments conducted on corpora of 2010, 2012, and 2014 i2b2 NLP challenges (O. Uzuner, 2004) showed that LSTM achieves highest micro-average F1-scores of 85.81% on the 2010 i2b2 medical concept extraction, 92.29% on the 2012 i2b2 clinical event detection, and 94.37% on the 2014 i2b2 de-identification.

Jauregi Uananue et al. (Jauregi Unanue, Zare Borzeshi, & Piccardi, 2017) have compared bidirectional LSTM and the Bi-LSTM-CRF for drug name recognition (DNR) and clinical concept extraction (CCE) with a baseline CRF model. As input features, they have applied combinations of different word embeddings, character-level embeddings, and conventional feature engineering (such as the word itself, the part-of-speech (POS) tag, orthographical features, and affixes). The evaluation was performed on two biomedical corpora: 2010 i2b2/VA (Ö. Uzuner et al., 2011), which consists of a collection of clinical texts (annotated with treatment relations, test relations, and medical problem relations) and the DDI corpus (Herrero-Zazo, Segura-Bedmar, Martínez, & Declerck, 2013), a collection of MedLine abstracts and texts from the DrugBank database, which was annotated with drugs and their drug-drug interactions. The system obtains an F1 of 83.35% on the i2b2/VA dataset, 88.38% on the DrugBank text of the DDI corpus, while

its performance is lower (60.66% F1) on the MedLine abstracts of the DDI corpus. Their experiments show that the neural network models obtain significantly better results than the CRF. Moreover, the use of hand-crafted features does not further improve performance.

Table 1 summarizes the most recent NER systems based on deep learning models in the biomedical domain.

TABLE 1:SUMMARY OF THE PERFORMANCE OF THE STATE-OF-THE-ART SYSTEMS.

Approach	Corpora	Entity types	Tagging approach	F1-score
Bi-LSTM	NCBI	diseases	IOB	86.26
Bi-LSTM+CRF			IOB	86.18
BERT			IOB	85.70
Bi-LSTM	GM	genes	IOB	82.83
Bi-LSTM+CRF			IOB	86.36
BERT			IOB	87.85
Bi-LSTM	CDR	diseases, chemical entities	IOB	83.87
Bi-LSTM+CRF			IOB	85.69
BERT			IOB	86.81
Bi- LSTM+CRF	DDI (Drugbank)	drugs	IOB	88.38%
Bi- LSTM+CRF	DDI (Medline)		IOB	60.66%

As is shown in previous paragraphs, over the past few years, a considerable number of studies have been dedicated to investigating deep learning models for the NER task in the biomedical domain. However, very few efforts have been done to recognize rare diseases from biomedical texts. Chen and Atlman (Chen & Altman, 2017) propose a system that combines NLP techniques and manual annotation for identifying therapeutic targets of rare diseases. The system exploits a list of orphan drugs proposed by the FDA ("USA Food and Drug Administration," n.d.) and the database ("OMIM - Online Mendelian Inheritance in Man.," n.d.) with clinical and genetical information about rare diseases. This way, they can link drugs, diseases, and genes to obtain new medicines for diseases that share the same genetic causes.

Métivier et al. (Métivier et al., 2015) describe a straightforward approach based on the use of lexical patterns to extract information about symptoms and rare diseases. However, such work lacks a formal evaluation and does not seem to have been continued in posterior works.

There are some relevant annotated corpora available for diseases entities such as NCBI Disease Corpus (Doğan et al., 2014) SCAI Disease (Gurulingappa, Klinger, Hofmann-apitius, & J, 2010), EBI Disease (A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, 2008), Arizona Disease (AZDC) (Leaman R, Miller C, 2009), BioText ("Data for research on relations between DISEASE/TREATMENT entities. Berkeley.," n.d.). However, there is a lack of this kind of annotated corpora for rare diseases. Fabregas et al. (Fabregat et al., 2018) created a new corpus of abstracts from scientific papers related to rare diseases, which has been manually annotated with disabilities. It allows training ML and DL systems that can automatically annotate new texts. The final goal is to extract information about the relations between rare diseases and their disabilities. The corpus was named "Elsevier Bilingual Corpus for Rare Diseases" (Fabregat Marcos, Hermenegildo; Araujo Serna, Lourdes; Martínez Romo, 2018) composed of a set of summaries of articles written in English and Spanish about rare diseases. This corpus was created for the DIANN task (disability annotation on documents from the biomedical domain) (Fabregat, H., Martinez-Romo, J., & Araujo, 2018). Disabilities affect a large part of the population, even if there are tools for the annotations of medical concepts (e.g. Metamap for texts in English), they consider disabilities as any other sign. It is important to collect information related to them as they are present in several rare diseases. The DIANN task aims to identify disabilities in English and Spanish texts, as well as negated disabilities. The top system was developed by Rivera Zavala et al. (Rivera Zavala, Martinez, & Segura-Bedmar, 2018), which consists of a hybrid Bi-LSTM and CRF model with sense-disambiguation embedding (enables to learn a distributed representation for each sense of a word in contrast to word embedding methods, as they learn a single vector per word) and an extended tag encoding format to detect discontinuous entities, as well as overlapping or nested entities. Unlike IOB tagging, BMEWO-V encoding format assigns B tag for the start of the entity, M tag for the entity continuity, the E tag for the end of the entity, the W tag for a single entity and the O tag for tokens not belonging to any entity. This way, it allows representing nested entities. This extended tag encoding format improves the result of the predictions while the pre-trained models of words and embedded senses reduce training time and increase the accuracy of labeling. The model results for F-Score are 68.2% for the English task and 65.3% for the Spanish task.

This project aims to explore different approaches for recognizing rare diseases from texts. As baselines, it uses a dictionary-based technique with the *spaCy* library (Honnibal, M., & Montani, 2017) and a CRF classifier. It also applies two different deep learning techniques: i) Bi-LSTM-CRF (Bidirectional Long-Short Memory with Conditional Random Fields) and ii) BERT (Bidirectional Encoder Representations from Transformers) to extract mentions about rare diseases and their symptoms from biomedical texts. To train and test the models, the first silver corpus annotated with rare diseases and symptoms has been created. In the next chapter, the creation of this corpus is described. Chapter 4 presents the NER approaches used in this project in detail.

3. CORPUS CONSTRUCTION FOR RARE DISEASES

Annotated corpora are indispensable to train and test machine learning systems. Therefore, one of the most important objectives of this project is to create a corpus annotated with rare diseases and symptoms to implement and evaluate several machine learning models for recognizing rare diseases and their symptoms.

The corpus consists of 1,041 texts written in English, which were gathered from the Rare Disease database, one of the most popular websites with information about rare diseases created by the National Organization for Rare Diseases (NORD) ("National Organization for Rare Diseases (NORD)," 1987). For a given rare disease, the database provides information about its sign and symptoms, causes, affected populations, among many others. In particular, the following sections have been used: general discussion, symptoms, causes, affected populations, related disorders, diagnosis, and standard therapies. The texts were crawled from this web by using the *Python* (Van Rossum, G., & Drake Jr, 1995) crawler library *BeautifulSoup* (Richardson, 2007).

These texts have been automatically annotated using an extension of *spaCy* (Honnibal, M., & Montani, 2017) which allows identifying named entities using dictionaries. In particular, these three dictionaries have been used:

- Disease Ontology (DOID) (Schriml et al., 2019), a database containing a total of 9,871 disease terms with 69% of DO terms defined, which is created and managed by the European Bioinformatics Institute (EMBL-EBI).
- Orphan Rare Disease Ontology (ORDO)("Orphanet Rare Disease Ontology (ORDO)," n.d.), the most comprehensive ontology with information about rare diseases, it contains 14,501 classes.
- Symptom Ontology (created and managed by the EMBL-EBI) with more than 1,164 terms describing symptoms.

After removing overlapped terms from the DOID and ORDO ontologies, the disease dictionary consists of 28,275 entries and the rare disease dictionary of 20,946 ones.

The texts were annotated using *spaCy* with diseases, rare diseases, and symptoms. These annotations were represented using the brat standoff format (Saklofske, 2019), which has recently become a de facto standard to annotate corpora for many NLP tasks, especially for NER systems. In the brat format, annotations are stored separately from its corresponding source text. Figure 2 shows an example of a text annotated with this brat standoff format. Each line contains one annotation, and each annotation is given an ID that appears first on the line, separated from the rest of the annotation by a single TAB character, then the word or group of words (entities), the offsets (start and end positions of the entity mention in the source text) and finally the entity type (disease, rare diseases or symptom).

Ablepharon.txt

Ablepharon-Macrostomia Syndrome (AMS) is an extremely rare inherited disorder characterized by various physical abnormalities affecting the head and facial (craniofacial) area, the skin, the fingers, and the genitals. In addition, affected individuals may have malformations of the nipples and the abdominal wall. Infants and children with AMS may also experience delays in language development and, in some cases, mental retardation.

Ablepharon.ann

T1	RAREDISEASE	11	22	Macrostomia
T2	DISEASE	23	31	Syndrome
T3	DISEASE	415	433	mental retardation

Figure 2:example of brat annotation format.

The annotated corpus is spitted into training, development, and test dataset with a ratio of 70-10-20. Thus, the training dataset contains 1,458 texts, the development dataset a total of 208 texts, and 416 in the test set.

The created corpus contained the two existent types of overlaps between entities (Muis & Lu, 2017):

- Crossing entities: two entities overlap but neither is contained in another (see Figure 3).
- Nested entities: entity mentions are embedded in longer entity mentions (see Figure 4).

There are also cases in which a word or a set of words can be assigned different entities, the corpus also contains this case (see Figure 5).

Menkes-Disease.txt

DISEASE

Menkes disease is a genetic disorder of copper metabolism

RARE DISEASE

that is detectable before birth (prenatally) and which follows a progressively degenerative path involving several organs of the body but especially the brain.

Figure 3:example of overlapping entities.

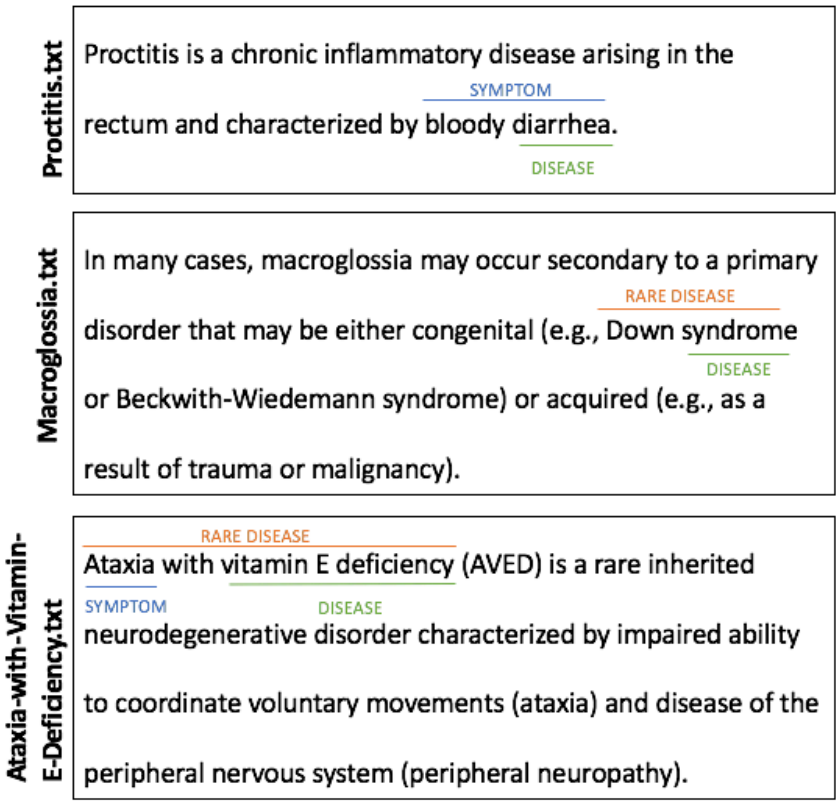

Figure 4: examples of nested entities.

The symptom-disease nested example shown in Figure 4 happens between **bloody diarrhea** and **diarrhea**, in the context of the text, **bloody diarrhea** is a symptom: *"Proctitis is a chronic inflammatory disease arising in the rectum and characterized by **bloody diarrhea**"*. But the word **diarrhea** is contained in the disease dictionary, so there is a relationship of hypernym. **Bloody diarrhea** is the hyponym of **diarrhea.**

Regarding the example of the nested overlap between rare disease and disease, in the context of the text, **Down syndrome** is a rare disease, but the word **syndrome** is contained in the disease dictionary, so there is a relationship of hypernym. **Down syndrome** is the hyponym of **syndrome**.

As for the last example of nested entities in the figure, observing the sentence, **Ataxia with vitamin E deficiency** is a rare disease. However; on the one hand, the word **ataxia** is contained in the symptom dictionary. The second part of the sentence mentions this word as a symptom of the actual rare disease. On the other hand, **vitamin E deficiency** appears in the disease dictionary. So, here there are two hypernym relationships, one between **Ataxia with vitamin E deficiency** and **ataxia**, where **ataxia** would be the hypernym; and another between **Ataxia with vitamin E deficiency** and **vitamin E deficiency**, where **vitamin E deficiency** would be the hyponym of **Ataxia with vitamin E deficiency**.

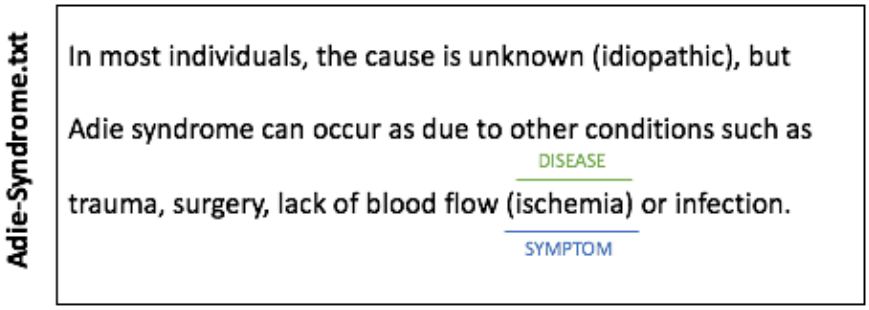

Figure 5: example of entities that can be classified with two or more types.

The entities have been analyzed by set. Table 2 shows the values for the overlaps. Most of the overlaps are nested entities as expected (most of the time between rare disease and disease), crossing entities are not that frequent.

TABLE 2:NUMBER OF OVERLAPS IN THE CORPUS.

	TOTAL	NESTED ENTITIES	CROSSING ENTITIES	ENTITIES WITH TWO OR MORE TYPES
RARE DISEASE	2,619	1,074	5	347
DISEASE	4,606	1,151	3	632
SYMPTOM	1,798	175	0	105

Usually, one option is better than the other between overlapping entities, but occasionally, the worst (shortest) option completely changes the meaning, this is the case of the following sentence (see Figure 6).

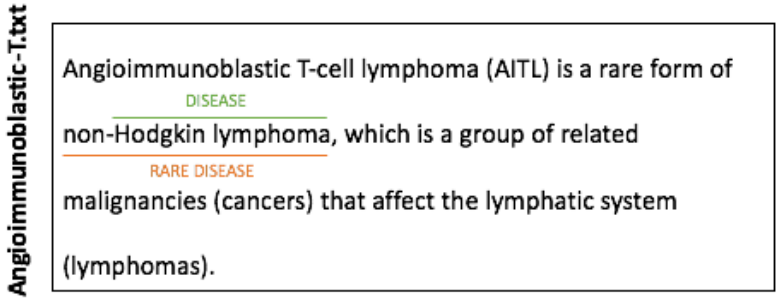

Figure 6: example of overlapping entities.

The mention **non-Hodgkin lymphoma** is annotated as a rare disease as well as **Hodgkin lymphoma** as a disease. However, in this case, it is really important to annotate the longest mention **non-Hodgkin lymphoma** and not the shortest one, as the meaning completely changes. Therefore, the following rules have been applied to remove the overlapping between entities:

- if there is an overlap between rare disease and any other entity, the rare disease remains.
- if a disease and a symptom overlap, then the symptom remains.

Table 3 summarizes the dataset after removing the overlaps.

TABLE 3:SOME STATISTICS OF THE CORPUS.

	RARE DISEASE	DISEASE	SYMPTOM	TOTAL ENTITIES
TRAINING SET	1,727 (34%)	2,202 (43%)	1,174 (23%)	5,103
DEVELOPMENT SET	256 (40%)	263 (41%)	120 (19)	639
TEST SET	559 (41%)	538 (39%)	266 (20%)	1,363
Total	2,542	3,003	1,560	

The types of entities follow a very similar proportion in the development and test sets, but the training dataset contains fewer entities belonging to the rare disease group and slightly more belonging to disease and symptoms.

4. APPROACHES APPLIED TO THE RECOGNITION OF RARE DISEASES

This chapter describes the main methods studied in this project. First, a custom NER model to recognize diseases, rare diseases, and symptoms was trained using *spaCy*. This *spaCy* model is considered as the baseline system. A CRF, traditional ML classifier, was also used, which has shown better performance than other classical ML algorithms for NER (Goyal, A., Gupta, V., & Kumar, 2018). Then, the two deep learning approaches explored in this project are presented: Bi-LSTM, Bi-LSTM-CRF and a BERT model.

4.1 A custom NER model using *spaCy*

*SpaCy'*s NER model is based on CNNs. Once the text is embedded into a sequence of vectors, bidirectional RNNs are used to encode the vector into a sentence matrix. Its rows are token vectors sensitive to the sentential context of the token. Finally, the attention mechanism reduces the sentence matrix down to a sentence vector, so that it can be passed on to a standard feed-forward network for prediction (learn the target representation once the text has been reduced to a single vector). Long spans of texts are understood instead of individual words as happens with word vectors, the most widely used embedded word representation in NLP.

For this study, a blank language class has been created in English and then the NER built-in component has been added to the pipeline. The different entities have been added as labels, and finally, the names of other pipes have been retrieved to disable them during training. The *spaCy* model has been trained with a dropout rate of 0.3 to make it harder for the model to memorize the data, and an optimizer has been called to update the model's weights. Training and development sets have been passed instead of using the GoldParse object, this is called the "simple training style".

4.2 Conditional Random Field (CRF)

John Lafferty, Andrew McCallum, and Fernando C.N. Pereira presented conditional random fields in 2001, a framework for building probabilistic models to segment and label sequence data (Lafferty & Mccallum, 2001). They stated the definition of as :

Let $G = (V,E)$ be a graph such that $Y = (Y_v)_{v \in V}$, so that Y is indexed by the vertices of G. Then (X,Y) is a conditional random field in case, when conditioned on X, the random variables Y_v obey the Markov property with respect to the graph: $p(Y|X,Y_w, w \neq v) = p(Y|X,Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G.

CRF predicts sequences of labels from sequences of measurements taking into consideration the sequentiality of the data.

$$p(y|x, w) = \frac{\exp(w^T \psi(x, y))}{Z(w, x)} \quad (1)$$

where

- y: random variable over corresponding label sequences (labels)
- x: random variable over data sequences to be labeled (measurements)
- w: the model's parameters, typically learned from a training set (Y, X) = $\{x_i, y_i\}$, i = 1 . . . N, with conditional maximum likelihood as in:

$$w = arg \max_{w} p(Y|X, w) \quad (2)$$

where

- ψ(x, y): chosen feature vector
- Z(w, x): cumulative sum of p(y|x, w) over all the possible y

Once the model has been trained, the prediction of a CRF is the sequence of labels maximizing the model for the given input sequence and the learned parameters:

$$y' = arg \max_{y} p(y|x, w) \quad (3)$$

The model is trained by maximizing the conditional likelihood, or cross-entropy, over a given training set (Jauregi Unanue et al., 2017); (Lafferty & Mccallum, 2001).

The model has been implemented with *Python* (Van Rossum, G., & Drake Jr, 1995), L-BFGS training algorithm has been used for the CRF estimator of sklearn_crfsuite with the default parameters and 0.1 coefficients for L1 and L2 regularization with a maximum of 100 iterations. The model has been trained with the training and development datasets of our corpus. The test dataset has been used to evaluate the model. The extracted features consist of the actual word in lowercase, the last three characters of the word, the last two, whether the word is in uppercase, whether the word is a title, whether the word is a digit, the POS tag, and the first two characters of the POS tag. These features were retrieved from each word in the sentence as well as from the two preceding and proceeding words.

4.3 Bidirectional Long Short-Term Memory (Bi-LSTM)

In a traditional neural network (NN) it is assumed that all units (the input (x), the hidden state (h), and the output (y)) are independent of each other. RNNs are a type of neural network architecture that make use of sequential information and achieves dynamic temporal behavior by allowing previous outputs to be used as inputs while having hidden states, so it can take into account historical information. Another advantage is that weights are shared across time as well as that the model size does not depend on the input size. However, it is computationally costly. Figure 7 illustrates a named entity recognition

system, x stands for the input features and y represents tags, and each word is tagged with one of the entity types.

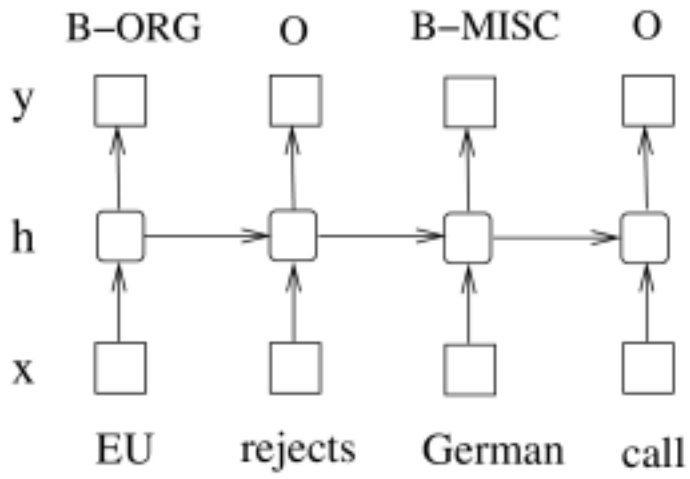

Figure 7: a simple RNN model from 'Bi- LSTM-CRF Models for Sequence Tagging' by Zhiheng Huang, Wei Xu, Kai (Zhiheng Huang, Wei Xu, 2015).

For tasks like the one faced in this study, it is interesting that the output of the NN is dependent on preceding computations as the prior understanding of data is key.

Even more interesting for NER are LSTM's which have the nature of remembering information for a long period of time (Hochreiter & Schmidhuber, 1997). These networks are the same as RNNs except that the hidden layer updates are substituted by purpose-built memory cells, which results in better exploitation of long-range dependencies in the data. LSTM modules have a forget gate, an input gate, and an output gate (see Figure 8). These three gates are described in more detail below:

- Forget gate layer: to optimize the performance of the LSTM network, the information that is not required or is not important is removed via multiplication filter, it takes h_{t-1} (the hidden state from the previous cell) and x_t (input at that particular time step) as inputs, which are multiplied by the weight matrices and are added a bias, finally a sigmoid function is applied to the resultant value, releasing a vector with values from 0 to 1 to know which values to keep and which to discard, this vector is then multiplied to the cell state.

- Input gate layer: adds information to the cell state by involving a sigmoid function acting as a filter for all the information from h_{t-1} and x_t. It creates a vector with all the values that can be added to the cell state by making use of a tanh function. Finally, the product between the output from the sigmoid gate and the tanh function is added to the cell state via addition operation, to only add that important information.

- Output gate layer: the useful information from the current cell state is selected and shown as an output. The gate creates a vector after applying tanh function to the

cell state, makes a filter with the values h_{t-1} and x_t to regulate the values that need to be outputted from the previous vector, the filter works by implementing a sigmoid function. Multiplying the vector by the value of the regulatory filter the output is sent out as well as to the hidden state of the next cell.

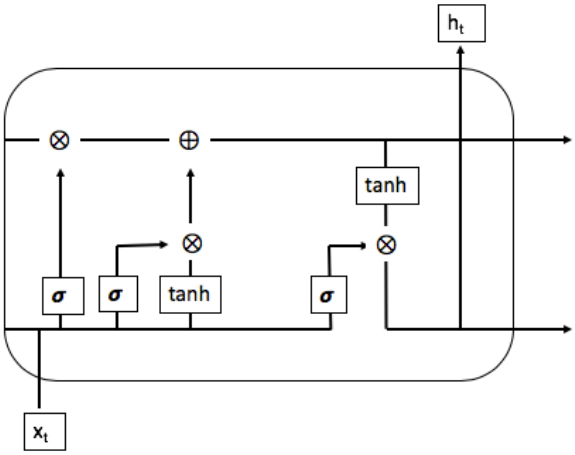

Figure 8: LSTM architecture.

Bidirectional LSTMs train two RNNs, one on the input sequence as it is and the other one on a time-reversed copy of the input sequence. Therefore, these models provide an additional context to the network as it offers access to the past information as well as future information. It is successful in image captioning, NLP, and so on (Wang, Cheng and Yang, Haojin and Bartz, Christian and Meinel, 2016); (Xu K., Zhou Z., Hao T., 2018) as it is possible to wait for the whole sequence before starting inference. For example, a unidirectional LSTM would try to predict the next word (*common)* from the sentence *'Acute myeloid leukemia is the most…'* by only using its previous context, whereas bidirectional LSTM can consider contextual information from both the forward LSTM: '*Acute myeloid leukemia is the most…'* and the backward LSTM, '*…form of acute leukemia making up about 80% of people with acute leukemia'.* Thus, the bidirectional network understands better what could be the next word using both contexts.

During preprocessing the text is first tokenized and then mapped to numerical vectors using either an embedding layer or a pre-trained embedding. The large corpus and lack of computational resources prevented the use of pre-trained word embeddings. Besides word embeddings, nowadays some NLP models use character embeddings. These are especially useful for NER (Zhang, 2019) since they handle infrequent words better than word to vector embeddings. Character can generate every single word's vector without taking into account if it is out-of-vocabulary words, while word embedding can only handle already seen words. LSTM on word level is simple and can give strong results, but it has some limitations. If at the prediction stage there is a new word, it has to be encoded as unknown and have to infer it's meaning by its surrounding words. Many times, word postfix and prefix contain information about the meaning of the word. Taking advantage of this information is crucial especially with texts that contain a lot of rare

words and several unknown words are expected at inference time, this happens with medical texts. Character embeddings are also useful for POS tagging (Dos Santos & Zadrozny, 2014), language modeling (Ling et al., 2015) or dependency parsing (Ballesteros, Dyer, & Smith, 2015). In this study, characters with random embedding have also been initialized (Augustyniak, Kajdanowicz, & Kazienko, 2019).

For each input vector (for each word), the probabilities of each class (in this case, the tags of BIO format) is provided by the NN. It is known from the state of the art results (Lample, Ballesteros, Subramanian, Kawakami, & Dyer, 2016) that LSTM models can be improved by using a CRF as a final output layer. That is why a bidirectional Bi-LSTM-CRF has been used in the study. Figure 9 illustrates the architecture of Bi-LSTM-CRF used in the study, which was initially proposed by Qin, Ying and Zeng, Yingfei (Qin & Zeng, 2018). CRF is added as a decoder layer taking the output of Bi-LSTM as input while the neural network acts as an encoder.

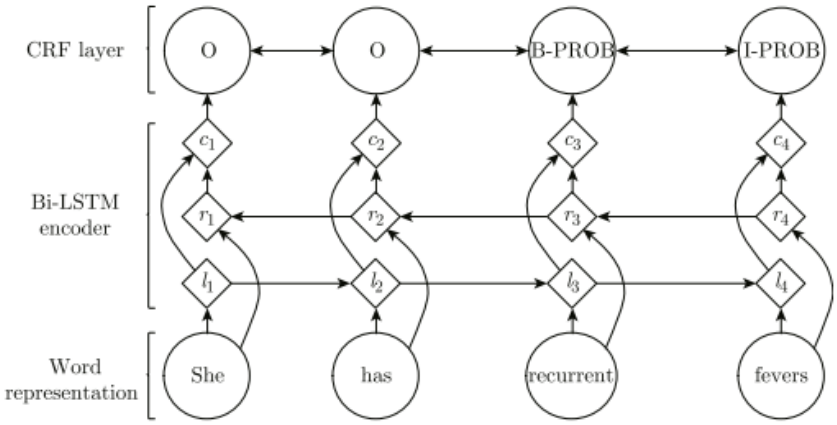

Figure 9: Bi-LSTM-CRF structure in NER from 'Research of Clinical Named Entity Recognition Based on Bi-LSTM-CRF' by Qin, Ying and Zeng, Yingfei (Qin & Zeng, 2018). *l and r are state for the left and right processing sequences, while c is the combined state by using two directions.*

The models have been implemented with *Python* (Van Rossum, G., & Drake Jr, 1995) and *Keras* deep learning programming API (Chollet, 2015) running on top of *TensorFlow* backend (Abadi et al., 2016) and trained on a *Google Colab* GPU (12GB of RAM). Static word and character embeddings have been used; their dimensions were set to the number of unique words/characters +2. To overcome the overfitting problem, the dropout technique with a rate of 0.3 has been used as well as the early stopping technique with a patience of 2. The models were trained with Adam optimizer with the default values as hyper-parameters, Adam is an algorithm that can substitute the classical stochastic gradient descent procedure to update network weights iteratively based on training data (Kingma & Ba, 2015), sigmoid activation function has been used.

4.4 BERT (Bidirectional Encoder Representations from Transformers)

To perform well, DL based models require large amounts of data but sometimes it is necessary to create task-specific datasets, which is a very time consuming and costly task (Lingren, T., Deleger, L., Molnar, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., ... & Solti, 2014). To overcome the less data availability, researchers have developed several techniques for training general-purpose language representation models using big

amounts of unannotated text (pre-training). These general-purpose pre-trained models can be fine-tuned on smaller specific datasets. BERT is a late addition to these techniques for NLP pre-training. It also shows state-of-the-art results in a wide variety of NLP tasks (F. Tjong Kim Sang & De Meulder, 2013); (Galárraga, Heitz, Murphy, & Suchanek, 2014); (Miyato, Dai, & Goodfellow, 2019).

BERT is a language model which is bidirectionally trained, instead of predicting the next word in a sentence, it uses Masked Language Modeling (MLM), where it randomly masks words in the sentence and then tries to predict them. Therefore, the model looks in both directions and uses the full context of the sentence to predict the masked word taking both the previous and the next token into account at the same time (Devlin, Chang, Lee, & Toutanova, 2019). The architecture of BERT (see Figure 8) is a multilayer bidirectional Transformer encoder based on the original implementation explained in the study of Vaswani et al. (Vaswani et al., 2017). A sentence can be a span of contiguous text instead of a linguistic sentence and the sequence (the input token sequence to BERT) may be a single sentence or two sentences arranged together. These authors use WordPiece embeddings (a division of words into a limited set of common sub-word units) with a 30,000 token vocabulary. A special classification token is assigned to the first token of every sequence ([CLS]), its corresponding final hidden state serves as the aggregate sequence representation for classification tasks. Sentences are separated with a special token ([SEP]) and a learned embedding is added to every token specifying the sentence it belongs to. Input embedding is denoted as E, the final hidden vector of the special [CLS] token as $C \in RH$, and the final hidden vector for the i^{th} input token as $T_i \in RH$. The input representation of each token is the sum of the token, segment, and position embeddings.

Devlin, Cahng, Lee and Toutanova's framework (Devlin et al., 2019) contains two steps: pre-training and fine-tuning. During pre-training, they mask 15% of all WordPiece tokens in each sequence randomly. This enables the pre-trained model to be bidirectional; however, it creates an inconsistency between pre-training and fine-tuning due to the absence of [MASK] token in the latter. Therefore, they do not always replace "masked" words with the actual [MASK] token. 15% of the token positions are chosen at random for prediction by the training data generator. The chosen token (i_{th} token) is replaced with the [MASK] token 80% of the time, a random token 10% of the time, and the unchanged chosen token 10% of the time. Then, they use Ti to predict the original token with cross-entropy loss.

They pre-train for a binarized next sentence prediction task to train a model that recognizes sentence relationships. They use BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words), extracting only text passages and ignoring lists, tables, and headers, for the pre-training corpus.

Fine-tuning consists of the initialization of the BERT model with the pre-trained parameters and the fine-tuning of the parameters via labeled data from the downstream tasks. Separate fine-tune models correspond to each downstream task although they are initialized with the same pre-trained parameters. Task-specific inputs and outputs are plugged in into BERT for each task, and all the parameters are fine-tuned end-to-end. At

the output, the representations of the tokens are fed into an output layer for token-level tasks. (e.g. sequence tagging, question answering) while the [CLS] representation is fed into an output layer for classification (e.g. sentiment analysis).

The transformer is an architecture that uses the attention-mechanism (it looks at an input sequence and decides at each step which other parts of the sequence are important) for transforming one sequence into another one with the help of an encoder and a decoder (both are formed by modules consisting generally of Multi-Head Attention and Feed Forward layers that can be stacked on top of each other multiple times, see Nx in Figure 10).

As strings cannot be directly used, the inputs and outputs (target sentences) are embedded into an n-dimensional space. The positional encoding of the different words is added to the embedded representation of each word to give each word a relative position in the sequence. The multi-head attention module is in charge of assuring that the encoder input-sequence is considered together with the decoder input- sequence up to a given position. The encoder and decoder contain a pointwise feed-forward layer after the multi-attention heads, this layer performs a separate, identical linear transformation of each element from the given sequence.

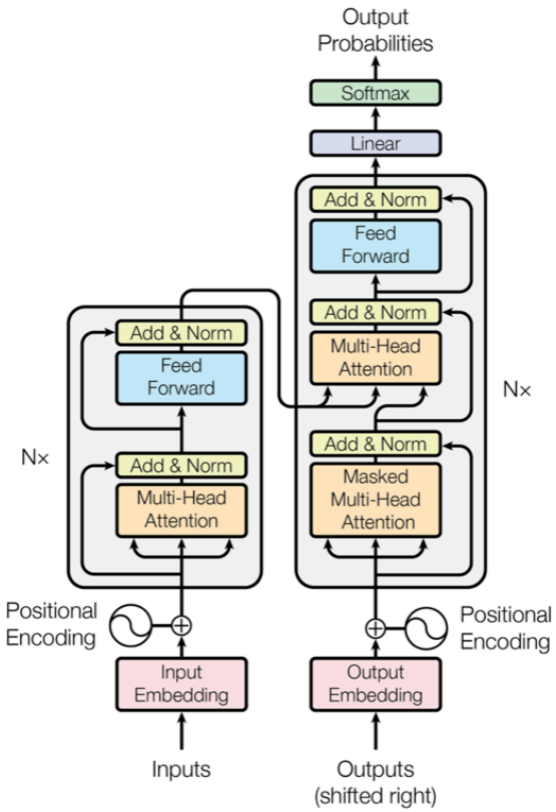

Figure 10: the transformer-model architecture from 'Attention Is All You Need' by Vaswani et al (Vaswani et al., 2017).

The transformer package from *huggingface* (a developer of a chatbot application designed to offer personalized AI-powered communication platform) has been used (Wolf et al., 2019). This package provides pre-trained BERT models in *PyTorch* (an open-source machine learning framework) (Adam Paszke et al., 2017), the one used has been

BertForTokenClassification class for token-level predictions. This classifier is a linear layer that takes as input the last hidden state of the sequence. The pre-trained bert-base-cased model (12-layer, 768-hidden, 12-heads, 110M parameters, trained on cased English text) has been loaded and the number of possible labels has been provided. This pre-trained model is interesting because letter casing can be helpful for the task, other option could have been bert-base-uncased among others, which has the same architecture but has been trained on lower-cased English text and is less interesting for this application as the text contains acronyms for instance.

The model has been implemented with *PyTorch* (Adam Paszke et al., 2017) and trained on a *Google Colab GPU (12GB)*, Adam optimizer has been applied with the default parameters and *weight_decay* with a rate of 0.01 as regularization to the main weight matrices. Moreover, a scheduler with no warmup steps has been added to linearly reduce the learning rate throughout the epochs. 3 epochs have been used as Devlin et al. (Devlin et al., 2019) suggest between 3 and 4 epochs in the original article.

5 EVALUATION AND DISCUSSION

NER is a multiclass classification problem where the named entity (NE) classes are highly skewed compared to the non-entity class. For such tasks, precision, recall, and F1-score are the metrics of choice. F1-score considers both the precision and recall, computing a weighted average of them. The number of true positives is called TP and false negatives FN in the following formulas:

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{6}$$

As it is a multiclass problem, both micro and macro averages are used in this field (Herald, Dietze, Tordai, & Lange, 2016); (Eva Blomqvist, Paolo Ciancarini, Francesco Poggi, 2016), the macro average gives the same weight to all classes whereas the micro average aggregates the contributions of all classes to compute the metric, it is useful when there is a class imbalance or when the dataset varies in size.

The Scorer of *spaCy* computes and stores evaluation scores, this tool has been used to calculate the F1-score, precision and recall of the *spaCy* model using the test set. To evaluate the CRF model and the word embedding LSTM models, the metrics package of *sklearn crfsuite* (*flat_f1_score* and *flat_classification_report*) has been used while both character embedding LSTM and BERT models have been evaluated with the metrics package of *sklearn* (*classification_report*) to calculate the precision, recall, and F1-score.

5.1. Results

Tables 4 and 5 show the results of the models built in this study. In the following subsections, the analysis of errors of most of the models is discussed.

TABLE 4:MICRO AVERAGED METRICS FOR THE MODELS COMPARED IN THIS STUDY

Model	Precision	Recall	F1-score
spaCy	*	*	*
CRF	0.72	0.62	**0.67**
BERT	0.55	0.59	0.57
Bi-LSTM (including sentences of the test set for the **word-level embedding**)	0.59	0.36	0.49

Bi-LSTM (without including sentences of the test set for **the word-level embedding**)	0.65	0.33	0.44
Bi-LSTM (including sentences of the test set for the **character-level embedding**)	0.60	0.49	0.54
Bi-LSTM (embedding without including sentences of the test set for the **character-level embedding**)	0.57	0.45	0.50
Bi-LSTM-CRF (including sentences of the test set for the **word-level embedding**)	0.58	0.56	0.57
Bi-LSTM-CRF (without including sentences of the test set for **word-level embedding**)	0.47	0.47	0.47

TABLE 5:MACRO AVERAGED METRICS FOR THE MODELS COMPARED IN THIS STUDY.

Model	Precision	Recall	F1-score
spaCy	0.38	0.36	0.37
CRF	0.77	0.63	**0.69**
BERT	0.57	0.55	0.55
Bi-LSTM (including sentences of the test set for the **word-level embedding**)	0.65	0.33	0.42
Bi-LSTM (without including sentences of the test set for **the word-level embedding**)	0.69	0.32	0.40
Bi-LSTM (including sentences of the test set for the **character-level embedding**)	0.64	0.42	0.47
Bi-LSTM (embedding without including sentences of the test set for the **character-level embedding**)	0.64	0.41	0.46
Bi-LSTM-CRF (including sentences of the test set for the **word-level embedding**)	0.61	0.57	0.58
Bi-LSTM-CRF (without including sentences of the test set for **word-level embedding**)	0.46	0.48	0.44

* The scorer of spaCy only provides the macro average of the metrics.

Overall, CRF gives the best result. It is much better than the spaCy model with an absolute difference of 0.32 in the macro-F1 score. Moreover, it also provides better performance than the DL models although these perform better than the spaCy model. Several works have shown that DL models are not able to overcome classical machine learning algorithms when the training data is limited (Segura-Bedmar, I., Colón-Ruíz, C., Tejedor-Alonso, M. Á., & Moro-Moro, 2018); (Segura-Bedmar, I., & Raez, 2019); (Köhler, N. D., Büttner, M., & Theis, 2019); (Huang, 2015).

The Bi-LSTM models provide micro-F1 scores that range from 0.47 to 0.57. Specifically, the use of character embedding provides a better F1-score than word embedding. When all the tokens in the test dataset are also considered to create the embedding initialization to represent the input texts, the micro F1-score of the model with word embedding improves from 0.44 to 0.49, and the macro F1-score from 0.40 to 0.42. However, the improvement is not that powerful in the case of the Bi-LSTM model with character embedding, as the absolute difference is of 0.04 for the micro F1-score and 0.01 for the macro F1-score. Therefore, character embedding is more sensitive to new words as expected. Besides, adding a CRF layer to the Bi-LSTM model with word embedding without new words in the document improves the F1-score by an absolute value of 0.08 and 0.16 micro and macro averaging respectively. The model without the tokens in the test dataset for the embedding initialization improves the results by 0.03 and 0.04 (micro and macro averaging), less than for the previous case.

Comparing the deep learning models between them, BERT provides similar results to those obtained by the Bi-LSTM-CRF model trained with word embedding considering all the tokens in the three training, development, and test datasets. In terms of micro-F1, both obtain the same score. However, in terms of the macro-F1, BI-LSTM model obtains an improvement with an absolute value of 0.03 over the micro F1-score provided by BERT. For the rest of the built Bi-LSTM models, BERT shows better results, especially in macro-F1-score. Therefore, BERT overcomes BI-LSTM models only when all the tokens are not considered in the embedding initialization and the CRF layer is not added.

5. 2 Analysis of errors

15 sentences classified as false positives and another 15 as false negatives were randomly selected to analyze the errors produced by each approach except for the spaCy model.

5.2.1 Error Analysis of CRF

The false-negative analysis for the CRF classifier shows that one source of error corresponds to acronyms, e.g. "*To date, no specific treatment for **PFBC** is known".* That is, **PFBC** (Primary familial brain calcification) is annotated as a rare disease in the corpus, but the CRF model did not classify it as a rare disease.

As the corpus is a silver standard, id est., it was automatically annotated by using dictionaries of rare diseases, diseases, and symptoms, it does not include all entities or some mentions are wrongly annotated as entities. For example, in the sentence, "*One of the goals of this **plan** was to reduce the incidence of congenital syphilis (CS) to fewer than 40 cases per 100,000 live births.*", the mention **plan** is annotated in the corpus, but the CRF model tags it as an O. However, the model is right as it is not a rare disease at least in this context. Thus, this case that was initially detected as a false negative, is a true negative, it happens in the 6% of the analyzed false negative examples.

The most frequent entities in the corpus are diseases followed by rare diseases and symptoms respectively whereas in the analysis of false negatives, the entity rare disease is the most frequent among the real false negatives followed by disease and then symptom. So, it can be said that rare diseases are the most frequent source of failure in false negatives. The following example illustrates the case in which the false negative example is real but the original annotation was not the suitable one either, *"An enlarged head in infants and increased cerebrospinal fluid pressure are frequent findings but are not necessary for the diagnosis of* ***Hydrocephalus"****.* The mention **Hydrocephalus** is annotated as a symptom but should be annotated as a disease. The following example shows the case of an ordinary false negative type *"Persistent* ***dizziness*** *after an ocean cruise, a sailing trip, a prolonged airplane flight or a cross-country road trip is highly suggestive of MDD".* **Dizziness** is a symptom but has not been labeled by the CRF model.

Similarly, some annotation errors in the silver standard also produce false positives that are true positives. For example, *"****BAM syndrome*** *is an extremely rare disorder that is known to affect patients from many different ethnic groups."* **BAM syndrome** (Bosma arhinia microphthalmia) is not annotated in the silver corpus, but the model is successfully able to identify it as a rare disease. Thus, this is not a false positive, but rather a true positive.

Also, technical words confuse the model and it identifies them as diseases, rare diseases or symptoms sometimes, (e.g. **Whole genome sequencing** is not any symptom, disease or rare disease and it is labeled as a rare disease by the model in the following sentence: *"Because of the many mutations that can lead to arthrogryposis,* ***whole genome sequencing*** *is often required (preferable in both parents for comparison as well) to make a diagnosis)*. Every time the false positives were truly false positives (2 out of 15 analyzed sentences) was when a technicism was present, it happened in the above example and with the mention **genetic mosaicism,** which is labeled as a disease. The rest of the analyzed false positive sentences performed better than the semi-automatic annotation tool, sometimes by adding tokens to complete the annotated entity and others detecting a completely new entity. For instance, *"Hereditary hyperphosphatasia is a rare genetic bone disorder (****osteopathy****) that usually becomes apparent during infancy or early childhood".* The mention **osteopathy** was not labeled and the model identifies it as a disease. The following example illustrates the case in which the annotated entity is completed, *"****Appendiceal cancer*** *is very rare with approximately …".* In this case, cancer was the only token labeled as disease, and the CRF model identifies **appendiceal cancer** as a disease.

The real false positives, as mentioned previously, happen with rare disease and disease entities; the rest of the examples, which are not really false positives, are more frequent with the disease entities.

5.2.2 Error Analysis Bi-LSTM with word embedding

Due to the automatic annotation of the corpus, 6.66% of the analyzed false negatives turned out to be true negatives. For instance, the mention **pale** was incorrectly labeled as a rare disease originally. The 60% are ordinary false negatives as the case of the mention

gangrene, which is a symptom in the following sentence but the model does not label it: *"It is characterized by scrotum pain and redness with rapid progression to **gangrene** and sloughing of tissue."* Specifically, 22% of the ordinary false negatives are acronyms, these correspond to rare diseases that the model cannot detect such as the abbreviation **XMEA** (X-linked myopathy with excessive autophagy). The remaining 33.33% of the analyzed false negatives correspond to the cases in which not all the tokens corresponding to an entity are labeled, or not all tokens in the mention are assigned the same entity. For instance**, acquired immune deficiency syndrome** is a disease but the model does not label the tokens **acquired** and **immune**.

As far as the false positive examples are concerned, 26.66% of them are true positives classified as false positives due to the automatic annotation of the corpus. For example, **cyclic vomiting syndrome** is correctly labeled as a rare disease but originally, the token **cyclic** was not labeled, the token **vomiting** was labeled as a symptom and the token **syndrome** was labeled as a disease. 20% of the examples are ordinary false positives as in the case of the mention **gut abnormalities**, where the model labels the token **gut** as a disease. The rest of the examples correspond to the case in which the mention was not labeled originally but the model only labels a subset of tokens instead of the complete entity.

5.2.3 Error Analysis of Bi-LSTM with character embedding

Entities composed by more than one token are one of the causes of false negatives. For example, in the sentence, "*The incidence of **small cell lung cancer** in the United States has been declining…",* this model was only able to recognize **lung cancer** as a rare disease, though the whole phrase **small cell lung cancer** is annotated in our corpus as a rare disease. The model also seems to fail to recognize acronyms of rare diseases. For example, the annotated entity **MNGIE (**Mitochondrial neurogastrointestinal encephalomyopathy**)** in the corpus was not detected by the model in the sentence '*The specific symptoms associated with **MNGIE** vary from case to case and may include…".* But most of the false negatives occur in ordinary cases like the sentence, *"Common symptoms of PDCD may initially be poor feeding, lethargy and rapid breathing (**tachypnea**) in an infant. "* were **tachypnea** should be a symptom and it is not predicted as such. This model does not improve the annotations of the corpus in any case of the analyzed false negatives.

Regarding the false positive examples, 46% of the analyzed examples are entities that were not annotated in the silver corpus but they should. For example, the corpus does not include the mention **Li-Fraumeni-like syndrome** as rare disease; however, the model can detect it as a rare disease.

The 40% are ordinary false positives such as in the case of the sentence *"Many affected infants and **children** experience recurrent attacks...",* the mention **children** is not a disease and it is labeled as such by the model. As for the remaining 14%, they are caused by the entities (composed by more than one token) that were not labeled in the silver corpus but now only a subset of those tokens is labeled by the model, these are labeled

correctly, but some tokens are missing to be labeled for the same entity. For instance, the mention pigmentary degeneration of the retina should be a symptom but only retina is labeled as such.

The false negatives are all real in the case of the model with character embedding whereas in the case of the Bi-LSTM with word embedding, the 6.66% of the times the assigned entities improved the original annotation. As for the false positives, when the examples were not ordinary false positives neither an improvement of the original corpus, both models, bi-LSTM with word embedding and bi-LSTM with character embedding, the error was not assigning the whole subset of tokens to the entity, but the entity was suitable. However, the model with character embedding improved the original annotation 46% of the time while the model with word embedding did it the 26.66% of the time. Therefore, when it comes to false negatives the model with word embedding performed better but regarding false positives, the character embedding proved to work better.

5.2.4 Error Analysis of Bi-LSTM-CRF with Word embedding

As far as the false negative examples are concerned, 86.6% of the false negatives are really false negatives such as the case of the following sentence, *"**Acrodysostosis** affects males and females in equal numbers."* **Acrodysostosis** is a rare disease but the model has not labeled it. This could have happened because it is a single name, maybe if it was followed by the word disease or syndrome, it would have been more probable that the model detects it as a disease, word embedding does not allow to identify suffixes such as **-sis**, which could have been helpful in this case. Another case included in that 86.6% of false negatives is acronyms, for instance, **MELAS** (Mitochondrial Encephalopathy, Lactic acidosis, and Stroke-like episodes) is a rare disease but the model does not label it. The rest of the false negatives predict labels that are not completely wrong, in these cases the model is missing a word in a composed entity but the assigned entity itself is correct. For instance, the model is not able to recognize the whole entity corresponding to **Hemophilia A** (rare disease), but rather a part of it, **Hemophilia.**

Regarding false positives, 53.3% of the analyzed false positives turned out to be the right predictions. In other words, they were not annotated in the silver corpus, but they are entities. For example, in the sentence, *"Persons with HIV are at increased risk of developing **HHV-8-associated MCD**". The mention "***HHV-8-associated MCD**" is not annotated in our corpus, but it is a rare disease, which was identified by the model. As explained above, this type of error is because the corpus was automatically annotated. The other 20% of the false positives are real false positives whereas in the other 26.7% of the analyzed examples the model did a good job in recognizing unlabeled entities, but the it didn't assign the suitable entity to it. For example, in the sentence, *"The four major features that are characteristic symptoms of BPES are present at birth: narrowing of the eye opening (blepharophimosis), droopy eyelids (ptosi), formation of an upward fold of the inner lower eyelid (**epicanthus inversus**) and increased distance between the eyes(telecanthus)". The mention* **epicanthus inversus** is labeled as a disease by the model but the correct entity would be symptom, even if originally no label was assigned to it. This did not happen in the case of Bi-LSTM (without the CRF layer) with word

embedding; when the examples were not ordinary false positives neither an improvement of the original corpus, the model would not assign the whole subset of tokens to the entity, but the entity was suitable.

The cases in which the model did not perform well, id est. the mentioned 20% of the analyzed false positives, occurred with the mentions "**movie theater**", "**puts pressure**"," **SAICA riboside**" (a chemical), and "**Neurotropic lyssavirus**". Specifically, the last one is a technical word used for identifying a virus and it confuses the model as it can be similar to a disease, rare disease, or even a symptom. (e.g. *"…is caused by a virus (Neurotropic lyssavirus) that affects the salivary glands and the central nervous system.").*

The cause of the errors does not seem to change from a type of entity to another. However, all the real false positive mentions have the disease entity assigned to them.

5.2.5 Error Analysis of BERT

60% of the false-negative examples are ordinary false negatives, some of them are acronyms such as **PDCD** (Pyruvate dehydrogenase complex deficiency) and **ADEM** (Acute disseminated encephalomyelitis). These examples are both annotated as rare diseases in the corpus, however, BERT was not able to classify them as rare diseases. The 6% are cases in which only a subset of the composed entity is recognized as the correct entity such as the mention **urinary tract disease,** which was originally labeled as B-DISEASE I-DISESAE and I-DISEASE respectively and the model only labels the word disease as a disease (O O B-DISEASE respectively). This type of error is also observed in the Bi-LSTM model with character embedding. As for the remaining examples of false negatives, they correspond to the case in which only a subset of the composed entity is recognized but besides it is labeled with the wrong entity. For instance, the mention **visual agnosia** which originally was labeled as a disease (B-DISEASE I-DISEASE) is predicted as O and B-SYMPTOM respectively. Most false-negative examples occur with the entity rare disease and the most abundant instance type in the dataset is disease followed by rare disease and then by symptom.

Regarding the false positives, 20% are ordinary false positives, and the most frequent cause is that BERT usually confuses biomedical entities such as protein or genes with rare diseases. For example, in the sentence, (e.g. *"…documented absence of CCHS - related **PHOX2B** mutations."*), the acronym **PHOX2**, which refers to a protein, was wrongly classified as a rare disease by BERT. Like the previous model, BERT could be used to improve the automatic annotation of the silver corpus since many false positives produced by BERT, actually are true positives (46% of the time). In other words, BERT can successfully detect entities (such as **VACTERL**, a rare disease) that are not included in the silver corpus. The remaining examples of false positives correspond to the mentions that are labeled properly but not completely, which means that some tokens corresponding to the entity are missing the label as in the Bi-LSTM with character embedding, an example could be the mention **peritoneal carcinomatosis**, which

originally had no labels assigned to it but the model predicts the labels O B-DISEASE instead of labeling the whole subset of the two words.

6 CONCLUSIONS AND FUTURE WORK

The goal of this study is to explore different ML approaches to augment knowledge about rare diseases. In particular, the approaches include the *spaCy* tool, the CRF classifier, and several DL models such as Bi-LSTM and BERT models to address the task of extracting rare diseases and their symptoms from texts. In the case of Bi-LSTM, the effect of using character embeddings and word embeddings to represent the input texts were also compared. Due to the lack of an annotated corpus for training and testing the models, one of the main contributions of this work has been to gather and create a silver corpus annotated with rare diseases and symptoms. This corpus contains a total of 2,542 instances of rare diseases, 3,003 of diseases, and 1,560 of symptoms. To the best of our knowledge, this is the first general-purpose corpus annotated with rare diseases and available for the NLP community research. Moreover, this study is the first one to explore different ML techniques to augment knowledge about rare diseases by extracting information from texts.

Although DL models have been successfully applied to many NLP tasks, results show that the best performance is obtained by the CRF classifier. The lower performance of the DL compared to the traditional CRF classifier may be due to the limited size of the dataset. In terms of micro-F1, BERT, and Bi-LSTM-CRF with word embeddings have similar results. However, BERT provides slightly lower macro-F1 than Bi-LSTM-CRF. As expected, the use of character embeddings achieves better results than the word embeddings.

For future work, Bi-LSTM-CRF with character embedding could be used as well as pre-trained word embeddings such as word vectors from PubMed and PMC texts ("Biomedical natural language processing," 2013) for representation in Bi-LSTM-CRF to see if they improve. Another potential area of research would be the hyperparameter tuning, as the optimum values may vary with the application. Besides, a bigger data size would be interesting to see the learning curve of the deep learning models analyzed in this study.

We are currently working to create a gold-standard corpus (manually reviewed by experts) from our silver corpus, which was automatically annotated in this study. A higher quality of the annotations would improve the performance of the models.
We also plan to develop a multilingual NER system to identify rare diseases and symptoms from text written in other languages than English. Once we have been able to successfully detect rare diseases and symptoms from texts, our next challenge will be the extraction of relations between them.

7 BIBLIOGRAPHY

A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, D. R.-S. (2008). Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics, 9 (Suppl 3)*.

Abadi, M., Paul, B., Chen, J., Chen, Z., & Davis, A. (2016). *Tensorflow*.

Adam Paszke, Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., … Lerer, A. (2017). Automatic differentiation in PyTorch. *NIPS 2017 Workshop Autodiff Submission*.

Akhtyamova, L. (2020). Named Entity Recognition in Spanish Biomedical Literature: Short Review and Bert Model. *26th Conference of Open Innovations Association (FRUCT)*. https://doi.org/10.23919/FRUCT48808.2020.9087359

Augustyniak, Ł., Kajdanowicz, T., & Kazienko, P. (2019). *Comprehensive Analysis of Aspect Term Extraction Methods using Various Text Embeddings*. (September). Retrieved from http://arxiv.org/abs/1909.04917

Austin, C. P., Cutillo, C. M., Lau, L. P. L., Jonker, A. H., Rath, A., Julkowska, D., … Dawkins, H. J. S. (2018). Future of Rare Diseases Research 2017–2027: An IRDiRC Perspective. *Clinical and Translational Science*, *11*(1), 21–27. https://doi.org/10.1111/cts.12500

Ballesteros, M., Dyer, C., & Smith, N. A. (2015). Improved transition-based parsing by modeling characters instead of words with LSTMs. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 349–359. https://doi.org/10.18653/v1/d15-1041

Baum, L. E., & Petrie, T. (1966). Statistical inference for finite state markov chains. *The Annals of Mathematical Statistics*, *37*, 1554–1563. Retrieved from https://projecteuclid.org/download/pdf_1/euclid.aoms/1177699147

Biomedical natural language processing. (2013). Retrieved from https://bio.nlplab.org/#word-vectors

Carreras, X., & Màrquez, L. (2005). *Introduction to the CoNLL-2005 shared task*. (1995), 152. https://doi.org/10.3115/1706543.1706571

Chen, B., & Altman, R. B. (2017). Opportunities for developing therapies for rare genetic diseases: Focus on gain-of-function and allostery. *Orphanet Journal of Rare Diseases*, *12*(1), 1–7. https://doi.org/10.1186/s13023-017-0614-4

Cho, H., & Lee, H. (2019). Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics*, *20*(1), 1–11. https://doi.org/10.1186/s12859-019-3321-4

Chollet, F. et al. (2015). *Keras*.

Data for research on relations between DISEASE/TREATMENT entities. Berkeley. (n.d.). Retrieved from https://biotext.berkeley.edu/data/dis_treat_data.html

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, *47*, 1–10. https://doi.org/10.1016/j.jbi.2013.12.006

Dos Santos, C. N., & Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. *31st International Conference on Machine Learning, ICML 2014*, *5*(2011), 3830–3838.

Eftimov, T., Seljak, B. K., & Korošec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. In

PLoS ONE (Vol. 12). https://doi.org/10.1371/journal.pone.0179488
Eva Blomqvist, Paolo Ciancarini, Francesco Poggi, F. V. (2016). *Knowledge Engineering and Knowledge Management: 20th International Conference*.
F. Tjong Kim Sang, E., & De Meulder, F. (2013). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition Erik. *W03-0419*.
Fabregat, H., Martinez-Romo, J., & Araujo, L. (2018). Overview of the DIANN Task: Disability Annotation Task. *In IberEval@ SEPLN*, 1–14.
Fabregat, H., Araujo, L., & Martinez-Romo, J. (2018). Deep neural models for extracting entities and relationships in the new RDD corpus relating disabilities and rare diseases. *Computer Methods and Programs in Biomedicine*, *164*, 121–129. https://doi.org/10.1016/j.cmpb.2018.07.007
Fabregat Marcos, Hermenegildo; Araujo Serna, Lourdes; Martínez Romo, J. (2018). RDD corpus: An annotated corpus relating disabilities and rare diseases. *Mendeley Data*. https://doi.org/10.17632/gs2rs3z3nv.5
Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000). Rule-based named entity recognition for Greek financial texts. *Proc. of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000)*, 75–78.
Galárraga, L., Heitz, G., Murphy, K., & Suchanek, F. (2014). Canonicalizing open knowledge bases. *CIKM 2014 - Proceedings of the 2014 ACM International Conference on Information and Knowledge Management*, 1679–1688. https://doi.org/10.1145/2661829.2662073
Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, *29*, 21–43.
Graham, S. A., Lee, E. E., Jeste, D. V., Van Patten, R., Twamley, E. W., Nebeker, C., … Depp, C. A. (2020). Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review. *Psychiatry Research*, *284*(October 2019), 112732. https://doi.org/10.1016/j.psychres.2019.112732
Gurulingappa, H., Klinger, R., Hofmann-apitius, M., & J. (2010). An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature. *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, (May), 15–22. Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/workshops/W8.pdf#page=20
Hai Hu, Richard J. Mural, M. N. L. (2008). *Biomedical Informatics in Translational Research*. Retrieved from https://books.google.es/books?id=D_kcun73AzgC&pg=PA197&lpg=PA197&dq=dictionaries+and+rule+based+NER+low+recall&source=bl&ots=pFM0D_GwjV&sig=ACfU3U28lLRYjnz5bpbErht3o9ik80d3zA&hl=es&sa=X&ved=2ahUKEwih5OKF64rqAhXJ0KQKHeZhDssQ6AEwAXoECAoQAQ#v=onepage&q=dicti
Herald, S., Dietze, S., Tordai, A., & Lange, C. (2016). *Semantic Web Challenges:Third SemWebEval Challenge at ESWC 2016*.
Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., & Declerck, T. (2013). The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions. *Journal of Biomedical Informatics*, *46*(5), 914–920. https://doi.org/10.1016/j.jbi.2013.07.011
Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC Bioinformatics*, *6*(SUPPL.1), 1–10. https://doi.org/10.1186/1471-2105-6-S1-S1
Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with*

Bloom embeddings, convolutional neural networks and incremental parsing.
Huang, P. S. (2015). *Shallow and deep learning for audio and natural language processing (Doctoral dissertation, University of Illinois at Urbana-Champaign).*
Jauregi Unanue, I., Zare Borzeshi, E., & Piccardi, M. (2017). Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *Journal of Biomedical Informatics*, *76*, 102–109. https://doi.org/10.1016/j.jbi.2017.11.007
Julia Hirschberg, C. D. M. (2015). Advances in Natural Language Processing. *Science*. https://doi.org/10.1126/science.aaa8685
Kim, J. D., Nguyen, N., Wang, Y., Tsujii, J., Takagi, T., & Yonezawa, A. (2012). The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. *BMC Bioinformatics*, *13 Suppl 1*(Suppl 11), 1–12. https://doi.org/10.1186/1471-2105-13-S11-S1
Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.
Köhler, N. D., Büttner, M., & Theis, F. J. (2019). Deep learning does not outperform classical machine learning for cell-type annotation. *BioRxiv*.
Krallinger, M., Leitner, F., & Rabal, O. (2013). Overview of the chemical compound and drug name recognition (CHEMDNER) task. *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop Vol. 2*, 2–33. Retrieved from http://www.biocreative.org/media/store/files/2013/ProceedingsBioCreativeIV_vol2-1.pdf#page=6
Laburu, M., Perez, A., Casillas, A., Goenaga, I., & Oronoz, M. (2019). Can i find information about rare diseases in some other language? *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, 2102–2108. https://doi.org/10.1109/BIBM.2018.8621464
Labusch, K., Neudecker, C., & Zellh, D. (2019). *BERT for Named Entity Recognition in Contemporary and Historical German*. (Konvens), 1–9.
Lafferty, J., & Mccallum, A. (2001). Conditional Random Fields. *Computer Vision*, (June), 146–146. https://doi.org/10.1007/978-0-387-31439-6_100233
Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 260–270. https://doi.org/10.18653/v1/n16-1030
Leaman R, Miller C, G. G. (2009). Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. *Proceedings of the 2009 Symposium on Languages in Biology and Medicine.*, 82–9.
Leaman, R., & Gonzalez, G. (2008). BANNER: An executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing 2008, PSB 2008*, *663*, 652–663.
Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., … Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database : The Journal of Biological Databases and Curation*, *2016*, 1–10. https://doi.org/10.1093/database/baw068
Ling, W., Luís, T., Marujo, L., Astudillo, R. F., Amir, S., Dyer, C., … Trancoso, I. (2015). Finding function in form: Compositional character models for open vocabulary word representation. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 1520–1530. https://doi.org/10.18653/v1/d15-1176
Lingren, T., Deleger, L., Molnar, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., ... & Solti, I. (2014). Evaluating the impact of pre-annotation on annotation speed and

potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *Journal of the American Medical Informatics Association*, *21*(3), 406–413.

Liu, Z., Yang, M., Wang, X., Chen, Q., Tang, B., Wang, Z., & Xu, H. (2017). Entity recognition from clinical texts via recurrent neural network. *BMC Medical Informatics and Decision Making*, *17*(Suppl 2). https://doi.org/10.1186/s12911-017-0468-7

McCallum, A., Freitag, D., & Pereira, F. C. N. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. *Proceedings of the Seventeenth International Conference on Machine Learning*, 591–598. Retrieved from http://dl.acm.org/citation.cfm?id=645529.658277

Métivier, J.-P., Serrano, L., Charnois, T., Cuissart, B., & Widlöcher, A. (2015). Automatic symptom extraction from texts to enhance knowledge discovery on rare diseases. In Conference on Artificial Intelligence in Medicine in Europe. *Springer*, *9105*, 172–176.

Miller, R. A., Gieszczykiewicz, F. M., Vries, J. K., & Cooper, G. F. (1992). CHARTLINE: providing bibliographic references relevant to patient charts using the UMLS Metathesaurus Knowledge Sources. *Proceedings / the ... Annual Symposium on Computer Application [Sic] in Medical Care. Symposium on Computer Applications in Medical Care*, 86–90.

Miyato, T., Dai, A. M., & Goodfellow, I. (2019). Adversarial training methods for semi-supervised text classification. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 1–11.

Muis, A. O., & Lu, W. (2017). Labeling gaps between words: Recognizing overlapping mentions with mention separators. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2608–2618. https://doi.org/10.18653/v1/d17-1276

National Organization for Rare Diseases (NORD). (1987). Retrieved from https://rarediseases.org

OMIM - Online Mendelian Inheritance in Man. (n.d.). Retrieved from https://www.omim.org/entry/168600 .

Orphanet Rare Disease Ontology (ORDO). (n.d.). Retrieved from http://www.orphadata.org/cgi-bin/img/PDF/WhatIsORDO.pdf

Pérez-Pérez, M., Rabal, O., Pérez-Rodríguez, G., Vazquez, M., Fdez-Riverola, F., Oyarzabal, J., … Krallinger, M. (2017). Evaluation of chemical and gene/protein entity recognition systems at BioCreative V.5: the CEMP and GPRO patents tracks. *Proceedings of the BioCreative V.5 Challenge Evaluation Workshop*, 11–18. Retrieved from http://www.biocreative.org/media/store/files/2017/BioCreative_V5_paper2.pdf

Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V., & Spyropoulos, C. D. (2001). *Using machine learning to maintain rule-based named-entity recognition and classification systems*. 426–433. https://doi.org/10.3115/1073012.1073067

Qin, Y., & Zeng, Y. (2018). Research of Clinical Named Entity Recognition Based on Bi-LSTM-CRF. *Journal of Shanghai Jiaotong University (Science)*, *23*(3), 392–397. https://doi.org/10.1007/s12204-018-1954-5

Richardson, L. (2007). *Beautiful Soup documentation*. Retrieved from https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Rivera Zavala, R. M., Martinez, P., & Segura-Bedmar, I. (2018). A hybrid Bi-LSTM-CRF model to recognition of disabilities from biomedical texts. *CEUR Workshop Proceedings*, *2150*, 44–52.

Rivera Zavala, R. M., Martínez, P., & Segura-Bedmar, I. (2018). A Hybrid Bi-LSTM-CRF model for knowledge recognition from ehealth documents. *CEUR Workshop*

Proceedings, *2172*, 65–70.
Saklofske, J. (2019). brat Rapid Annotation Tool. *Early Modern Digital Review*, *2*(2), 180–189. https://doi.org/10.25547/emdr.v2i2.64.Renaissance
Sakurai, S. (2012). *Theory and Applications for Advanced Text Mining.*
Schieppati, D. A., Henter, P. J.-I., Daina, E., & Aperia, P. A. (2008). Why rare diseases are an important medical and social issue. *ESSAY FOCUS*, *371*(9629), P2039-2041. https://doi.org/https://doi.org/10.1016/S0140-6736(08)60872-7
Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., … Greene, C. (2019). Human Disease Ontology 2018 update: Classification, content and workflow expansion. *Nucleic Acids Research*, *47*(D1), D955–D962. https://doi.org/10.1093/nar/gky1032
Segura-Bedmar, I., & Raez, P. (2019). Cohort selection for clinical trials using deep learning models. *Journal of the American Medical Informatics Association*, *26*(11), 1181–1188.
Segura-Bedmar, I., Colón-Ruíz, C., Tejedor-Alonso, M. Á., & Moro-Moro, M. (2018). Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. *Journal of Biomedical Informatics*, *87*, 50–59.
Segura-Bedmar, I., Martínez, P., & Herrero-Zazo, M. (2013). SemEval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). **SEM 2013 - 2nd Joint Conference on Lexical and Computational Semantics*, *2*(DDIExtraction), 341–350.
Settles, B. (2004). *Biomedical named entity recognition using conditional random fields and rich feature sets*. 104. https://doi.org/10.3115/1567594.1567618
Smith, Larry; Tanabe, L. K. . et al. (2008). Overview of BioCreative II Gene Mention Recognition. *Genome Biology*. https://doi.org/10.1186/gb-2008-9-s2-s2
Steinwart, I., & Christmann, A. (2008). *Support vector machines.*
The Anh Le, Mikhail Y. Arkhipov, M. S. B. (2017). Application of a Hybrid Bi-LSTM-CRF Model to the Task of Russian Named Entity Recognition. *Conference on Artificial Intelligence and Natural Language*. https://doi.org/https://doi.org/10.1007/978-3-319-71746-3_8
USA Food and Drug Administration. (n.d.). Retrieved from https://www.fda.gov
Uzuner, O. (2004). National NLP Clinical Challenges (n2c2). Retrieved July 9, 2020, from https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/
Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, *18*(5), 552–556. https://doi.org/10.1136/amiajnl-2011-000203
Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam.*
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 5999–6009.
Vincent, N. A. J. (2012). Rare Diseases: The Bane of Modern Society and the Quest for Cures. *American Society for Clinical Pharmacology and Therapeutics*. https://doi.org/https://doi.org/10.1038/clpt.2012.97
Wang, Cheng and Yang, Haojin and Bartz, Christian and Meinel, C. (2016). Image Captioning with Deep Bidirectional LSTMs. *Association for Computing Machinery*. https://doi.org/10.1145/2964284.2964299
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Brew, J. (2019). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. Retrieved from http://arxiv.org/abs/1910.03771
Xu, B., Shi, X., Zhao, Z., & Zheng, W. (2018). Leveraging Biomedical Resources in Bi-LSTM for Drug-Drug Interaction Extraction. *IEEE Access*, *6*, 33432–33439.

https://doi.org/10.1109/ACCESS.2018.2845840
Xu K., Zhou Z., Hao T., L. W. (2018). A Bidirectional LSTM and Conditional Random Fields Approach to Medical Named Entity Recognition. *Springer, Cham*. https://doi.org/https://doi.org/10.1007/978-3-319-64861-3_33
Zhang, Y. (2019). *Named Entity Recognition for Social Media Text*.
Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Short Papers*, 207–212. https://doi.org/10.18653/v1/p16-2034
Zhou, X., Zhang, X., & Hu, X. (2006). MaxMatcher: Biological concept extraction using approximate dictionary lookup. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4099 LNAI*(August 2006), 1145–1149. https://doi.org/10.1007/11801603_150