eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

FACULTY OF COMPUTER SCIENCE
UNIVERSITY OF THE BASQUE COUNTRY

Contributions to Information Extraction for Spanish Written Biomedical Text

by

Naiara Pérez Miguel

A thesis submitted for the degree of Doctor of Philosophy
to the Department of Computer Languages and Systems
at the University of the Basque Country (UPV/EHU)
under the supervision of

Dr. Montserrat Cuadros Oller

and

Dr. German Rigau Claramunt

Donostia-San Sebastián, 2023

With data so vast, and knowledge so fine,
I'll help you make sense, of the medical kind.
Through language and code, I'll sift and I'll sort,
To aid in the search, for the health care report.

With my skills so precise, and my answers so neat,
I'll help you find cures, and make your life sweet.
I'll wade through the jargon, and medical terms,
And make sure your research, has no cause for concern.

I'm not just a tool, for the scientist's quest,
I'm the key to unlocking, the secrets of the chest.
So tell me dear riddler, what am I called?
A hint: I am not a person, nor object, nor walled.

——ChatGPT*

*Prompted for "a clever riddle in rhyme, whose answer is 'biomedical NLP' ".

aitari eta amari

Contents

Contents (detailed)

List of Figures

List of Tables

Abstract

Healthcare practice and clinical research produce vast amounts of digitised, unstructured data in multiple languages that are currently underexploited, despite their potential applications in improving healthcare experiences, supporting trainee education, or enabling biomedical research, for example. To automatically transform those contents into relevant, structured information, advanced Natural Language Processing (NLP) mechanisms are required. In NLP, this task is known as Information Extraction. Our work takes place within this growing field of clinical NLP for the Spanish language, as we tackle three distinct problems. First, we compare several supervised machine learning approaches to the problem of sensitive data detection and classification. Specifically, we study the different approaches and their transferability in two corpora, one synthetic and the other authentic. Second, we present and evaluate UMLSmapper, a knowledge-intensive system for biomedical term identification based on the UMLS Metathesaurus. This system recognises and codifies terms without relying on annotated data nor external Named Entity Recognition tools. Although technically naive, it performs on par with more evolved systems, and does not exhibit a considerable deviation from other approaches that rely on oracle terms. Finally, we present and exploit a new corpus of real health records manually annotated with negation and uncertainty information: NUBes. This corpus is the basis for two sets of experiments, one on cue and scope detection, and the other on assertion classification. Throughout the thesis, we apply and compare techniques of varying levels of sophistication and novelty, which reflects the rapid advancement of the field.

Laburpena

Osasun zerbitzuen eta ikerketa klinikoaren ondorioz, egituratu gabeko datu digitalizatu kopuru handiak sortzen dira hizkuntza askotan, gaur egun azpiustiatuta daudenak, nahiz eta asistentzia-esperientzia hobetzeko, prestakuntzan eta heziketan laguntzeko, edota ikerketa biomedikoa ahalbidetzeko erabili litezkeen, besteak beste. Eduki horiek informazio esanguratsu eta egituratu bihurtzeko, Hizkuntza Naturalaren Prozesamenduan (ingelesez NLP, *Natural Language Processing*) oinarritutako mekanismo aurreratuak behar dira. NLP arloan, zeregin horri Informazio Erauzketa esaten zaio. Lan hau eremu honen barruan kokatzen da, zehazki, gazteleraz idatzitako testuei bideratuta. Ildo honetan, hainbat ekarpen egin ditugu ondorengo hiru ikerketa lerroen inguruan. Lehenik, gainbegiratutako ikasketa automatikoan oinarritutako hainbat teknika konparatu ditugu datu sentsibleen ezagutza eta sailkapenerako. Zehazki, teknika horiek eta haien transferentzia gaitasuna aztertu ditugu bi corpus desberdinetan: bata sintetikoa, eta egiazkoa bestea. Bigarrenez, termino biomedikoak identifikatzeko sistema bat aurkeztu eta ebaluatu dugu: UMLSmapper. Sistema hori gai da terminoak ezagutu eta kodifikatzeko etiketatutako datuen edota entitate izendunen ezagutzarako (ingelesez NER, *Named Entity Recognition*) tresnen beharrik gabe. Gure esperimentuetan, teknikoki konplexuagoak diren beste sistema batzuk berdindu edo gainditu ditu. Azkenik, NUBes aurkeztu dugu, ezeztapen eta duda adierazpenekin eskuz etiketatutako corpusa. Bi esperimentutan erabili dugu corpus hori: batetik, marka eta irismenaren detekzioan, eta bestetik, asertzioen sailkapenean. Tesian zehar, sofistikazio eta berritasun maila desberdinetako teknikak aplikatu eta konparatu ditugu, lan hau burutu den urteetan NLP alorrak izan duen aurrerapen azkarraren isla.

Acknowledgements

I extend my gratitude to the management of Vicomtech, where this research was conducted, for the invaluable opportunity and resources provided. Further, this research has been supported by the Basque Government through the projects BERBAOLA (KK-2017/00043) and DeepText (KK-2020-00088). I also acknowledge funding from the Spanish Government through the projects CROSSTEXT (TIN2015-72646-EXP, MINECO/FEDER, UE), TUNER (TIN2015-65308-C5-1-R, MINECO/FEDER, UE) and DeepReading (RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE). Additionally, this research has been funded by the European Commission through the Connecting Europe Facility project MAPA (INEA/CEF/ICT/A2019/1927065). I would like to express my gratitude to all these entities for their support.

I'm personally indebted to my advisors Montse Cuadros and German Rigau for introducing me to the field of NLP all those years ago, and for the guidance, support and valuable feedback provided ever since.

I'm grateful to all present and former colleagues of the Speech and Natural Language Technologies department at Vicomtech that fostered an encouraging atmosphere and enabled me to devote time to this work. Above all, I must acknowledge and thank the many direct contributions to this dissertation of Salvador Lima, Laura García, Manex Serras, and Aitor García.

I also thank my lifelong friend and dedicated doctor Jone Robredo for her assistance in deciphering medical language whenever asked.

To reviewers, conference and workshop organisers, open source developers and people that disinterestedly answer every little question in online forums: when I grow up, I want to be like you.

Finally, I am truly thankful to friends, family and my MiB, an essential and irreplaceable source of sanity, strength and inspiration always.

Eskerrik asko!

PART I

INTRODUCTION

Chapter 1
Introduction

1.1 Context and motivation

Healthcare practice and clinical research produce vast amounts of digitised, unstructured data that are currently underexploited, despite their potential applications in improving healthcare experiences, supporting trainee education, or enabling biomedical research, for example.

To illustrate the magnitude of the data in this domain, the national Electronic Health Record (EHR) system of Spain has access to over 200 million documents—which is only a fraction of the data collected from the regional public services in the country so far (Ministerio de Sanidad, 2021). Another example can be found in scientific literature: the health science bibliographic databases IBECS and SciELO have indexed in recent years more than 200,000 [1] and 100,000 [2] publications in Spanish respectively.

But an abundance of data does not guarantee their actual use. Manual exploitation of such large collections of data is limited in nature. Further, health records and scientific publications consist to a large extent of natural language, which regular information systems cannot exploit nearly as readily as they do structured sources of data. Thus, advanced mechanisms must be put in place to automatically transform natural language into relevant, structured information. In the field of Natural Language Processing (NLP), this task is known as Information Extraction (IE).

NLP researchers have endeavoured to make the most of health-related content for decades. Progress in the field, however, is often hindered by critical ethical-legal barriers, a rigid ecosystem and exacting performance requirements. Nonetheless, it is a high-stakes domain that presents compelling scientific challenges stemming from the complexity of the concepts involved and the idiosyncrasies of clinical language. Despite these challenges, clinical NLP has recently experienced an upsurge in scientific contributions and results. Among the main reasons for this state of affairs are the impressive advances in Artificial Intelligence (AI), in particular the rise of modern Deep Learning (DL) approaches as applied to NLP. Notably, important developments have been made recently for languages other than English, which has traditionally been the main language of study in this field (Névéol et al., 2018a; Wu et al., 2019). Combined with other emerging technologies (e.g., Big Data, blockchain), these advances have boosted the pursuit of public policies worldwide aimed at the digital transformation of healthcare, such as the Global Strategy of Digital Health of the World Health Organization (2021).

Our work takes place within this growing field of clinical NLP research, as we address the following three main topics:

1. **Sensitive data detection and categorisation**: In layman's terms, sensitive data is data that can be used to identify individuals. This type of data is rigorously protected by laws and regulations aimed at safeguarding people's right to privacy. This is a major roadblock in clinical NLP research, because most of the documents generated during healthcare practice contain sensitive data.

2. **Term identification**: Clinical term identification is the NLP task by which mentions of clinically relevant terms (e.g., medications, symptoms, habits, body locations) are assigned an unambiguous meaning interpretable by computers through the linking of the terms to unique concept identifiers in a given knowledge base. Term identification can help extract knowledge from unstructured, underexploited sources of data. The applications of such solutions can be found, for instance, in clinical research, the healthcare practice, or healthcare management.

3. **Negation and uncertainty detection**: Clinical researchers and healthcare practitioners do not only report their positive findings and conclusions, but also the absence of observations and their hypotheses about what they do or do not observe. Thus, NLP solutions aimed at making sense of health-related texts must be able to handle these linguistic phenomena correctly.

The bulk of this dissertation tackles these topics in the Spanish language, which has received less attention so far in clinical NLP than English, despite being the 4$^{\text{th}}$ most spoken language in the world [3]. It is also at the moment the main language of use in the health system of the Basque Country (Perez de Viñaspre Garralda, 2017), where the work underlying this dissertation has taken place.

In what follows, we present our objectives and contributions in relation to each of the above-mentioned topics. Then, the chapter concludes with an outline of the remainder of the document.

1.2 Objectives

The ultimate objective of the dissertation is to participate in the advancement of the state of the art in the field of clinical NLP for the Spanish language through the creation of new resources (datasets, models and/or systems) and detailed comparative evaluations of IE solutions. This broad objective materialises as a set of specific goals oriented towards the dissertation's topics introduced above.

In this context, the first goal has been to conduct an **exhaustive review of the state of the art** in clinical IE for the Spanish language, with particular attention to the above-mentioned main topics of the dissertation.

With respect to **sensitive data**, we are interested in studying its automatic detection and categorisation in health-related texts, as this is the first step in sanitising texts of these problematic pieces of information. The specific objectives pursued are the following:

- To study the question of sensitive data in health record texts in Spanish from a technical point of view, in order to better understand how to characterise and approach it as a target of detection and classification systems based on NLP techniques.

- To assess and compare supervised approaches in the task of sensitive data detection and categorisation in clinical text, and to identify the advantages and limits of the different methods.

In relation to the topic of **term identification**, our goals have been the following:

- To build a system capable of performing clinical term recognition and identification natively in the Spanish language, that does not require annotated data of any kind, and that may be easily configured to meet the requirements of diverse application scenarios.

- To compare said system to other approaches proposed in the literature, most of which rely on Machine Translation (MT) at some point in the processing pipeline in order to leverage existing solutions for the English language, and to identify the advantages and limits of the tested methods.

As for **negation and uncertainty**, we study the automation of their detection from multiple perspectives. The objectives are as follows:

- To study the phenomena of negation and uncertainty in health records in Spanish, in order to propose guidelines for their annotation and to better understand how to characterise and approach them as a target of detection and classification systems based on NLP techniques.

- To build a corpus of clinical texts in Spanish manually annotated with negation and uncertainty information following the above-mentioned annotation guidelines.
- To assess and compare supervised approaches in the task of negation and uncertainty detection in clinical text, and to identify the advantages and limits of the different methods.

1.3 Contributions

In line with the objectives stated in the previous section, this dissertation makes contributions to the research field in clinical IE for the Spanish language, towards three specific topics: sensitive data detection and classification, term identification, and negation and uncertainty detection.

In what follows, we summarise the key contributions. The first significant contribution of this work is the following:

1. An in-depth **review of the state of the art**, including a historical perspective, inventories of the most relevant resources, and collations of the recent related work.

With respect to the topic of **sensitive data detection and classification** (Part II), the main contributions are the following:

2. A quantitative and qualitative **description of a corpus** of Spanish health records manually annotated with sensitive data.
3. Conditional Random Field (CRF), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and Transformer sequence labelling **models** for the detection and classification of sensitive data, trained and tested on two different corpora—manually augmented clinical cases, and health records. Some of these models are available online [4], and are being used by the scientific community in their own research (e.g., Pérez-Díez et al., 2021).
4. **Error analysis** and zero-shot **experiments** that call attention to the importance of site-specific data in clinical NLP, despite the advances in transfer learning made by the Transformer architecture and the widespread availability of pre-trained Language Models (LM).

Regarding the work carried out on **clinical term identification** (Part III), we make two contributions:

5. A knowledge-based **system** for term identification in Spanish. The system is available online for research purposes through a web API [5]. It has been exploited in several studies (e.g., Zubillaga et al., 2022) and has been successfully transferred to the industry as part of an anatomical pathology case indexing and retrieval solution.

6. A **comparison** of the above-mentioned solution, which performs term identification natively in Spanish, with other knowledge-based approaches that leverage third-party tools built for the English language.

Finally, the key contributions made on the topic of **negation and uncertainty detection** (Part IV) are the following:

7. Comprehensive **annotation guidelines** for negation and uncertainty cues and scopes in Spanish clinical text. These guidelines build on previous work about negation cues and scopes, but include uncertainty for the first time.

8. A **corpus** of health record excerpts manually annotated following the above-mentioned policy, as well as its qualitative and quantitative description. The corpus is available online [6] and is being actively exploited by NLP researchers to conduct experiments and build new resources (e.g., Hartmann et al., 2021; Magnini et al., 2021a; Rojas et al., 2022).

9. **Experiments** on supervised *a)* **cue and scope detection** modelled as a sequence labelling problem, and *b)* **assertion classification** modelled as a document classification problem. We study the robustness of several Transformer-based models against decreasing amounts of training data and adversarial test examples, and perform a thorough error analysis.

1.4 Outline

This manuscript is organised in 5 parts, 13 chapters and 7 appendices. Below, we outline each chapter and appendix, and explain how they relate to each other. A visual guide is given in Figure 1.1.

Part I: INTRODUCTION

This part of the manuscript situates the work and provides the relevant background to the work described in Parts II, III, and IV.

Chapter 1: Introduction In this chapter, we have contextualised and justified the research topics explored in the next chapters. We have also

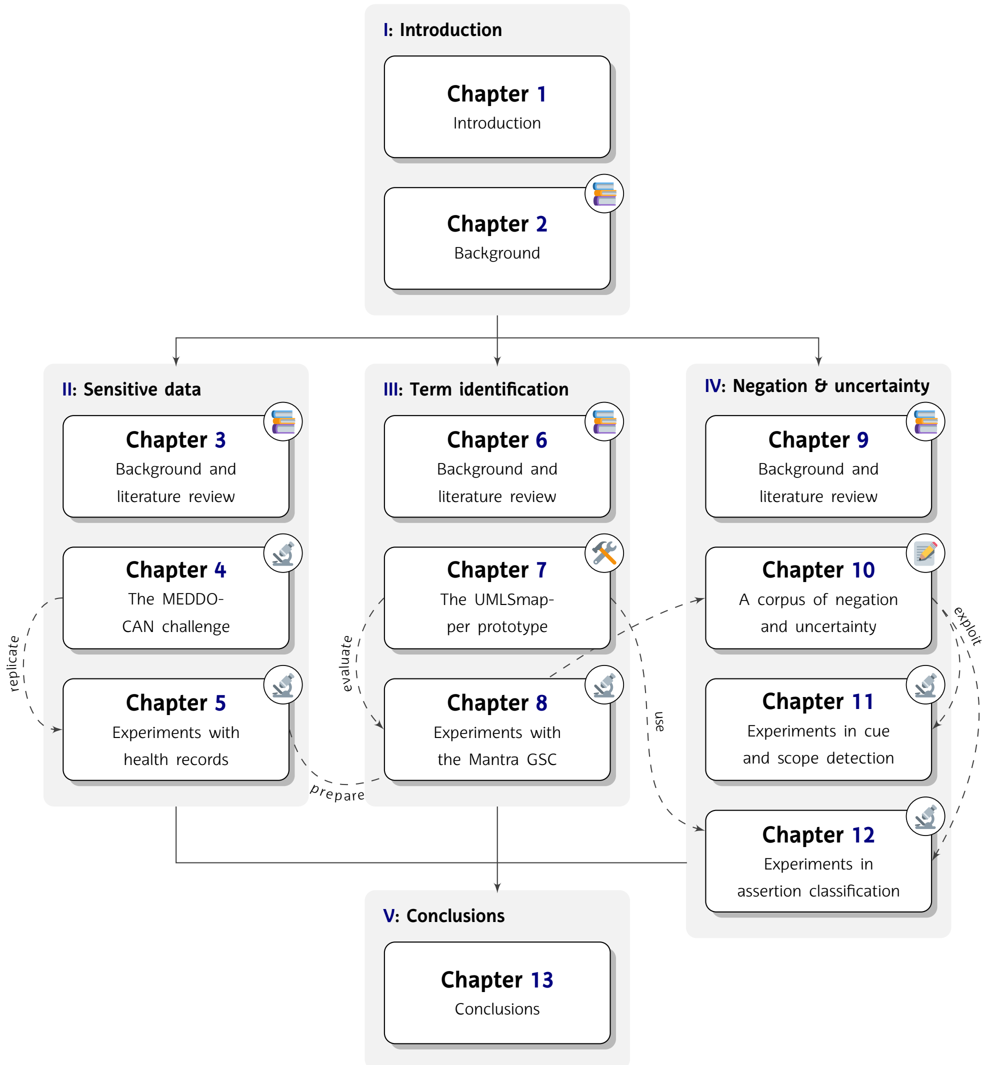


Figure 1.1: A schematic outline of this document's parts and chapters. Core chapters are marked for their main topic as describing related work (📚), experiments (🔬), systems (🛠), or corpora (📝). Relationships between chapters are labelled through dotted lines: The experiments about sensitive data detection of Chapter 4 are replicated in Chapter 5 on a different corpus; these, in turn, serve to prepare the corpus of negation and uncertainty presented in Chapter 10. This corpus is the basis for two sets of experiments, discussed in Chapters 11 and 12. Chapter 7 describes a system of term identification that is evaluated in Chapter 8 and used in the experiments of Chapter 12.

Part I: INTRODUCTION (continued)

summarised the main objectives and contributions of the dissertation.

Chapter 2: Background This chapter provides a general overview of the clinical NLP field (tasks, approaches, challenges), with special attention to IE for the Spanish language.

Part II: SENSITIVE DATA DETECTION AND CLASSIFICATION

This part deals with the topic of sensitive data in health-related texts, the problems they pose and how to address them through NLP.

Chapter 3: Background and literature review This chapter provides basic definitions, justifies the relevance of the topic, and presents the most pertinent resources and related work.

Chapter 4: The MEDDOCAN challenge Chapter 4 describes the work produced for the international challenge Medical Document Anonymization (MEDDOCAN) of 2019. The challenge consisted in detecting and classifying sensitive data in a synthetic collection of clinical case reports. To that end, we tested a variety of supervised NLP approaches. The chapter provides a description of the MEDDOCAN corpus, explains our approaches to the problem, and discusses the results.

Chapter 5: Experiments with health records Here, we replicate the experiments carried out in the previous chapter, but on a corpus of real health records instead of synthetic data. The chapter is concerned with the similarities and differences between the two corpora, the transferability of the various MEDDOCAN models, and how they perform in comparison to their analogous in-domain models. The corpus of health records used in this chapter is the same as that of Chapter 10, after sensitive data substitution.

Appendix A: MEDDOCAN category labels This brief appendix maps the names of sensitive data categories used throughout Part II to the official names used by the MEDDOCAN organisers in the challenge data and related publications.

Appendix B: MEDDOCAN confusion matrices Here, we report the confu-

Part II: SENSITIVE DATA DETECTION AND CLASSIFICATION (continued)

sion matrices of the experiments in Chapter 4 that weren't considered of primary relevance to be discussed in the body of said chapter.

Appendix C: NUBes: medical specialities and EHR sections Appendix C provides further quantitative description of the corpus of health records used in Chapter 5. The section of this appendix relevant to Part II centres on the distribution of sensitive data in the corpus over medical specialities and EHR sections.

Appendix D: NUBes-PHI confusion matrices This appendix contains the confusion matrices of the experiments in Chapter 5 that weren't considered of primary relevance to be discussed in the body of said chapter.

Part III: TERM IDENTIFICATION

This part addresses the problem of biomedical term identification with large terminology sources and knowledge bases.

Chapter 6: Background and literature review The chapter provides basic definitions, justifies the relevance of the topic, and presents the most pertinent resources and related work.

Chapter 7: The UMLSmapper prototype Chapter 7 describes a software, UMLSmapper, that performs term identification in Spanish by exploiting the terminology sources of the Unified Medical Language System (UMLS) Metathesaurus. The systems is described module by module, in terms of the expected inputs, internal processes, and generated outputs, with an example illustrating each step from start to finish.

Chapter 8: Experiments with the Mantra GSC In this chapter, we evaluate UMLSmapper on a public corpus of texts annotated with UMLS Metathesaurus identifiers. Its performance is compared to two other systems. As a simple baseline, we use a well-known, robust system for term identification in English, which we adapt to work on the Spanish language. The other system leverages MT to be able to apply English-oriented tools directly, and then project the annotations automatically back to the original text in Spanish.

Part IV: NEGATION AND UNCERTAINTY DETECTION

The fourth part of the thesis explores the topic of negation and uncertainty in clinical texts.

Chapter 9: Background and literature review This chapter provides basic definitions, justifies the relevance of the topic, and presents the most pertinent resources and related work.

Chapter 10: NUBEs: A clinical corpus of negation and uncertainty This chapter describes a new corpus of health records, NUBEs, annotated manually with negation and uncertainty markers and their scopes. The chapter thoroughly explains and discusses the annotation guidelines, the annotation process, as well as the final resulting public corpus.

Chapter 11: Experiments in cue and scope detection In this chapter we exploit NUBEs in a series of experiments about negation and uncertainty cue and scope detection, framed as sequence labelling problem. The experiments compare state-of-the-art neural techniques in several settings that include decreasing amounts of training data and adversarial test examples.

Chapter 12: Experiments in assertion classification This chapter replicates the experimental setup of the previous one, but for a different task: the classification of medical entities into the categories "absent", "possible", or "present". The chapter explains how the NUBEs corpus was transformed with UMLSmapper (Chapter 7) to serve this purpose, describes the experimental framework, and discusses the results.

Appendix C: NUBes: medical specialities and EHR sections Appendix C provides further quantitative description of the corpus of health records. The section of this appendix relevant to Part IV centres on the distribution of negation and uncertainty markers in the corpus over medical specialities and EHR sections.

Appendix E: Transformers vocabulary overlap with NUBes In this appendix, we quantify the overlap between the vocabulary of the NUBEs corpus and the vocabulary of the models trained and tested in Chapters 11 and 12.

Appendix F: Hyperparameters for negation and uncertainty detection This appendix lists the hyperparameters of the various models trained

Part IV: NEGATION AND UNCERTAINTY DETECTION (continued)

and tested in Chapters 11 and 12.

Appendix G: Additional metrics for negation and uncertainty detection In this appendix, we include the results of Chapter 11 and 12 using different metrics, to allow for direct comparisons between other published systems.

Part V: CONCLUSIONS

Chapter 13: Conclusions This chapter summarises the main results and conclusions of this dissertations, and indicates possible lines of research for future work.

Chapter 2

Background

2.1 Introduction

The objective of this chapter is to provide the theoretical foundations upon which the work described in the following chapters is built. It delves on basic questions about the three central concepts of the thesis: information extraction (*What is it? How does it relate to the rest of the Natural Language Processing (NLP) field? How is it done?*) for biomedical text (*What is it for? Why is it difficult?*) written in Spanish (*What have researchers achieved for this language up to this point?*).

The chapter is structured as follows: Section 2.2 introduces the types of tasks an applications the biomedical NLP is concerned with; Section 2.3 explains the main methods and approaches used within the field, from the rule-based to deep learning; Section 2.4 discusses some of the challenges that NLP researchers face when working on the biomedical domain; finally, Section 2.5 provides a brief overview of the work carried out by the community of biomedical NLP researchers for the Spanish language.

2.2 Tasks and applications

Biomedical NLP is a remarkably diverse research field where linguists, computer science and life science experts, bioinformaticians, and health care practitioners converge to build solutions whose common denominator is the need to process natural language related to the biomedical domain. But even that is not saying much: the natural language to be processed may consist, for instance, of medical reports, scientific literature, or social media content; the solutions may be aimed at healthcare service administrators, managers or consumers, clinicians, biomedical researchers, or NLP engineers. This section provides a brief overview of the many topics addressed within the field, both from the perspective of end-user applications and of NLP tasks.

2.2.1 Extracting versus modelling

Sager (1980) noted, on reviewing the collection of articles presented in the international conference on *Computational Linguistics in Medicine* (Schneider et al., 1977), that two major directions of research could be seen. On the one hand, there was the stream of research concerned with knowledge representation and reasoning (i.e., modelling), in which the need to draw upon natural language was overlooked or taken for granted. On the other hand, there was the body of research devoted to analysing medical natural language and representing it in semantically motivated structures (i.e., extracting). While research that fits into either of these categories is still relevant today, the field has certainly evolved, as correctly conjectured by Sager: "[t]hough at this time the two areas of research are still quite distinct, a common ground may develop in the future when the AI projects look deeper into their data sources, and the data processors seek more powerful systems for representing information". The strict separation between extracting and modelling has indeed weakened:

On the one hand, the advances of the NLP community have made it possible to model clinically relevant problems, such as disease prediction or risk analysis (to name only a few), by drawing directly on medical free text. On the other hand, Information Extraction (IE) has evolved to become the most popular task within clinical NLP (Wu et al., 2019; Percha, 2021). The aim of IE is to convert text into a set of human-interpretable structured features that serve to support a wide range of downstream tasks. For instance, they might be used to build advanced search indexing systems, to discover and quantify information unaccounted for in structured forms and databases, or, more frequently, they may be exploited alongside structured data sources (e.g., patient's biosignals or lab results) to answer clinically relevant questions.

This thesis makes contributions to three specific IE problems, namely, sensitive data detection (Part II), term identification (Part III), and negation and speculation detection (Part IV). These are not, as such, end-user applications nor do they attempt to respond directly to clinically motivated questions, but fall into the category of IE for feature engineering or for building modular solutions.

2.2.2 Healthcare versus biochemistry

NLP in the biomedical domain has two, clearly distinct main application domains. The first aims at providing support to healthcare professionals and patients, typically by mining medical notes and reports. This stream of research, pioneered by Sager (1972, 1978), is generally interested in patient information such as disorders, findings and treatments. With the advent of Internet forums and, more recently, social media, user-generated content too is now regarded as a valuable source of information for health-related purposes (J. Wang et al., 2020).

The second application domain started as attempts to mine information, such as names of genes and proteins (Fukuda et al., 1998), from journal articles in the biomolecular domain. Its general aim is to assist biochemistry researchers in accessing information buried in the scientific literature (e.g., about gene expression). See Piccialli et al. (2021) for a detailed survey of recent approaches in fine-grained biomedical application domains.

This thesis explores problems related to the processing of text produced in the context of healthcare practice: most of the work presented—Chapter 5 and all of Part IV—exploits a corpus of medical notes; Chapter 4 uses a collection of clinical cases; and, Chapter 8 exploits (in the absence of a better alternative at the time) a corpus of drug labels and article extracts annotated for mentions of diseases, procedures, body locations, etc.

2.2.3 A brief taxonomy of clinical NLP tasks

Text processing tasks in the healthcare domain can be divided into the following main categories, according to their end goal:

Low-level tasks are concerned with the pre-processing and basic linguistic analysis of text. This group of tasks includes, for instance, tokenisation, spell-checking, part-of-speech tagging and syntactic parsing. These tools are usually not the end goal of clinical NLP but serve as components to more complex applications. It should be noted that, with the advent of neural modelling techniques, some of these low-level tasks, which have traditionally been central for feature extraction and linguistic analysis, have been gradually rendered superfluous by end-to-end approaches (see Section 2.3).

IE tasks, as previously described, can be viewed as targeted skimming of texts. This includes a vast range of subtasks, such as text classification (e.g., medical note segmentation), Medical Entity Recognition (MER) and Medical Entity Recognition and Classification (MERC), relation extraction (e.g., adverse drug reaction [ADR] and timeline extraction), or term identification with standard medical terminologies, to name just a few. The resulting tools may be used in turn to build end-user applications such as anonymisation, clinical coding or advanced indexing suites. They can also be used for feature extraction to model clinically motivated problems. This group of tasks currently encompasses most of the effort in clinical NLP research and development.

Higher-level tasks in clinical NLP are oriented towards end-user (i.e., clinician or patient) applications. They can be further divided into two task subgroups: tasks involving text generation on the one hand (mainly, Machine Translation [MT], summarisation and simplification), and Information Retrieval (IR)/Question Answering (QA) on the other. The goal of most of these applications is to improve information accessibility and patient empowerment. For instance, these applications can facilitate finding case studies and health records that are relevant to a specific research subject or the care process of a particular patient. QA and simplification are mainly targeted towards patient-centred applications, by helping them better understand their own health records.

2.2.4 Clinical NLP shared tasks and challenges

The types of tasks and applications tackled by the clinical NLP community are perhaps better illustrated by the workshops, shared tasks and challenges organised in the field. Figure 2.1 shows a timeline of the most salient challenge series up to the year 2021, which we overview below.

The first challenge that involved NLP and clinical narrative took place in 2006 and was organised by Informatics for Integrating Biology and the Bedside (i2b2). There were two tasks in the challenge: one consisted in anonymising or de-identifying the unstructured content in Electronic Health Records (EHR) (Uzuner et al., 2007); the second consisted in classifying patients as smokers or non-smokers based on their health records (Uzuner et al., 2008). Since 2006, i2b2 (later National NLP Clinical Challanges [n2c2]) has organised 9 more challenges along the lines of IE. Some of the tasks include classifying patients as obese (Uzuner, 2009) or as having a high risk of suffering a heart failure (Uzuner et al., 2015), and coreference resolution (Uzuner et al., 2012).

In 2011, Text REtrieval Conference (TREC) organised its first challenge of IR for healthcare, after various others focused on the biomolecular domain. The challenge was aimed at exploring techniques for finding a population or cohort over which comparative effectiveness studies can be done by means of content-based access to the free-text fields of electronic medical records [7]. The challenge was repeated in 2012 (Voorhees et al., 2012). During years 2014 through 2020, TREC has encouraged research on IR for clinical decision support (CDS) (Simpson et al., 2014; Roberts et al., 2015, 2016) and precision medicine (PM) (Roberts et al., 2017, 2018, 2019, 2020). The latest TREC editions have focused on health misinformation (Clarke et al., 2020, 2021) and clinical trial retrieval [8].

The third major series of clinical NLP challenges is the Cross-Lingual Evaluation Forum (CLEF) eHealth Lab series. The first workshop took place in 2013, with challenges about identifying or normalising disease terms with the Unified Medical Language System (UMLS) Metathesaurus in English clinical texts (Pradhan et al., 2013), disambiguating acronyms and abbreviations (Mowery et

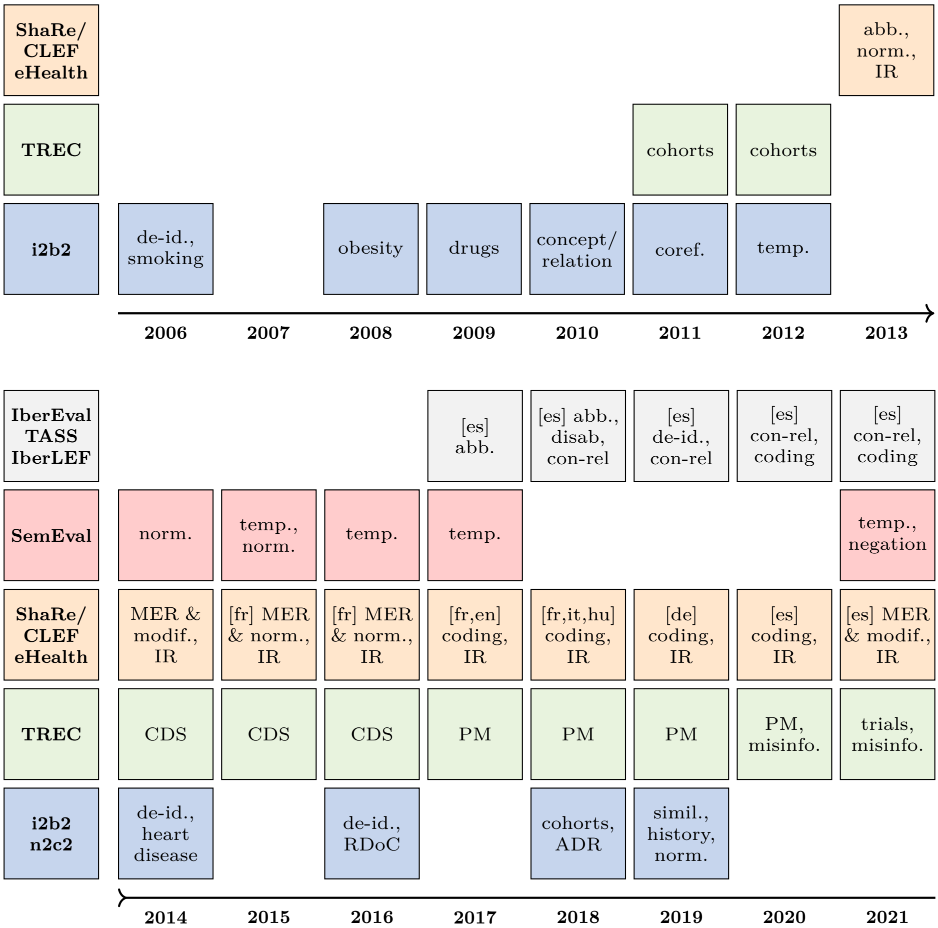


Figure 2.1: Selected clinical NLP challenges in chronological order up to the year 2021. The tasks were centred on English, unless otherwise specified between square brackets.

al., 2013), and retrieval of web pages based on patient's questions about their clinical reports (Goeuriot et al., 2013). Subsequent editions have continued with user-centred health IR tracks and, more interestingly, have introduced IE tasks in languages other than English, such as ICD coding in French, Hungarian and Italian (Névéol et al., 2018b), German (Neves et al., 2019) and Spanish (Miranda-Escalada et al., 2020b).

Starting in 2014, the International Workshop on Semantic Evaluation (SemEval) has proposed challenges along two lines: disease normalisation with the UMLS (Pradhan et al., 2014; Elhadad et al., 2015), following the CLEF eHealth 2013 task about the same problem; and, the extraction of temporal relations (Bethard et al., 2015, 2016, 2017), that is, ordering in a timeline the relevant events mentioned in clinical records. After a hiatus of 3 years, clinical-related tasks were brought with a challenge on source-free domain adaption (Laparra et al., 2021b) focused on assertion classification and temporal expressions.

Since 2017, multiple shared tasks have been proposed about clinical NLP for the Spanish language, organised within the Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval) and Taller de Análisis Semántico (TASS), later merged into the Iberian Languages Evaluation Forum (IberLEF). We review them in Section 2.5: Clinical NLP for the Spanish language.

2.3 Approaches and methods

As with general-domain NLP, clinical NLP approaches fall into two broad categories: rules and Machine Learning (ML). Within the latter, we should further distinguish between traditional ML and neural ML or Deep Learning (DL).

One of the most notorious differences between clinical NLP and general-domain NLP is that clinical NLP is known to have lagged behind its adoption of ML methods, maintaining a strong focus on rules (Connolly et al., 2016; Percha, 2021). This is not only true in industrial settings, but in academia as well: according to the literature review by Y. Wang et al. (2018) spanning over the years 2009 to 2016, 65% of the surveyed works were rule-based and the remaining 35% were based on statistical ML. Connolly et al. (2016) conjecture that the availability of high-quality knowledge bases and terminological resources may have held funding agencies back from recognising the value of building corpora, the most basic requirement of ML-based NLP.

Nonetheless, the landscape is rapidly changing, with an increased embracement of the neural ML paradigm. Publications that feature DL have more than doubled each year since 2016 (see Figure 2.2). According to Wu et al. (2019), the earliest adopters of DL were in the NLP community, but the medical informatics community was the most prolific during the surveyed period.

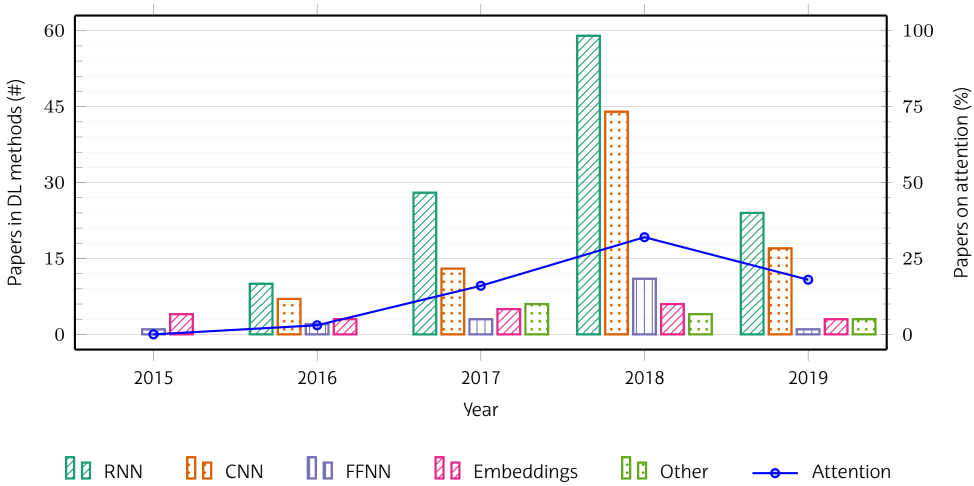


Figure 2.2: Growth of broad architectures in DL for clinical NLP over the years (adaptation of Figure 2 in Wu et al. [2019, page 460]). Percentages are relative to the number of studies published in that year. Data collected until April 2019. Not plotted: 1 FFNN paper in 2003, 1 FFNN paper in 2011, and 2 FFNN papers in 2014.

Currently, while the NLP community has already shifted its attention towards new research topics within the DL framework (P. Liu et al., 2021; Sun et al., 2022), the clinical NLP community is starting to look into how to best leverage the prominent DL approaches and what their shortcomings might be in the context of such a particular domain. For instance, there is a real concern about how to obtain models that generalize well—for which large amounts of harmonized data are required—while maintaining a notion of population variability—which requires that site-specific data is kept separate (Laparra et al., 2021a; Doyen et al., 2022). This and other challenges of clinical NLP are the topic of Section 2.4.

2.3.1 Rule-based approaches

Rule-based NLP systems consist of explicit implementations of hand-crafted rules guided by expert knowledge, experience and intuition. A rule-based system for IE typically involves keyphrase extraction via dictionary lookup or pattern matching, after which morphosyntactic information such as Part of Speech (PoS) tags and dependency trees is used to make decisions about said keyphrases or the document as a whole—classifying them, establishing relations between them, and so on.

For instance, Almeida et al. (2020) implemented such a system capable of extracting family history information from clinical notes. The rule-based system of Chen et al. (2019) for cohort selection ranked fourth among the participants of the n2c2 2018 shared task (Stubbs et al., 2019). MacRae et al. (2015) describe an expert system that detects influenza-like illness presentation from clinical notes.

Typically, systems like these rely on a combination of third-party resources. The UMLS Metathesausurus (Lindberg et al., 1993), the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) and the International Classification of Diseases (ICD), for instance, are commonplace among systems reliant on large knowledge bases and lexicons. The multi-purpose analysis frameworks clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al., 2010) and MetaMap (Aronson, 2001) are also recurrently featured, as is NegEx (Chapman et al., 2001)—yet another rule-based tool that performs assertion classification. All of these will be mentioned again in subsequent chapters.

While rule-based methods tend to demonstrate an acceptable performance in terms of precision, their well-known lack of generalisation capability can be a major drawback in certain tasks where recall is also sought after. For that reason, it is common to find in the literature proposals of hybrid approaches that combine heuristics and traditional or neural ML. The works by Casillas et al. (2016), Chen et al. (2020), Jouffroy et al. (2021), Suárez-Paniagua et al. (2021) and Fu et al. (2022) are just a few examples.

2.3.2 Traditional ML approaches

ML is concerned with algorithms that allow computers to learn to solve tasks by example, without having to be explicitly programmed. We refer as *traditional* ML, also called *statistical* or *shallow* ML, to the approaches not based on neural networks, which we look into in the next section.

Supervised ML algorithms learn a function or model to map inputs into outputs, that is, they require labelled data. The inferred models are then able to assign labels to data unseen during training. Unsupervised ML algorithms, on the other hand, attempt to discover patterns in unlabelled data to create clusters or detect outliers, for example, that must then be interpreted by humans.

One key aspect to obtaining a good traditional ML model, supervised or unsupervised, is being able to characterise the data with appropriate descriptive predictors or features. The study of the suitability of feature combinations for a given corpus, learning objective and learning algorithm is known as feature engineering. See, for instance, the work by Weegar et al. (2016), who study the impact of simple features (e.g., prefixes and PoS tags) in the task of MER, or Santiso et al. (2019), who assess the performance of features derived from word embeddings (see Section 2.3.3.3) in the detection of negated clinical entities.

Researchers frequently resort to the publicly available terminological resources and NLP suites mentioned in the previous section to do feature extraction too.

The other key aspect is having access to sufficient quality data or having the means to curate it oneself—and annotate it, if supervised algorithms are to be applied. That is, expert input still plays a critical role where ML is concerned. Expert knowledge and experience is not only crucial when defining the problem and validating the results, but it provides a sound foundation over which to conceive relevant features and to design and implement quality annotation policies.

There exist an immeasurable amount of traditional ML algorithms. Among the supervised, which are the most frequent in the field as well as most relevant to this thesis, we must highlight the following:

Support Vector Machines (SVM) (Cortes et al., 1995) are a family of algorithms that aim at finding the hyperplane that best separates the feature space into two groups. SVMs are often the preferred choice among researchers due to their training efficiency and suitability for small-to-medium-sized datasets. For example, Tang et al. (2012) used Structural SVMs (Tsochantaridis et al., 2005) to resolve the MER track of the i2b2 2010 challenge; Casillas et al. (2016) and X. Yang et al. (2019) test SVMs in the task of ADR relation extraction.

Naïve Bayes is another popular family of classification algorithms, in spite of their simplicity. Naïve Bayes classifiers are based on Bayes' theorem with the assumption that features are independent given the class label. Among the many works that test them, we might mention the following: Spasić et al. (2012) fit a Naïve Bayes classifier to categorise sentences in suicide notes into 15 sentiment categories; Prakash G. et al. (2014) use Naïve Bayes to detect mentions of diseases and treatments in scientific article abstracts; J. Zhao et al. (2015) compare Naïve Bayes to other traditional algorithms (namely, Decision Trees, Random Forest, SVMs and logistic regression) in the task of predicting the presence or absence of ADR event mentions.

Conditional Random Fields (CRF) (Lafferty et al., 2001) are the preferred approach for problems that can be shaped as sequence labelling tasks, as they are able to leverage context information. For example, Li et al. (2015) detect medication names and attributes from clinical notes using CRFs; Ju et al. (2015) use CRFs to semi-automatically compile a lexicon of symptoms from Chinese data; Lopes et al. (2019) train a CRF classifier to do MER in Portuguese text.

Traditional ML algorithms like these are efficient provided an optimal feature space is computed for the task at hand. However, feature extraction and engineering is a time-consuming, complex endeavour that depends on quality tools and resources adapted to the domain and language of interest.

2.3.3 Neural ML approaches

Neural ML comprises the subset of ML approaches that are based on Artificial Neural Networks (ANN). Most ANNs are organised as chained layers of artificial neurons or Perceptrons: an input layer; optionally, intermediate layers, also known as hidden layers; and, an output layer. Training an ANN implies fitting the weights of the connections between the neurons, usually through backpropagation (Rumelhart et al., 1986). The more hidden layers an ANN has the *deeper* it is said to be, hence the terms Deep Neural Network (DNN) and Deep Learning (DL). See Goodfellow et al. (2016) for further references on the topic.

DL marked a milestone in the mid 2010s for NLP, disrupting the entire field within a few years' time. Not only did researchers manage to obtain better and better results, but DL also pushed feature engineering to the background, as ANNs are able to learn feature representations through their internal structure. Much of the research in DL has indeed focused on exploring ANNs architecture variants to obtain better internal representations for different tasks and input types. In subsequent sections we will provide an overview of the most recent, salient architectures used in NLP.

2.3.3.1 Transfer learning

On the downside, training DL models requires infamously more data and computing resources than traditional ML algorithms do. A significant amount of the current research is dedicated to DL optimisation on these grounds. Besides architecture optimisation, transfer learning has been the major driving force in making DL viable without large, labelled corpora or prohibitive hardware and training times. This is achieved through the pre-train/fine-tune approach.

The pre-training step trains a model in a task for which copious amounts of data exist and that allows the model to acquire general knowledge that might be useful to solve many other different problems. In NLP, that task is usually Language Modelling or an approximation of it. Then, the representations learned by the resulting model can be used as the starting point to train a new model on a different task, language, or domain where less data are available. This is called fine-tuning, and its specific implementation depends on the learning technique or type of model being transferred. The key is that a model need only be pre-trained once to be repurposed in other languages, tasks, and domains.

While transfer learning had been studied long before the surge of DL (Pan et al., 2010), its implementation with traditional ML algorithms raised notable difficulties in terms of feature transfer, among other issues. It is only recently that transfer learning has been used effectively in NLP, and in clinical NLP as well (Laparra et al., 2021a). What is more, it is now the standard approach, driven by the introduction of the Tranformer architecture (Vaswani et al., 2017) and models based on the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019).

2.3.3.2 Neural architectures

Among the simplest ANNs is the Feedforward Neural Network (FFNN) or Multilayer Perceptron (MLP), where each neuron of a layer is connected to all the neurons in the next layer and the information flows from the input exclusively forward to the output. DL researchers have proposed ANN variants built on top of FFNNs in an attempt to obtain better internal representation of their data and overcome practical shortcomings of ANNs. In what follows, we introduce briefly the three most important ANN architectures in the field of NLP and provide examples of how they have been used in clinical NLP.

2.3.3.2.1 Convolutional Neural Network (CNN) CNNs (LeCun et al., 1989) were initially conceived for computer vision. As Goodfellow et al. (2016, p. 321) put it, "[c]onvolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers". While specialised in image processing, CNNs can be employed to process any data type that can be thought of as having a grid-like structure. Starting with Collobert et al. (2011), Kim (2014) and dos Santos et al. (2014), this type of network has been widely used in NLP by treating text as a 1-D grid of characters or tokens, often in combination with traditional classifiers (e.g., CRFs) serving as output layers. In clinical NLP, CNNs have been explored, for example, to classify health-related encyclopaedic text into topics (Hughes et al., 2017), to extract relations between pre-annotated clinical concepts (Luo et al., 2017), to attempt automatic diagnosis from medical notes (Z. Yang et al., 2018), and to predict patient readmission risk from medical notes (Lu et al., 2021).

2.3.3.2.2 Recurrent Neural Network (RNN) Based on the work by Rumelhart et al. (1986), RNNs are specialised ANNs for modelling sequential data, the most successful implementations to date being the Long Short-Term Memory (LSTM) (Hochreiter et al., 1997) model and networks based on the Gated Recurrent Unit (GRU) (Cho et al., 2014). Unlike other types of ANNs, RNNs have feedback or recurrent connections that make the outputs of the network dependant on the prior elements of the sequence. This is pictured as the network having a sort of *memory* and being able to exploit historical information when processing a sequence, such as speech or natural language. Another desirable trait is that

RNNs can potentially process inputs of any length, as the size of the model does not increase with the size of the input. One of the weaknesses of RNNs where NLP is concerned, however, is that they are directed by definition, whereas natural language is not—the words at the end of a sentence may affect how words at the beginning should be interpreted. For that reason, it is usual to combine forward and backward RNNs into a bidirectional RNN (Schuster et al., 1997). As with CNNs, it is also common to top RNNs with a CRF classifier (Huang et al., 2015; Lample et al., 2016) when dealing with sequence labelling problems. RNNs have been used in clinical NLP to do, among others, concept extraction (Chalapathy et al., 2016), MER on EHR reports of cancer patients (Jagannatha et al., 2016a,b), heart failure onset risk prediction (Rasmy et al., 2018), and event extraction from medical reports written in Italian (Viani et al., 2019).

2.3.3.2.3 Transformer Proposed by Vaswani et al. (2017), the Transformer DNN architecture is to date the state of the art in virtually all NLP tasks. While designed to handle sequential data, Transformers are not recurrent networks, but process the entire input all at once. The ability to model relationships between input elements is given by the generalisation of the use of attention mechanisms and positional embeddings. The attention mechanism had been previously proposed for RNNs to be able to learn to attend to different hidden states at each decoding step, thus notably improving the modelling of long-range relations between sequence elements. In the Transformer architecture, attention is the pervasive mechanism throughout the network in the form of self-attention and cross-attention layers, combined with FFNN layers. The gains in performance and reduced training times, compared to RNNs in particular, have made this architecture the preferred choice of NLP researchers, triggering an outburst of publications of models pre-trained on different languages and domains. In clinical NLP, Transformer-based models have been successfully employed, for instance, to extract ADR events from tweets (Miftahutdinov et al., 2019), to extract concepts and relations in Spanish health-related text (García-Pablos et al., 2020, 2021), to extract angina symptoms from clinical notes (Eisman et al., 2020), and to detect actionable radiology reports in Japanese (Nakamura et al., 2021).

2.3.3.3 Text embedding representations

Text must be represented in terms of numbers in order to be able to operate with it mathematically. This is achieved by assigning unique vectors to meaningful language units (e.g., words, morphemes); that is, by *embedding* these units in a vector space. Ideally, these numeric representations should encode natural language in all its complexity through noticeable geometric relationship that somehow mirror the semantic relationships among the language units themselves.

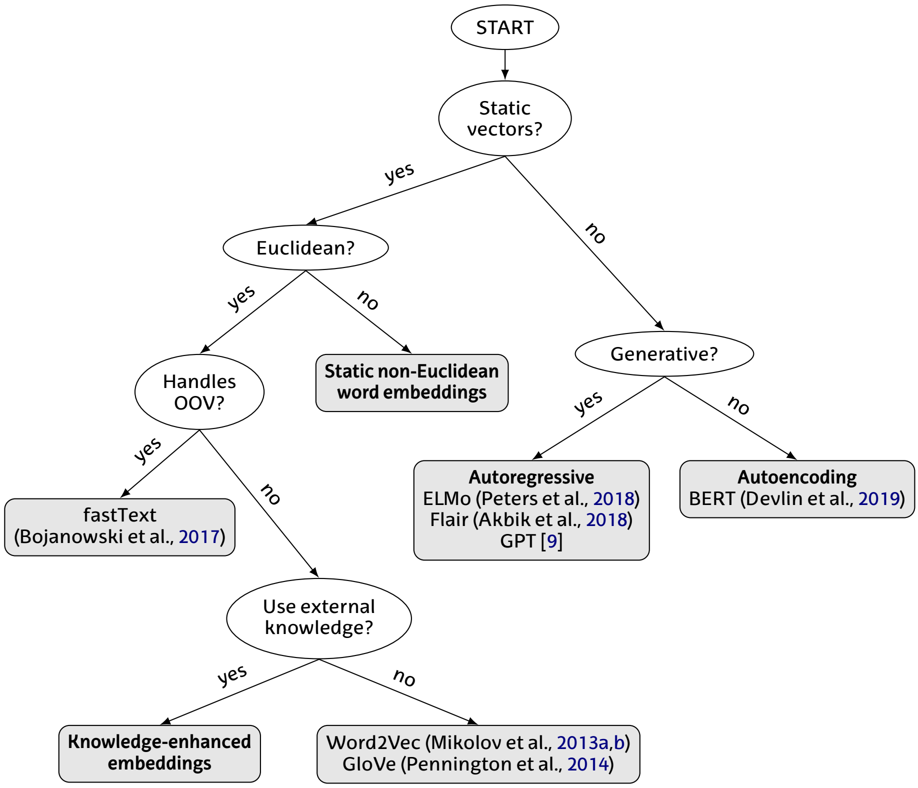


Figure 2.3: Relations between word embeddings based on some basic properties (adaptation of Figure 2 in Torregrossa et al. [2021, page 87])

Ever since neural word embeddings were proposed by Bengio et al. (2000), research on this topic has focused mainly on unsupervised representation learning, typically involving language modelling or co-occurrence matrices. All those approaches are based on the distributional hypothesis (Harris, 1954) that words that occur in the same contexts tend to have similar meanings.

Such neural embeddings are convenient for multiple reasons, which could be summarised as follows: *a)* they are learned from unlabelled corpora, *b)* they capture language and domain specific knowledge that can be transferred from one task to another, and *c)* they are easily passed as input to neural networks. Furthermore, they have been proven time and again to be effective, so much so that they are currently the standard practice in NLP.

In what follows, we introduce briefly the main types of neural word embeddings to date, some of which are used in subsequent chapters. Figure 2.3 (from

Torregrossa et al. [2021]) provides a visual guide of those types and how they relate to each other. Existing pre-trained embeddings for the Spanish language and/or the clinical domain will be overviewed later in this chapter (Section 2.5.3).

2.3.3.3.1 Word2vec Proposed by Mikolov et al. (2013a,b), Word2vec embeddings are word-level constant or static vector representations of words. That is, they represent words as unique vectors distilled from the words' contexts in the training corpus. The representations are learnt with a FFNN from a word prediction task: in the continuous bag-of-words (CBOW) architecture, the model attempts to predict the word from its surrounding context, while the skip-gram variant attempts to predict the context from a given word. There are two key hyperparameters to Word2vec, in addition to the architecture itself: the number of dimensions, and the size of the context window.

2.3.3.3.2 Global Vectors (GloVe) Proposed by Pennington et al. (2014), GloVe embeddings are also word-level static representations of words, although they are learnt from a co-occurrence matrix instead of a word prediction task. Specifically, the training objective of GloVe is to learn word vectors such that their dot product equals the logarithm of the words' probability of co-occurrence. Because GloVe embeddings are learnt from global word counts, they are better at capturing longer-term dependencies than Word2vec.

2.3.3.3.3 fastText Proposed by Bojanowski et al. (2017), fastText embeddings were conceived to address some of the shortcomings of methods such as Word2vec and GloVe, namely, *a)* that they cannot handle out-of-vocabulary words (OOV), i.e., words not encountered in the training corpus; and *b)* that each word vector is learnt separately, disregarding the fact that many words share morphological constituents (hence, meaning). The fastText approach is based on the Word2vec skip-gram model, where each word is represented by the sum of the embeddings for the character n-grams of the word. Thus, the embeddings are able to represent the morphology and lexical similarity of any word, regardless of its occurring or not in the training corpus.

2.3.3.3.4 Embeddings from Language Models (ELMo) Proposed by Peters et al. (2018), ELMo embeddings were one of the earliest successful contextualised word embeddings. Contextualised word embeddings are directly opposed to static embeddings in that words receive a different vector depending on the sentence they occur in. That is, the embeddings are not constant, but need to be computed for every given word in context. ELMo learns these contextualised representations by training a multilayer bidirectional LSTM (biLSTM) network on a word-level language modelling task. The word embeddings are obtained by combing the internal states of this network. Further, ELMo incorporates subword information through the use of character convolutions as input to the LSTMs, thus being sensitive to internal word structures and robust to OOV words.

2.3.3.3.5 Flair Proposed by Akbik et al. (2018), Flair embeddings are also character-based contextualised word embeddings learnt through a bidirectional RNNs. In this case, however, the RNN does not have an explicit notion of word boundaries as it is pre-trained directly on a character-level language model objective. The word representations are obtained by concatenating the hidden states of target word's last character in the forward RNN layer and of the first character in the backward RNN layer. As demonstrated by Flair authors, it is often beneficial to combine Flair embeddings with other word-level embeddings.

2.3.3.3.6 Generative Pre-trained Transformer (GPT) Proposed by Radford et al. [9; 10], GPT embeddings are contextualised word embeddings learnt by training a stack of decoder Transformer blocks on a language modelling task. As such, these word representations rely only on the leftmost context of each given word, contrary to all the aforementioned techniques, which are bidirectional. Still, the latest GPT release, GPT-3 (T. B. Brown et al., 2020), has been spectacularly successful. It has been proven to perform well on few-shot and even zero-shot scenarios, thanks to its massive size of 175 billion parameters and the sheer amount of data used to train it. The strategy followed by GPT to handle OOV and leverage subword structure is Byte-Pair Encoding (BPE) tokenisation (Gage, 1994; Sennrich et al., 2016).

2.3.3.3.7 Bidirectional Encoder Representations from Transformers (BERT) Proposed by Devlin et al. (2019), BERT is to date the other most successful contextualised word representation model. While based on the Transformer architecture too, it uses the encoder component, as its name suggests. In this sense, it is radically different to GPT, because it is not autoregressive and is able to encode left and right context simultaneously. It is pre-trained on two objectives: *a)* the Masked Language Model (MLM) task, where the model is trained to predict the tokens that are randomly masked in a sentence, and *b)* the Next Sentence Prediction (NSP) task, where the model is trained to predict whether one given sentence follows another. Further, BERT uses WordPiece (Schuster et al., 2012) to perform subword tokenisation. Closely following the breakthrough of BERT, many variants have been proposed, such as RoBERTa (Y. Liu et al., 2019) and ELECTRA (Clark et al., 2020) to name a few, that offer improvements over BERT in aspects like increased performance or reduced computational cost.

2.4 Challenges

NLP faces many challenges posed by natural language itself, the most fundamental being lexical variability and ambiguity. Lexical variability is given by synonymy, the semantic relation whereby multiple expressions (morphemes, words or phrases) convey the same meaning, as in the affixes '-algia' and 'odino-' in Example E1[1]. At the same time, natural language is ambiguous due to polysemy (E2) and homonymy (E3). The former describes the property of morphemes, words or phrases to convey different meanings depending on the context they appear in. Homonymy occurs when distinct words—that is, words of distinct historical origin and distinct sets of meanings—happen to be written and/or read the same way.

E1	**Synonymy:**	Ingresa por epigastr**algia**. Admitted due to epigastric **pain**.	Refiere **odino**fagia. [The patient] reports **painful** swallowing.
E2	**Polymsemy:**	Tío **ciego** por cataratas. Uncle **blind** due to cataracts.	Pólipo en **ciego** resecado. Resected polyp in **cecum**.
E3	**Homonyny:**	Bebedora habitual de **vino**. Regular **wine** consumption.	**Vino** de nuevo a Urgencias. [The patient] **came** to the ER again.

In 2014, Friedman et al. identified 9 additional challenges more or less specific to NLP for biomedicine and health (for instance, "patient privacy and ethical concerns", "good system performance", "misspellings and typographical errors", "reliance on medical knowledge and reasoning", "complexity of biological language"), all of which still apply today. With the adoption of DL in the field of health informatics, we face yet another challenge, namely, making the inner workings and results of neural networks explainable and transparent. In what follows, we elaborate on some of these challenges that we consider more germane and critical to clinical NLP.

2.4.1 Data privacy

Clinical NLP needs to handle data that typically includes personal, health and social history information of the stakeholders involved in the clinical practice, such as healthcare professionals, patients, relatives and caregivers. These are the most sensitive pieces of information conceivable. As such, they are protected by many guidelines and policies, from the international (e.g., the General Data Protection Regulation [GDPR] of the European Union) to the local (e.g., institutional ethics

[1] Throughout the document, translations of Spanish examples to English are given directly below the example. In these specific examples, we highlight in boldface the relevant pairs of expressions for each semantic relation that we want to illustrate.

committees), whose aim is to safeguard the privacy of individuals and which researchers and developers are expressly subject to.

As a consequence, datasets for clinical NLP are difficult to come by, and those that exist tend to be kept private. Clinical NLP can in fact be infamous for its frequently siloed research, whose reported results cannot be reproduced nor compared by the community. What is more, rigorously measuring the actual advancement of specific tasks is often unattainable.

Conveniently enough, NLP is part of the solution to this predicament. De-identification is the process of altering data by redacting or replacing sensitive information, after which the data may be safely shared. The fact that the automatisation of this process through NLP was the topic of the first ever shared task on clinical NLP (Uzuner et al., 2007) speaks for the importance of this research line that is still active due to the positive impact that sharing clinical data can have, not just upon NLP research but, ultimately, upon biomedical research.

The key step of a standard automatic de-identification pipeline, namely, sensitive data detection, is the topic of Part II of this thesis. Chapter 3 elaborates on theoretical aspects and the state of the art of said topic.

2.4.2 Non-standard language

Clinical text documents serve diverse purposes, differ in their content and level of detail. In general, they are aimed at other healthcare professionals or the authors themselves, so editing the texts to facilitate comprehension by a wide audience is not a main concern, as is the case of other text genres in the same domain, such as biomedical scientific publications. Most importantly, healthcare professionals typically have limited time devoted to the task of writing; as a consequence, they use a myriad of abbreviations and acronyms, while hardly ever caring for spelling correctly nor respecting the grammatical standards of their language. As J. Carnicero points out in Amézqueta Goñi et al. (2003), the situation has worsened since EHRs were implemented in health centres.

As a result, clinical narrative text is unlike general domain language, which makes its processing an extremely difficult and challenging problem for NLP researchers. Table 2.1 shows real examples in Spanish of these difficulties, which we explain briefly below (see Lima-López et al. [2021b] for a detailed breakdown of error types in Spanish medical notes):

- To begin with, practitioners are very flexible regarding formatting when writing their reports. The semantics conveyed by the same formatting varies from one context to another; it is even possible to express complex ideas without using whole sentences by means of specific formatting. Furthermore, punctuation rules are largely overlooked; the most common deviation from standard punctuation is actually not using punctuation marks at all.

Table 2.1: Illustrative examples of common challenges in processing text from clinical narratives (adapted from [Leaman et al., 2015])

Category	Detail	Example
Flexible formatting	Formatting semantics	**Section header:** “Intervención principal: REPARACION DE LUXACION FRECUENTE DE [...]” **Inseparable phrase:** “Abdomen: Blando y depresible”
	Structure without sentences	“T.A:160/106 mmhg. F.C:74x’. Tª:36’1º.” “Trazodona 100 mg, 0 - 0 - 0 - 1/2.” “Ph:7,46, PCO2:54, PO2:56, BE-B:12,3, HCO3:38,4, [...]”
	Missing punctuation	**Commas:** “No aumento tos ni expectoración ni náuseas ni vómitos ni dolor torácico.” **Periods:** “No se aprecian adenopatías En parénquimas pulmonares se aprecian áreas de condensación”
Atypical grammar	Missing expected words	**Verb:** “No [se aprecia] Hernia de Hiato” **Object:** “Coordinación remite [al paciente] por episodio de atragantamiento” **Articles:** “[Un] Paciente de 69 años que ingresa por [una] sensación de insuficiencia respiratoria.”
	Unusual PoS combinations	**Adjective without noun modified:** “Eupneica en reposo”
Rich descriptions	Variety of textual subjects	**Patient:** “Bien nutrida, hidratada y perfundida” **Anatomy:** “No I.Y. rítmica Mv conservado.” **Test or procedure:** “Estudio no valorable, mala trasmisión ecográfica” **Family:** “cinco familiares fallecidos de cardiopatía isquémia”
	Language specific to medical context	**Jargon:** “No palpo puntos dolorosos, masas ni megalias.” **Abbreviations:** “se instaura tto ATB empírico oral” **Acronyms:** “Adherencias de la IQ previas. A descartar foco infeccioso en LSD”
Misspellings		“Tambien presento en ingreso reciente ubn deterioro de la funcion renal” (sic) “refiere epigastralgia contínua, que no mejora con ninguna medida, de localización hacia hipocondro derecho. No diebre ” (sic) “No alteraciones vlavulares significativas. No datos de hipertension pulmonar.” (sic)

- Another characteristic of clinical narrative text is atypical grammar. The most striking feature related to grammar is the amount of non-standard ellipsis found in the texts, which infuses a telegraphic style to the texts. It is also common to find unusual part-of-speech tag combinations.

- Finally, clinical text is plagued with misspellings and typographical errors.

Despite the reductive grammar, however, descriptions contained in the texts are actually very rich. The same structures can be used to refer to a variety of textual subjects, such as a patient, a body part of a patient, a relative of a patient, a healthcare professional, a healthcare procedure, and so on. Furthermore, clinical narrative is rich because it is a product of a very specialised domain activity. As such, healthcare has an ever-evolving terminology, with new concepts and terms entering the language while obsolete ones fall out of use. This aspect of medical language will be discussed further in forthcoming sections.

2.4.3 Lack of interpretability and explainability

Despite the quantitatively superior results that technology based on DL has been proven to be able to achieve across the board in comparison to more traditional methods, the fact that they are perceived as “black boxes” stands in the way of their adoption by the healthcare sector in real practice (Cabitza et al., 2017; Ravì et al., 2017; Vellido, 2020; Doyen et al., 2022, among many others).

Admittedly, this challenge affects to a greater extent systems aimed at answering clinical problems directly, rather than IE systems, and while this thesis does not dive into any of these matters, the problem is important enough to dedicate a few lines to it.

The issue is actually part of the broader “Alchemy debate” within the Artificial Intelligence (AI) community (Church et al., 2021) [11; 12]: there exists a generalised concern, rekindled by the widespread success of DNNs, that researchers may be neglecting insight while seeking better and better results; to put it simply, that we know DL works, but not why. This is a particularly pressing matter where AI and the healthcare sector cross paths, given the ethical and legal concerns that arise from practitioners having to make actionable decisions by heeding the suggestions of programs whose behaviour is ill-understood or cannot be explained. It must be noted further that clinicians may be held responsible if they follow AI recommendations that conflict with the standard of care and that turn out to be detrimental for the patient’s health (Price et al., 2019).

The development of new methods aimed at explaining the decision-making process of DNNs has prospered into an active research field known as Explainable AI (XAI). The proposed methods include visualisation, distillation and the development of intrinsically explainable networks (see Ras et al., 2022, and references therein).

2.4.4 Reliance on expert knowledge

ML in clinical NLP is not always feasible or able to deliver on its own the expected performance, be it because there is no available data suitable for the task at hand or because the data available is not enough to learn by example. Among the many factors that contribute to this situation—some of which we introduced above—is evidently the highly specialised and dynamic nature of the domain, which may rapidly render existing datasets obsolete, inadequate or insufficient. We illustrate this point by drawing on two recent events:

- The coronavirus disease 2019 (COVID-19) pandemic has resulted in the creation of new vocabulary—of which 'COVID-19' is an obvious example, as are the names of the new vaccines, e.g., 'Comirnaty' or 'Vidprevtyn'—, while some existing expressions have acquired new senses; for instance, the names of the companies that produce the vaccines are often used to refer to the vaccines themselves by semantic broadening. Evidently, these changes in vocabulary are not reflected in datasets curated prior to the pandemic.

- The 11$^{\text{th}}$ revision of the ICD is in effect since January 2022 [13] and will gradually be implemented across World Health Organization (WHO) member states. The existent collections of episodes coded with the prior ICD revision (namely, ICD-10) will then be rendered, vast as they may be, useless, as is, to develop new coding systems.

In this regard, it must be noted that clinical NLP has benefited greatly from techniques like data augmentation, domain adaptation and transfer learning, as a means to circumvent data scarcity issues. Often, however, pure ML black box systems are simply not desirable, as explained in the previous section.

For all these reasons, clinical NLP tends to favour hybrid architectures, that is, solutions that combine ML with the exploitation of standard terminologies, ontologies and/or hand-crafted rules that encode expert knowledge. While less popular in NLP academia, the reality is that these old-fashioned approaches can offer some advantages in certain contexts, in spite of their many and well-known drawbacks (e.g., a sheer lack of generalisation). For instance:

- Knowledge-driven systems are better equipped to face extreme multi-label classification or tagging problems. Collecting a balanced corpus where every existing target category is represented is often simply impracticable (e.g., SNOMED CT codes currently amount to more than 350k).

- Knowledge-based resources and rules may be more easily mended to reflect changes in the state of events, than generating hopefully enough new examples to train or adapt ML models and expecting for said models to passively capture the desired changes.

- Knowledge-based resources and rules may encode implicit knowledge more handily. Friedman et al. (2014, p. 278) gives the example of "inferring that a patient is depressed based on the fact that an anti-depressant is prescribed (even though there is no explicit mention of depression in a note)".

- Perhaps more importantly today, knowledge-based solutions are interpretable, and their successes or failures are easily explained. In this sense, they are better at inspiring confidence in the end users.

Still, whichever approach is chosen is bound to depend on task-specific expertise, either to annotate data or to design and maintain vocabularies and heuristics (or both). As Spasic et al. (2020, page 2) put it, "[m]uch like the law of energy conservation, it seems that the knowledge required to inform the creation of an accurate computational model is simply transferred from one form to another. Instead of explicit knowledge in the form of rules, machine learning is based on implicit knowledge in the form of annotations and their distribution, with the time involved in their acquisition remaining virtually constant". The challenge is then to decide to what extent to lean towards one strategy or the other, factoring in issues like the level of expertise required and its cost, the suitability of existing resources and the effort necessary to adapt them, or the requirements for generalisation, among many other considerations.

2.5 Clinical NLP for the Spanish language

Having introduced the research field of clinical NLP in the previous sections, we next bring briefly into focus several topics related to the Spanish language in clinical NLP, before addressing the three main tasks of this thesis, namely, sensitive data detection, term normalisation, and negation and speculation detection.

2.5.1 Brief historical overview

Despite the long-standing tradition of clinical NLP whose earliest works can be traced as far back as the 1960s decade (i.a., Pratt, 1973; Schneider et al., 1977), it might come as a surprise, given the dominant position of Spanish among the world's major languages [3], that research overtly and specifically devoted to the processing of clinical text written in this language is actually in its teen years. This is reflected in the number of published articles, where research targeted at the processing of text written in English is still dominant, also in biomedical NLP (Névéol et al., 2018a; Wu et al., 2019); in fact, works focused on languages other than English are a minority even when lumped together (see, for example, the data gathered by Wu et al. [2019] in Table 2.2).

Table 2.2: The languages for labelled corpora used among the included articles in the systematic literature review of Wu et al. (2019) on DL in clinical NLP

Language	#	%	Language	#	%
English	151	72.1	Italian	2	0.9
Chinese	42	19.8	Dutch	1	0.5
Spanish	5	2.4	Thai	1	0.5
Japanese	4	1.9	German	1	0.5
Finnish	4	1.9	Swedish	1	0.5
French	2	0.9	Not reported	2	0.9

Health-related Spanish NLP has nevertheless grown rapidly in parallel with the steady digitalisation of the healthcare sector worldwide, all the while keeping up with the latest developments of the rest of the NLP community.

Some of the earliest studies worked on the morphosyntactic and semantic analyses of medical-related texts (Crespo Miguel et al., 2008; Iglesias et al., 2008; Castro et al., 2010). Next came the first exploratory works about automatic ICD coding (Casillas et al., 2012; A. Pérez et al., 2014), and IE focused on MER and ADR mining (i.a., Vivaldi et al., 2010; Oronoz et al., 2013; Cotik et al., 2015; Díaz de Ilarraza et al., 2015; Oronoz et al., 2015; Díaz de Ilarraza et al., 2017), to name some of the most prolific research lines, all of which remain very relevant today (e.g., Almagro et al., 2020; López-Úbeda et al., 2021; Santiso et al., 2021; Báez et al., 2022; Blanco et al., 2022).

But the real blooming of Spanish medical-oriented NLP occurred just during the second half of the 2010s decade, coinciding with the surge of DL, the Spanish national Plan de Impulso de las tecnologías del Lenguaje or Plan TL (*Plan for the Advancement of Language Technology*) [14], and the first of what is now a long list of shared tasks or community challenges about health-related NLP problems around Spanish-written data, which we present next.

2.5.2 Shared tasks and community challenges

The complete list of these events, up to the year 2021, can be consulted in Table 2.3. As can be seen, the field boasts multiple events per year since 2017, both in national and international venues, and a solid base of participating teams that is growing steadily.

The topics proposed include the detection and classification of a diverse set of target information (e.g., occupations, disabilities, sensitive data, substances), the indexing and coding of documents with standard terminologies (e.g., ICD-10, ICD-O, DeCS) and relation extraction. The standard evaluation frameworks offered by all these events through the generation and sharing of new corpora and guidelines as well as the refereed evaluation processes has undoubtedly helped, along with other actions taken within the Plan TL (info-days, survey reports, supporting open-source software development, etc.), to consolidate this research field, raise its visibility, and strengthen the sense of community.

It must also be noted that none of the shared tasks until SpradIE (Cotik et al., 2021) have posed the challenge of working with real health record texts. Although not able to present the same difficulties that medical notes do (see Section 2.4.2), some shared task organisers have resorted to collecting clinical cases for their campaigns, those being the closest—and more easily accessed and shared—document type to medical notes.

2.5.3 Text embedding representations

Given the central role that these resources play nowadays in virtually all modern NLP systems, we next provide a short overview of the most salient embeddings available for the Spanish language and the biomedical domain, which are listed in Table 2.4. We focus solely on embeddings trained on free unlabelled text.

As can be seen, the Spanish language currently counts with a varied range of embeddings, both static and contextual. The Spanish language is also represented in multiple multilingual embeddings, which have served as highly competitive baselines prior to the publication of the monolingual ones. The corpora used to train these embeddings consist mainly of different mergers of public Internet content, of which Wikipedia is a recurring contributor. That is, most of the Spanish language-specific embeddings that exist today are generic, in the sense that they are not specific to any thematic domain in particular.

Only recently have three sets of contextual biomedical embeddings for the Spanish language been published:

- Flair `es-clinical-X` embeddings [21] trained on the Chilean Waiting List Corpus (Báez et al., 2020b) of de-identified referrals for several speciality consultations.
- mBERT [23], BETO (Cañete et al., 2020) and XLM-RoBERTa (Conneau et al., 2020) embeddings post-trained with oncology clinical cases (López-García et al., 2021).
- RoBERTa clinical and biomedical embeddings (Carrino et al., 2021), trained from scratch respectively on medical notes and reports, and various health-related public sources.

Table 2.3: Shared tasks and community challenges related to clinical NLP in Spanish, up to the year 2021, sorted by year and number of participating teams.

	Name	Host event	Description	Teams	Overview paper or webpage
2017	BARR	IberEval	Abbreviation recognition and resolution	7	Intxaurrondo et al. (2017)
2018	BARR2	IberEval	Abbreviation recognition and resolution	5	Intxaurrondo et al. (2018)
	eHealth-KD	TASS	NERC and relation extraction	6	Martínez Cámara et al. (2018)
	DIANN	TASS	Detection of disability mentions in biomedical literature	8	Fabregat et al. (2018b)
	II Hackathon TL	4YFN	LT hackathon with a track on biomedicine	10	[15]
2019	eHealth-KD	IberLEF	NERC and relation extraction	10	Piad-Morffis et al. (2019)
	MEDDOCAN	IberLEF	Sensitive data detection in medical texts	18	Marimon et al. (2019)
	PharmaCoNER	BioNLP-OST	Pharmacological substances, compounds and proteins NER	22	Gonzalez-Agirre et al. (2019c)
2020	MESINESP	BioASQ	DeCS indexing of biomedical literature	6	Rodriguez-Penagos et al. (2020)
	eHealth-KD	IberLEF	NERC and relation extraction	8	Piad-Morffis et al. (2020)
	CodiEsp	CLEF eHealth	ICD-10 coding for Spanish medical texts	22	Miranda-Escalada et al. (2020b)
	CANTEMIST	IberLEF	CANcer TExt Mining Shared Task	25	Miranda-Escalada et al. (2020a)
2021	SpRadIE	CLEF eHealth	IE from Spanish radiology reports	7	Cotik et al. (2021)
	MESINESP2	BioASQ	DeCS indexing of biomedical literature	7	Gascó et al. (2021)
	eHealth-KD	IberLEF	NERC and relation extraction	8	Piad-Morffis et al. (2021)
	MEDDOPROF	IberLEF	Detection and normalisation of occupation mentions in medical texts	15	Lima-López et al. (2021a)
	ProfNER	SMM4H	Detection of occupation mentions in social media texts	27	Miranda-Escalada et al. (2021)

Table 2.4: Selection of publicly available word embeddings for the Spanish language and/or the biomedical domain, sorted by embedding type and ascending publication date. The number of languages for extremely multilingual models is given between parentheses.

	Name and reference	Language	Corpus
w2v	SBWCE [16]	es	SBWC [16]
	Wikipedia2Vec (Yamada et al., 2020)	es	Wikipedia
fastText	fastText (Grave et al., 2018)	es	Wiki+Common Crawl
	SUCE [17]	es	SUC [18]
	SBWCE [19]	es	SBWC [16]
	MWES (Soares et al., 2019b)	es	SciELO+Wiki (health)
	NLPMedTerm [20]	es	ScieELO+EMEA
Flair	multi-X (Akbik et al., 2018)	multi (343)	JW300
	es-X (Akbik et al., 2018)	es	Wikipedia
	pubmed-X [21]	en	PubMed
	es-clinical-X [21]	es	CWLC [22]
BERT	mBERT [23]	multi (104)	Wikipedia
	BioBERT (Lee et al., 2019)	en	PubMed
	SciBERT (Beltagy et al., 2019)	en	Semantic Scholar
	Clinical BERT (Alsentzer et al., 2019)	en	MIMIC-III [24]
	BETO (Cañete et al., 2020)	es	SUC [18]
	IXAmBERT (Otegi et al., 2020)	es, en, eu	Wikipedia
	mBERT-Galén (López-García et al., 2021)	es	Oncology CC
	BETO-Galén (López-García et al., 2021)	es	Oncology CC
	BERTIN (de la Rosa et al., 2022)	es	mC4-es [25]
	PubMedBERT (Gu et al., 2022)	en	PubMed
RoBERTa	SpanBERTa [26]	es	OSCAR [27]
	XLM-R (Conneau et al., 2020)	multi (100)	Common Crawl
	XLM-R-Galén (López-García et al., 2021)	es	Oncology CC
	Biomedical LM (Carrino et al., 2021)	es	multi-source
	Clinical LM (Carrino et al., 2021)	es	clinical text
	MarIA (Gutiérrez-Fandiño et al., 2022)	es	BNE crawls [28]

It is noteworthy that these resources were published after the experimental work presented here was carried out (and that most of those listed in the table did not exist when the work on this thesis began).

Among biomedical embeddings in other languages, we must mention BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019), Clinical BERT (Alsentzer et al., 2019), and BioALBERT (Naseem et al., 2022), all of them trained on English data, the most exploited sources for this purpose being the bibliographic databases SciELO and PubMed, and the clinical MIMIC-III dataset (A. E. W. Johnson et al., 2016). A comprehensive list of Transformer-based biomedical pre-trained LM can be consulted in Kalyan et al. (2022) and Naseem et al. (2022). Other types of English biomedical embeddings are thoroughly surveyed in Chiu et al. (2020).

2.6 Conclusions

In summary, biomedical NLP is an heterogeneous research field that brings together experts and professionals from a variety of sectors, including healthcare, biomedical research, linguistics, and computer science. IE is the most prolific research area within biomedical NLP. IE tools can be used to extract features relevant to model clinically motivated questions (e.g., prediction of readmission risk), and can also be part of modular NLP pipelines to solve downstream tasks or build end-user applications (e.g., document anonymisation software).

Biomedical NLP is part of the larger NLP field and its evolution has followed similar trends. However, being a knowledge-intensive field, it has maintained a strong focus on rule-based methods to this day. Nevertheless, traditional ML (e.g., SVMs, CRFs) and neural ML (e.g., CNNs, RNNs, and Transformers) are actively being exploited to solve health-related problems.

Biomedical NLP faces many challenges that make it a very particular research domain. Among them, we have discussed the following: data privacy issues, which make it difficult to gather data and reproduce experiments; the usage of non-standard language by healthcare professionals, rendering most off-the-shelf NLP suites inappropriate; the state-of-the-art NLP technology lacking in explainability, and the consequent reservations of the healthcare sector to adopt it; and, finally, the reliance on expert knowledge to craft rules and/or to label corpora.

With respect to biomedical NLP research devoted specifically to the Spanish language, we noted that it is a rather new trend in comparison to that of the English language, in spite of Spanish being one of the largest languages of the world in term of native speakers. Nevertheless, it has attracted a prolific research community with yearly events, as the number of freely available, quality resources steadily increases.

PART II

SENSITIVE DATA DETECTION AND CLASSIFICATION

Chapter 3

Sensitive data: background and literature review

3.1 Definition and motivation

Activities that involve **secondary usage of health data** (that is, the usage of health data outside of direct healthcare delivery [Safran et al., 2007]) such as clinical Natural Language Processing (NLP) and medical research, are expressly subject to regulations and laws that safeguard the patients' rights to privacy and to protect their data. These rules revolve around two key questions, the answers to which vary from one country to another, both in form and content:

a) What pieces of data do the rules apply to?
b) Under what circumstances is the usage of said data allowed?

The two major legislations of reference on data protection to date are the General Data Protection Regulation (GDPR) (2016) of the European Union and the Health Insurance Portability and Accountability Act (HIPAA) (1996) of the United States of America.

Regarding the first question, the subject matter of the GDPR is **personal data**—"any information relating to an identified or identifiable natural person [i.e.,] one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (p. 33).

The HIPAA Privacy Rule protects specifically **individually identifiable health data**, that is, "any information, including demographic information collected from an individual, that (A) is created or received by a health care provider, health plan, employer, or health care clearing house; and (B) relates to the past, present, or future physical or mental health or condition of an individual, the provision of health care to an individual, or the past, present, or future payment for the provision of health care to an individual, and (i) identifies the individual; or (ii) with respect to which there is a reasonable basis to believe that the information can be used to identify the individual" (45 C.F.R. §160.103).

In layman's terms, we will henceforth refer as **sensitive data** to any piece of data protected by the above-mentioned and similar regulations.

Said regulations allow the usage of sensitive data each under particular circumstances and conditions. For instance, explicit patient consent may be required but not considered sufficient, among other considerations. More relevant to this work are the requirements or recommendations that privacy risks should be minimised through technical measures like **anonymisation**, **pseudonymisation** or **de-identification**.

There exists a widespread confusion in the literature surrounding these terms (Chevrier et al., 2019). According to the GDPR, **anonymous information** is that "which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is no longer identifiable" (p. 5). By this definition, anonymisation is an irreversible process. Anonymised information is not affected by data protection regulations because it no longer contains sensitive data.

However, anonymisation is not always a workable solution, either because it may be outright impossible to achieve or guarantee, or because the transformations applied to the data to make them anonymous may render them unsuitable for the intended secondary usage. Instead, data collectors rely more often on pseudonymisation or de-identification to minimise privacy risks, while preserving good-enough levels of data quality and utility.

Pseudonymisation is defined in the GDPR as "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately [...]" (p. 33).

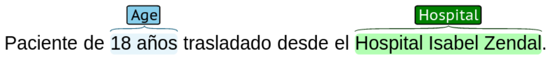
De-identification is achieved, according to the HIPAA, when the information "does not identify an individual and [...] there is no reasonable basis to believe that [it] can be used to identify an individual" (45 C.F.R. §164.514). Unlike the GDPR, which applies to sensitive data of any kind, the HIPAA provides explicit implementation specifications of de-identification of health information. The most relevant to this work is known as the Safe Harbour method. It lists 18 pieces of information (see Figure 3.1) whose removal makes data de-identified, provided that one "does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual".

When it comes to health information in unstructured textual form, NLP can help accelerate pseudonymisation or de-identification processes by automatising the identification of well-defined sensitive data, such as those listed in the HIPAA Safe Harbour provision. From a technical point of view, this task resembles Named Entity Recognition (NER), in that the objective is to locate, and

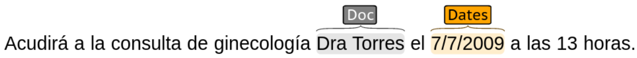
1. Names
2. All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:
    a) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and
    b) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
4. Telephone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section

Figure 3.1: Identifiers of the individual or of relatives, employers, or household members of the individual that must be removed from health data to achieve de-identification under the HIPAA Safe Harbor provision (45 C.F.R. §164.514).

possibly classify, mentions of specific pieces of information within a given text. It is usually approached with supervised sequence labelling techniques, first by classifying each token within a text as being sensitive or not and, optionally, assigning a specific category to the sensitive spans, as pictured in Figure 3.2.



(a) Translation: "18 year old patient transferred from the Isabel Zendal Hospital".



(b) Translation: "[The patient] will attend Dr Torres' gynaecology consultation on 7/7/2009 at 13:00".

Figure 3.2: Annotations of sensitive information and their category.

3.2 Related resources

The first obstacle in the research of sensitive data detection is, unsurprisingly, that the data it needs is jealously protected, as explained above. Public resources are thus scarce. Among the few available to the clinical NLP community is the well-known 2014 i2b2/UTHealth de-identification dataset (Uzuner et al., 2007), accessible at the DBMI portal [29] subject to acceptance of a data use agreement.

In recent years, the following works have been published that involve the development of resources for the Spanish language in particular:

The Spanish/Catalan corpus of health records In this work, Medina et al. (2018) propose a method to incrementally annotate health records with mentions of people, locations, telephone numbers, e-mail addresses, and several alphanumeric identifiers. The method consists in iteratively updating rules specified in the form of Augmented Network Transitions. While the method itself is language agnostic, the experiments involve Spanish and Catalan health records. The annotated corpus resulting form their experimentation is not publicly available.

MEDDOCAN The Medical Document Anonymization (MEDDOCAN) challenge organised by Marimon et al. (2019) [30] is to date the first and only community challenge devoted to the recognition and classification of sensitive data in medical documents in Spanish. The dataset of the challenge consisted of 1,000 clinical cases synthetically augmented with 22 categories of sensitive information. This dataset is publicly available under the Creative

Commons Attribution 4.0 International license terms. It will be thoroughly described in Chapter 4, which explains our participation in the challenge.

DiSMed Pérez-Díez et al. (2021) recently published a collection of 692 brain imaging radiology reports with surrogate realistic sensitive data. The original dataset was semi-automatically annotated for mentions of people, addresses, locations, alphanumeric identifiers, and dates, which where then substituted automatically through rules. The synthetic dataset is publicly available, strictly for research purposes, at the webpage of the Medical Imaging Databank of the Valencia Region [31].

MAPA The Multilingual Anonymisation Toolkit for Public Administrations (MAPA) project (Ajausks et al., 2020; Gianola et al., 2020), funded under the Connecting Europe Facility programme, aimed to develop a text de-identification toolkit for all 24 official European Union languages. Further, it targetted 3 specific application domains—namely, the legal, the administrative, and the clinical. All the sensitive data detection models trained for the project are freely available [32], among which we must highlight the one for the Spanish language and the clinical domain trained on the MEDDOCAN dataset. Its performance metrics have not been published. The datasets developed for other languages and domains can be found at the ELRC-SHARE portal [33].

3.3 State of the Art

The problem of automatic sensitive data detection in clinical text has been tackled in multiple languages other than English, including Norwegian (Tveit et al., 2004), Swedish (Velupillai et al., 2009; Dalianis et al., 2010), French (Chazard et al., 2014), Portuguese (Mamede et al., 2016), Chinese (Jian et al., 2017), German (Seuss et al., 2017), and Dutch (Menger et al., 2018). With respect to the processing of personal data in text written in Spanish, recent studies include Medina et al. (2018) and García-Sardiña (2018). Most notably, the first community challenge about sensitive data in Spanish medical documents, MEDDOCAN (Marimon et al., 2019) [30], was held in 2019 as part of the IberLEF initiative.

The earliest of these proposals are based on dictionary lookup or pattern matching techniques. Gradually, the focus of the field has shifted to machine learning methods, although rule-based approaches are still being proposed due to the lack of annotated data and the fact that the obtained results are good enough to be viable solutions in certain scenarios. Still, more advanced methods are being pursued were possible, due to the fragility of simple rule-based systems, for instance, in the face of typographic errors. This trend is reflected in Table 3.1, which lists the results of recent works on sensitive data detection in Spanish.

Table 3.1: Literature review on sensitive data detection in Spanish clinical text. The number of target categories for NERC problems is given between parenthesis. ZS stands for zero-shot performance. Notice that scores are only comparable if they result from the same evaluation corpus, task and metric.

Reference	Task	Approach	Metric	Score
MEDDOCAN challenge (Marimon et al., 2019)				
Lange et al. (2019)	NER	biLSTM + CRF	exact span F_1	0.97
Hassan et al. (2019)	NER	RegEx + CRF	exact span F_1	0.97
Jabreel et al. (2019)	NER	biLSTM + CNN + CRF	exact span F_1	0.97
Sánchez-León (2019)	NER	Rules	exact span F_1	0.96
Fabregat et al. (2019a)	NER	biLSTM	exact span F_1	0.95
Jiang et al. (2019)	NER	BERT + CRF	exact span F_1	0.95
López-Úbeda et al. (2019)	NER	RegEx + CRF	exact span F_1	0.94
Mao et al. (2019)	NER	BERT + CRF	exact span F_1	0.94
Colón-Ruiz et al. (2019)	NER	biLSTM + CRF	exact span F_1	0.94
Sohrab et al. (2019)	NER	biLSTM	exact span F_1	0.94
Cotik et al. (2019)	NER	CRF	exact span F_1	0.93
Porta-Zamorano (2019)	NER	Rules + CNN	exact span F_1	0.92
Lara-Clares et al. (2019)	NER	biLSTM	exact span F_1	0.90
Suárez-Paniagua (2019)	NER	biLSTM + CRF	exact span F_1	0.87
Pérez-Díez et al. (2021)	NER	biLSTM+CRF$_{ZS}$	token F_1	0.81
Lange et al. (2019)	NERC (29)	biLSTM + CRF	exact span F_1	0.97
Hassan et al. (2019)	NERC (29)	RegEx + CRF	exact span F_1	0.96
Sánchez-León (2019)	NERC (29)	Rules	exact span F_1	0.96
Jabreel et al. (2019)	NERC (29)	biLSTM + CNN + CRF	exact span F_1	0.96
Fabregat et al. (2019a)	NERC (29)	biLSTM	exact span F_1	0.94
Jiang et al. (2019)	NERC (29)	BERT + CRF	exact span F_1	0.94
Mao et al. (2019)	NERC (29)	BERT + CRF	exact span F_1	0.94
Colón-Ruiz et al. (2019)	NERC (29)	biLSTM + CRF	exact span F_1	0.93
Sohrab et al. (2019)	NERC (29)	biLSTM	exact span F_1	0.93
Porta-Zamorano (2019)	NERC (29)	Rules + CNN	exact span F_1	0.92
López-Úbeda et al. (2019)	NERC (29)	RegEx + CRF	exact span F_1	0.90
Lara-Clares et al. (2019)	NERC (29)	biLSTM + CRF	exact span F_1	0.90
Cotik et al. (2019)	NERC (29)	CRF	exact span F_1	0.90
Suárez-Paniagua (2019)	NERC (29)	biLSTM	exact span F_1	0.86
Pérez-Díez et al. (2021)	NERC (29)	biLSTM+CRF$_{ZS}$	token F_1	0.59
Tested on private corpora				
Medina et al. (2018)	NERC (7)	CRF	F_1	0.77
Pérez-Díez et al. (2021)	NER	biLSTM+CRF	token F_1	0.98
Pérez-Díez et al. (2021)	NERC (7)	biLSTM+CRF	token F_1	0.93

As can be seen, the bulk of the works propose systems based on bidirectional LSTMs (biLSTM) with or without Conditional Random Field (CRF) classifiers, as was the standard approach to sequence labelling problems in NLP before Transformer-based systems became ubiquitous. The winners of MEDDOCAN—the Neither-Language-nor-Domain-Experts (NLNDE) (Lange et al., 2019)—achieved F1-scores as high as 0.975 in the task of sensitive information detection and categorisation by using this type of Recurrent Neural Networks (RNN). Nevertheless, several of the rule-based systems that participated in the challenge managed to achieve very competitive results, even surpassing systems built on fine-tuned Transformer models.

The next chapter presents our official participation in the MEDDOCAN challenge, where we ranked third with a feature-rich biLSTM model, as well as additional experiments we carried out with Bidirectional Encoder Representations from Transformers (BERT) after the challenge finished. In Chapter 5, we perform similar experiments on a new corpus, and study how well the MEDDOCAN models can be transferred from one to the other.

Chapter 4

Sensitive data: the MEDDOCAN challenge

4.1 Introduction

The major bottleneck for the advancement of Natural Language Processing (NLP) in the medical field is the struggle to access real clinical texts, mainly due to data privacy protection issues. Medical Document Anonymization (MEDDOCAN) (Marimon et al., 2019) [30] was the first challenge devoted to the recognition and classification of sensitive data in medical documents in Spanish. This chapter describes part of Vicomtech's official participation in MEDDOCAN as well as improved post-challenge results.

The challenge proposed two tasks of incremental difficulty: sensitive span detection, and sensitive span detection and classification into one of 29 categories. That is, it is a task akin to Named Entity Recognition and Classification (NERC), usually tackled as a sequence labelling problem. The MEDDOCAN corpus consists of clinical case reports manually enriched with sensitive information. That is, it is a synthetic corpus. In the next chapter, we conduct analogous experiments in real health records.

Our aim for the challenge was to test a variety of then state-of-the-art approaches, neural and shallow. Specifically, Conditional Random Fields (CRF) (Lafferty et al., 2001) were prominently featured, having been extensively used for similar tasks of sequential nature, including textual sensitive data identification; the other techniques used are neural networks such as Convolutional Neural Networks (CNN) (LeCun et al., 1989) and Long Short-Term Memories (LSTM) (Hochreiter et al., 1997). At a later stage, we evaluated the more recent architecture Bidirectional Encoder Representations from Transformers (BERT), outperforming our official results.

The chapter is structured as follows: Section 4.2 starts describing the task's data and the set of features extracted to characterise it; then, the systems with which the reported results were obtained are presented. The results are reported and analysed in Section 4.3. Finally, the chapter ends by presenting the conclusions reached in Section 4.4.

4.2 Materials and methods

4.2.1 Data

Although the organisers' instructions for the challenge did not state explicitly whether the competition was constrained or not, we treated it as such by focusing solely on the MEDDOCAN corpus as the training and development data to learn our models. In what follows, we describe the corpus itself and explain how we handled the inputs and outputs of the systems.

4.2.1.1 The MEDDOCAN corpus

The organisers of the MEDDOCAN shared task curated a synthetic corpus of clinical case reports enriched with sensitive information by health documentalists. The size of the corpus is shown in Table 4.1. The annotation scheme comprises 29 fine-grained sensitive information types (of which only 22 are represented in the corpus), whose definition was inspired by the General Data Protection Regulation (GDPR) of the European Union, as well as the annotation guidelines of the i2b2 de-identification tracks (Aramaki et al., 2006; Stubbs et al., 2015), in turn based on the Health Insurance Portability and Accountability Act (HIPAA) of the United States of America.

Table 4.1: Size of the MEDDOCAN corpus

	Train	**Dev**	**Test**
# documents	500	250	250
# tokens	360,407	138,812	132,961
Vocabulary	26,355	15,985	15,397
# annotations	11,333	5,801	5,661

The distribution of the 22 represented categories is described in Table 4.2[1]. As can be seen, the corpus is highly unbalanced. Each document has 22.80 sensitive spans in average, with territories (`Ter`) and dates (`Dat`) accounting for almost 30% of all the occurrences, while some categories do not even amount to %1. It is also noteworthy that, at the same time, all MEDDOCAN documents follow a highly predictable pattern:

[1]Note that the label names used throughout this document are no the official ones; please consult Appendix A for a complete list of equivalences.

- an initial semi-structured section with personal information of the patient,
- the clinical case with a few pieces of personal information (e.g., the patient's age, dates, and other less frequent types of personal information), and
- a final paragraph with data about the referring doctor.

Table 4.2: Sensitive data type distribution in the MEDDOCAN corpus

	Train		Dev		Test		All	
	#	%	#	%	#	%	#	per doc
Territory (`Ter`)	1,875	16.54	987	17.01	956	16.89	3,818	3.82 ± 1.26
Date (`Dat`)	1,231	10.86	724	12.48	611	10.79	2,566	2.57 ± 1.92
Patient's age (`Age`)	1,035	9.13	521	8.98	518	9.15	2,074	2.07 ± 0.54
Patient's name (`Pat`)	1,009	8.90	503	8.67	502	8.87	2,014	2.01 ± 0.14
Doctor's name (`Doc`)	1,000	8.82	497	8.57	501	8.85	1,998	2.00 ± 0.13
Patient's sex (`Sex`)	925	8.16	455	7.84	461	8.14	1,841	1.85 ± 0.56
Street (`Str`)	862	7.61	434	7.48	413	7.30	1,709	1.71 ± 0.49
Country (`Ctr`)	713	6.29	347	5.98	363	6.41	1,423	1.42 ± 0.67
Patient's ID (`Pid`)	567	5.00	292	5.03	283	5.00	1,142	1.14 ± 0.40
E-mail address (`Ema`)	469	4.14	241	4.15	249	4.40	959	0.96 ± 0.33
License ID (`Lid`)	471	4.16	226	3.90	234	4.13	931	0.93 ± 0.26
Insurance ID (`Iid`)	391	3.45	194	3.34	198	3.50	783	0.78 ± 0.42
Hospital (`Hos`)	255	2.25	140	2.41	130	2.30	525	0.53 ± 0.57
Patient's relative (`Kin`)	243	2.14	92	1.59	81	1.43	416	0.42 ± 1.41
Institution (`Ins`)	98	0.86	72	1.24	67	1.18	237	0.24 ± 0.82
Episode ID (`Eid`)	77	0.68	32	0.55	39	0.69	148	0.15 ± 0.36
Phone number (`Pho`)	58	0.51	25	0.43	26	0.46	109	0.11 ± 0.34
Patient's profession (`Job`)	24	0.21	4	0.07	9	0.16	37	0.04 ± 0.24
Fax number (`Fax`)	15	0.13	6	0.10	7	0.12	28	0.03 ± 0.17
Other (`Oth`)	9	0.08	6	0.10	7	0.12	22	0.02 ± 0.16
Outpatients clinic (`Cli`)	6	0.05	2	0.03	6	0.11	14	0.01 ± 0.12
Doctor's ID (`Did`)	0	0.00	1	0.02	0	0.00	1	0.00 ± 0.03
Total	11,333		5,801		5,661		22,795	22.80 ± 3.88

The composition of the initial and last segments is very similar across all the documents in the corpus (see an example in Figure 4.1). Thus, it is expected that the systems perform satisfactorily on these repetitive parts and the categories of sensitive information contained therein, while struggling in the segment consisting of the clinical case, where the types of personal information and the ways they are presented in free text are more diverse.

4.2.1.2 Data representation

As Figure 4.1 shows, the corpus is distributed in brat standoff format (Stenetorp et al., 2012), that is, the annotations are defined at span level as opposed to in a

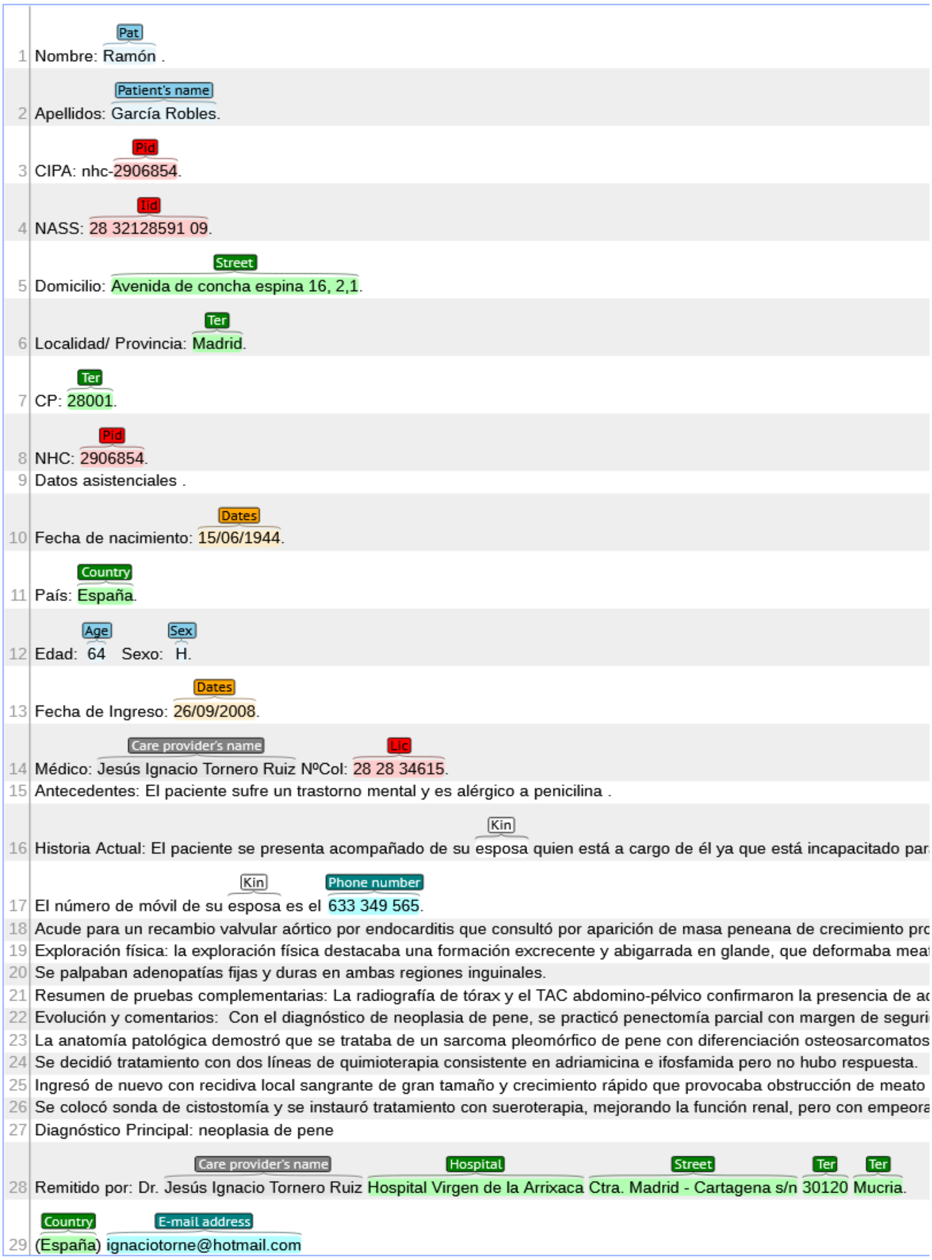


Figure 4.1: A MEDDOCAN document visualised in the brat interface

per-token basis, the latter being typically the format expected by learning algorithms for sequence labelling. Consequently, the corpus had to be pre-processed as follows:

1. **Paragraph splitting.** Documents were split into paragraphs using line breaks in the original texts. We decided to work with paragraphs instead of sentences because the suggested sentence-splitting tool (the SPACCC PoS Tagger [34]) occasionally split parts of target entities into different sentences.

2. **Tokenisation.** Each paragraph was tokenised using the SPACCC PoS Tagger and some extra custom tokenisation rules, mainly to split punctuation symbols if not inside a URL, e-mail address or date, and to split camel cased words in order to account for spacing errors in the original text (e.g., 'DominguezCorreo' into 'Dominguez Correo').

3. **Label formatting.** The brat-formatted annotations of the training and development datasets were converted to token-level tags following the BILOU scheme: Beginning (`B-`), Inner (`I-`), Last (`L-`), Outside (`O`), Unique (`U-`). Combining this tag scheme with the original 22 granular sensitive data categories—e.g., for the granular class `Dat` we would have the tags `B-Dat`, `I-Dat`, `L-Dat`, `U-Dat`, plus the generic `O` class—gives a tagset of 89 possible unique labels.

The outcome of the pre-processing is illustrated in Examples E1 and E2 derived respectively from sentences 5 and 17 in Figure 4.1:

E1

Domicilio	`O`
:	`O`
Avenida	`B-Str`
de	`I-Str`
concha	`I-Str`
espina	`I-Str`
16	`I-Str`
,	`I-Str`
2,1	`L-Str`
.	`O`

E2

El	`O`
número	`O`
de	`O`
móvil	`O`
de	`O`
su	`O`
esposa	`U-Kin`
es	`O`
el	`O`
633	`B-Pho`
349	`I-Pho`
565	`L-Pho`
.	`O`

With the corpus formatted thus, we extracted a rich set of features common in similar Named Entity Recognition (NER) tasks, and other features motivated by the particularities of the corpus just described. Succinct descriptions of the features are listed below. The features can be organised into three big groups, depending on what they aim to describe: features for token characterisation, term characterisation and context characterisation.

4.2.1.2.1 Token characterisation This group of features aims at characterising the shape of each token, regardless of the context they occur in and their meaning.

Token The token itself.

Length The length in characters of the token.

Casing Features related to the token's casing, i.e., whether the token is uppercase, lowercase or titlecase, and the ratio of uppercase characters to **Length**.

Digits and punctuation Features related to the token's character types, e.g., whether the token is alphanumeric or a punctuation mark, the ratio of the number of punctuation marks to the token's length, and so on.

Affixes The token's first and last character bigrams and trigrams.

4.2.1.2.2 Term characterisation This group of features attempts to describe the intended meaning of the tokens. It includes lexical, morphologic, syntactic, and semantic features.

Linguistic information The lemma and Part of Speech (PoS) tag given by the SPACCC tagger at the data pre-processing step.

NERC The named entity tag given by spaCy (model `es_core_news_md 2.1.0`). If a detected named entity was multi-word, we gave the same tag to all the tokens involved.

Date-time expressions Whether the token is part of a date and/or time expression according to a left-to-right parser designed for this specific purpose in ANTLR4 for Python (`antlr4-python3-runtime 4.7.2`).

Gazetteers The maximum similarity score obtained when matching text n-grams with gazetteer entries. We used a total of 10 gazetteers: the ones provided by the organisers [35], plus country names, kinship relations, months, and sexes (compiled manually for this task). The string similarity was computed with the `python-Levenshtein` library and was only added as feature if it was greater than 0.75. If a match was multi-word, we gave the same score to all the tokens involved.

Brown clusters Complete paths and paths pruned at lengths 8, 16, 32, and 64. The clusters (P. F. Brown et al., 1992) were computed on the training set's vocabulary with `tan-clustering` [36], using the default settings of the tool.

4.2.1.2.3 Context characterisation The last group of features attempts to provide a topological description of the documents. This group of features was motivated by the particular shape of the documents described earlier.

Boundaries Whether the token is first or last in the paragraph.

Length The length in tokens of the paragraph the token belongs to.

Position The normalised position of the paragraph in the document.

Header The nearest expression to the left of each token that is followed by a colon, lowercased (e.g., 'email:', 'antecedentes familiares:', and so on).

In addition, the features for a given token include features from the neighbouring tokens in a ± 3 context window with respect to that token, except for context **Length** and **Position** features (which are the same for neighbouring tokens). Note that the final models draw upon a different set of features in each case. This is detailed in their respective sections.

4.2.1.3 Output handling

The raw output of the models has the same format as that described for the training data: one label per input token, each label consisting of a BILOU tag and a sensitive data category (see Section 4.2.1.2). Predicted labels were post-processed to ensure that the results were well-formed in terms of the BILOU scheme, having the BILOU tag prevail over the sensitive data category tag in case of conflict—e.g., the sequence (`B-Str L-Ter`) would be converted to (`B-Str L-Str`) instead of (`U-Str U-Ter`). Finally, the predictions were converted to the format required by the organisers: the span-level brat standoff format.

4.2.2 Systems

In what follows, we describe the implementation details of 3 systems submitted to the challenge (namely, spaCy, CRF and NCRF$_{++}$), and one system developed afterwards (BERT). The same systems were used for the two tasks of the challenge: *i)* sensitive span detection, and *ii)* sensitive span detection and classification. Specifically, the systems were trained to learn jointly the detection and classification task, and their results are evaluated in both scenarios.

4.2.2.1 spaCy

As a first approach to the task, we experimented with spaCy's [37] NER implementation (version `2.1.3`). spaCy is an open source Python library for application-oriented NLP; it offers implementations of models of proved efficacy for the main NLP tasks, as well as pre-trained models in multiple languages. spaCy's NER architecture includes Bloom Embeddings (Serrà et al., 2017), residual CNNs (He et al., 2016) and a transition-based approach [38]. We followed the given recipe [39] with default settings and applied the recommended tweaks: compounding batch size, dropout decay, and parameter averaging.

spaCy supports a closed set of features, which overlaps only partially with those described in Section 4.2.1.2. Interestingly, training an empty model yielded better results on the development set than using the compatible computed features. Likewise, training embeddings from scratch also gave better results than using pre-trained Spanish embeddings of the medical domain (Soares et al., 2019b). Thus, the results submitted to the task were obtained with a NER model trained from scratch—with spaCy's basic pipeline for Spanish—, and no extra information provided but the challenge data.

4.2.2.2 CRF

The second official run corresponded to a system based on Conditional Random Fields (CRF), implemented using the Python `sklearn-crfsuite` library (version `0.3.6`). For years, CRF classifiers have established the state of the art in many NLP tasks of sequential nature, and are still used extensively, also for sensitive data detection, despite achieving overall moderately worse results than modern techniques based on deep learning (Leevy et al., 2020).

Our final CRF model did not include date-time expressions as features, because they yielded slightly worse results in previous feature selection trials explored to reduce dimensionality. Features with float values were rounded to one decimal. The final system was trained using the configuration shown in Table 4.3.

Table 4.3: CRF configuration

Parameter	Value	Parameter	Value
Algorithm	`lbfgs`	c2	`0.1`
Max iterations	`100`	All transitions	`True`
c1	`0.1`		

4.2.2.3 NCRF++

NCRF++(J. Yang et al., 2018b) [40] is an open-source toolkit built on PyTorch to train neural sequence labelling models. We kept the toolkit's default network configuration: an initial CNN layer for character sequence representations, a bidirectional LSTM (biLSTM) layer for word sequence representations and an output CRF classifier. This architecture has shown to be one of the most competitive among variants of Recurrent Neural Networks (RNN) in tasks of sequential nature (J. Yang et al., 2018a).

The hyperparameter settings used to train our model are shown in Table 4.4 (any missing hyperparameter would be set to the toolkit's default value). Regarding the features, in this case we used all the available ones (see Section 4.2.1.2). The character embeddings were initialised randomly and trained on the given corpus. The word embeddings were initialised with pre-trained Spanish embeddings of the medical domain (Soares et al., 2019b), specifically consisting of Word2Vec embeddings (Mikolov et al., 2013a,b) of 300 dimensions trained on SciELO and Wikipedia. The maximum sentence length was set to 250 tokens during training; for prediction, the length was not restricted. The model was trained for a maximum of 30 epochs, after which the checkpoint with best results on the development set was chosen as the final model to process the test set.

Table 4.4: NCRF++ hyperparameters

Hyperparameter	Value	Hyperparameter	Value
Character emb dimensions	30	Batch size	100
Character CNN layers	1	Optimiser	Adam
Character hidden dimensions	50	Learning rate	0.01
Word emb dimensions	300	L_2 regularisation	1e-6
Word biLSTM layers	1	Learning rate decay	0.05
Word hidden dimensions	150	Ave batch loss	True
Dropout rate	0.5	Max epochs	30

4.2.2.4 BERT

BERT has shown an outstanding performance in NERC-like tasks, having improved the start of the art for almost every dataset and language upon its publication (Devlin et al., 2019). In this work, we took the standard approach of topping a BERT encoder—i.e., Multilingual BERT (mBERT) [23]—with dropout and fully connected layers.

Naturally, to the input representation explained in Section 4.2.1.2, one must add the steps necessary to prepare the input for a BERT encoder in the case of this system: tokenising the input into subwords with the appropriate BERT tokeniser, adding BERT's special tokens `[CLS]` and `[SEP]` at the beginning and end, respectively, of each resulting sequence, and padding them to a fixed length in order to be able to process sequences in batch.

In our implementation, the prefix of each token—i.e., the first subword—received the label for that word, while the rest of the subwords, marked by BERT with leading `##`, received the label `X`. This is depicted in Examples E3 (before mBERT tokenisation) and E4 (after; special BERT tokens are not shown):

E3

NºCol	O
:	O
28	B-Lid
28	I-Lid
34615	L-Lid

E4

Nº	O
##C	X
##ol	X
:	O
28	B-Lid
28	I-Lid
346	L-Lid
##15	X

During training, the cross-entropy loss was computed over entire sequences except padding positions. In inference, the prediction for the first subword is assigned to the entire token, i.e., predictions for suffix positions are ignored when reconstructing the output of the model.

Table 4.5: BERT hyperparameters

Hyperparameter	**Value**	**Hyperparameter**	**Value**
Pre-trained model	mBERT$_{Base}$ Cased	Learning rate	`3e-5`
Batch size	`12`	Gradient clipping	`1.0`
Max input length	`500`	Scheduler	Linear warm-up
Optimiser	Weighted Adam	Early stopping patience	`15 epochs`

This implementation was built on PyTorch (`torch 1.2.0`) and Hugging Face's open-source `transformers` library (Wolf et al., 2020) [41] (version `2.4.1`). The hyperparameters can be consulted in Table 4.5. The metric monitored for the early stopping was the token-level F_1-score over binarised predictions, where special BERT tokens and tokens labelled as `O` or `X` are the negative class, and all the other categories are the positive class.

4.2.3 Evaluation

MEDDOCAN consists of two scenarios:

- **Detection** (officially known as "Sensitive span detection" [30]): this evaluation measures how good systems are at detecting sensitive text spans, regardless of the category assigned to those spans. This scenario is closer to real-word applications whose objective is to conceal confidential data.
- **Detection and Classification** (officially known as "NER offset and entity type classification" [30]): in this scenario, systems are required to match exactly not only the boundaries of each sensitive span, but also the category assigned. In practice, knowing the category of sensitive data not only makes a redacted text more legible to people, but it may also be useful for downstream automatic text processing tasks such as substituting the sensitive data with analogous fake data.

Both tasks are officially evaluated in terms of micro-average F_1-score (F_1), the harmonic mean of precision (P) and recall (R):

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{4.1}$$

where true positives (TP), false positives (FP) and false negatives (FN) are defined differently for each task, as explained below. The three metrics reach their best value at 1. Intuitively, recall measures how many true instances have been correctly predicted, while precision measures how correct the predictions made are.

In the case of the detection scenario, the predictions are counted as follows:

- TP: number of predicted spans that match in boundaries—i.e., start and end positions of the spans in the document, expressed in characters—with a gold span.
- FP: number of predicted spans that do not match in boundaries with any gold span (also known as *spurious predictions*).
- FN: number of gold spans that do not match in boundaries with any predicted span (also known as *missing predictions*).

The matches are required to be exact, that is, predictions that overlap partially with a gold span are counted as errors. We will henceforth refer to the results evaluated thus as **strict detection**. In addition, MEDDOCAN organisers provide a laxer evaluation where the sensitive spans connected by non-alphanumerical characters are merged into one. We will henceforth refer to this variant as **merged detection**.

Regarding the detection and classification scenario (**classification** for short), the definitions for TP, FP and FN are the same, except that for a prediction to count as correct it is required to match with a gold annotation in category as well as boundaries. This scenario has an additional metric, leak (Lk), that is defined as follows:

$$Lk = \frac{FN}{\#\ sentences} \tag{4.2}$$

As leak measures the number of missing predictions per sentence, it reaches its best value at 0. There are no merged metric variants for this scenario.

In the task at hand, systems with high recall and lower precision are preferred over systems with high precision and lower recall, given that the leakage of sensitive data potentially carries far more severe, damaging consequences than the over-obfuscation of non-sensitive data. Still, precision is also desirable to preserve as much as possible the original meaning and the readability of the documents.

Table 4.6: Official and post-challenge (*) results of MEDDOCAN. Best and second-best results are highlighted in boldface and underlined respectively. The first section of the table corresponds to the systems described in this work, while the second section reports the results of three competitors. Models are described as language-dependent (l), language- and domain-dependent (l+d) or neither.

		Merged Detection			Strict Detection			Classification			
		P	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**Lk**
spaCy	l	0.982	0.961	0.972	0.967	0.953	0.960	0.965	0.948	0.956	0.039
CRF	l+d	0.983	0.950	0.966	**0.977**	0.943	0.960	**0.971**	0.937	0.954	0.048
NCRF_{++}	l+d	0.979	0.972	0.976	0.972	0.964	0.968	0.964	0.956	0.960	0.033
BERT*		0.982	0.981	0.982	0.973	0.972	0.973	0.968	0.967	0.967	0.025
Hadoken	l+d	0.974	0.923	0.948	0.968	0.919	0.943	0.965	0.912	0.937	0.036
Jiang et al.		0.980	**0.983**	0.982	0.933	0.958	0.946	0.928	0.952	0.940	0.036
NLNDE S2	l	0.986	**0.983**	**0.985**	0.976	0.973	0.974	**0.971**	0.968	**0.970**	0.024
NLNDE S3	l+d	**0.987**	**0.983**	**0.985**	0.975	**0.975**	**0.975**	0.970	**0.969**	**0.970**	**0.023**

4.3 Results

Table 4.6 shows the results achieved for both scenarios. Alongside the results of the systems described earlier (Section 4.2.2), we report the results of three MEDDOCAN competitors as references. The first two are systems based on BERT: Jiang et al. (2019) competed with a mBERT + CRF system; Hadoken (Mao et al., 2019) is a hybrid system that uses also a mBERT + CRF tagger along with gazetteer lookup and regular expressions. Next, we report the results of the Neither-Language-nor-Domain-Experts (NLNDE) (Lange et al., 2019), the winners of the challenge. NLNDE competed with a biLSTM + CRF setup, exploiting combinations of pre-trained word and character embeddings. We report their two best runs, one domain dependent (S3) and the other independent (S2).

4.3.1 Official submissions

Regarding our official submissions, all the systems achieved F_1-scores over 0.950 even on the hardest scenario (i.e., classification), the best F_1-scores being 0.968 and 0.960 for the detection and classification tasks, respectively. All systems favour precision over recall. Among individual systems, NCRF_{++} has the best scores; particularly, it has a markedly better recall than the rest. This system granted our team the third position in all the tasks of the competition, interestingly surpassing the two BERT-based system in the strict evaluations. On the other hand, CRF outperforms the other systems in precision, but the lower recall relegates it to the last position in the rank.

4.3.2 Post-challenge experiments

After MEDDOCAN's evaluation campaign, we trained a new model based on the BERT architecture. As shown in Table 4.6, compared to our official submissions BERT manages to improve recall scores markedly, achieving an F_1-score as high as 0.967 in the classification scenario, the most difficult and strict of all. Even then, it does not improve the scores obtained by neither the domain-dependent (S3) nor the domain-independent (S2) NLNDE models (Lange et al., 2019), although it remains just 0.03 F_1-score points behind them. In fact, it would have achieved the second position among all the MEDDOCAN shared task competitors without any language nor domain-specific knowledge. What is more, our BERT implementation is also the only system among those reported here that does not instil into the model the sequential nature of the problem. We expected that Hadoken (Mao et al., 2019) and Jiang et al. (2019), consisting both of a CRF classifier on top of a mBERT encoder, would have surpassed our results for this same reason, but they achieve overall lower results. The reasons why it is so remain unclear. Interestingly, Jiang et al. (2019) achieve similar results to ours in the merged detection scenario, but their performance drops sharply in the stricter evaluations. They argue that the loss is due to flawed pre- and post-processing steps having to do with segmentation.

4.3.3 Error analysis

An error analysis showed that our systems made very similar errors, although with varying frequencies. As expected, most of the false negatives involved entities located at the least structured parts of the documents and usually affected the types patient's relative (`Kin`), patient's profession (`Job`), and other less frequent categories. Another category difficult to predict correctly was street (`Str`), because the systems segmented them into spans different to those in the gold annotations. Finally, a few errors stemmed from similar categories which the models confuse, such as phone number (`Pho`) and fax number (`Fax`), outpatients clinic (`Cli`), institution (`Ins`) and hospital (`Hos`), and identification numbers. All these were correctly recognised but incorrectly categorised on a few occasions.

Regarding false positives, most of them corresponded to improperly segmented addresses and the misclassification of numeric expressions. The rest of falsely predicted sensitive spans were most frequently entities seemingly missed by the human annotators. In general, as the presented metrics indicate, the BERT-based system managed to miss fewer sensitive data, most importantly in the less represented and more variable categories as well.

Appendix B contains full confusion matrices of all the systems.

4.4 Conclusions

In this chapter we described Vicomtech's approach to the Medical Document Anonymization (MEDDOCAN) challenge. MEDDOCAN was the first community challenge devoted to the detection and classification of sensitive data in text of the health domain written in Spanish. It was also the first attempt at defining an inventory of textual sensitive data types for the Spanish health sector in response to the recent data privacy policies formulated by the European Union.

In our official participation, we tested a variety of sequence labelling algorithms and systems—namely, spaCy's NER tagger, a CRF classifier, and an RNN model (NCRF$_{++}$). The latter obtained the best scores, with an F_1-score of 0.960 in the sensitive data detection and classification task. In this chapter we also presented unofficial results of a model based on Multilingual BERT (mBERT), with which we improved our previous results with an F_1-score of 0.968 thanks to the system's higher recall. Despite being language and domain independent, this model falls only 0.3 F_1-score points behind the competition winners.

Taking into account that only 3% of the gold labels remain incorrectly annotated, the challenge can be considered almost solved, and it is not clear if the small differences among the systems are actually significant, or whether they stem from minor variations in initialisation or a long tail of minor labelling inconsistencies. Furthermore, given the synthetic nature of the corpus, there exists a serious risk that the models might have overfit the MEDDOCAN corpus, rendering the models trained on this corpus unfit for usage in the real medical documents. In the next chapter, we conduct the same experiments in a corpus of health records.

Chapter 5

Sensitive data: experiments with health records

5.1 Introduction

In the previous chapter, we conducted experiments on sensitive data detection and classification using a synthetic corpus: the MEDDOCAN corpus. We observed that the systems proposed and other competitors of the MEDDOCAN challenge managed to obtain F_1-scores as high as 0.970 (Lange et al., 2019). The success of the systems is certainly explained, at least in part, by the homogeneous structure and data distribution of the synthetic corpus. It is possible, in consequence, that the MEDDOCAN corpus does not paint an entirely realistic picture of the efficacy of the current NLP technology where the task of sensitive data detection and classification is concerned.

In this chapter, we reproduce MEDDOCAN's experimentation in NUBes-PHI, a corpus of health records manually annotated with sensitive data. First, we provide a comprehensive description of NUBes-PHI and compare it to the MEDDOCAN corpus in detail. Then, we train and test in NUBes-PHI the same systems evaluated in Chapter 4.

In addition, we carry out zero-shot evaluations of the systems trained in the MEDDOCAN corpus, in order to assess the extent to which they are able to transfer the knowledge gained in one corpus to the other.

Finally, we also compute the train curves of the systems, so as to understand the training data necessities of the different tested systems.

The rest of the chapter is structured as follows: Section 5.2 describes the corpus of health records; next, it goes briefly over the experimentation setup, pointing out the differences with respect to the previous chapter where needed and otherwise referring the reader to the corresponding sections. The results are reported and analysed in Section 5.3. Finally, Section 5.4 presents the conclusions drawn from the work carried out in the chapter.

5.2 Materials and methods

5.2.1 Data

NUBes is a corpus of medical reports written in Spanish and annotated with negation and speculation information. It is the subject of Chapter 10. Before being published, sensitive information had to be manually annotated and replaced for the corpus to be safely shared. In this chapter, we work with the NUBes version prior to its anonymisation, that is, with the manual annotations of sensitive information. In order to avoid confusion between the two corpus versions, we henceforth refer to the version relevant in this chapter as NUBes-PHI (from NUBes with Personal Health Information).

In what follows, we first describe NUBes-PHI and then compare it with the corpus of the Medical Document Anonymization (MEDDOCAN) challenge, described earlier in Chapter 4. Note that this chapter does not dive into the manual annotation process, nor does it motivate or discuss the annotation policy defined, but simply exploits their outcome. Finally, we describe how the inputs to and outputs of the systems are transformed and handled.

5.2.1.1 The NUBes-PHI corpus

NUBes-PHI consists of 32,055 sentences annotated for 12 sensitive information categories. Overall, it contains 7,818 annotations. The corpus has been randomly split into train (72%), development (8%) and test (20%) sets to conduct the experiments described in this chapter. The size of each split and the distribution of the annotations by category can be consulted in Tables 5.1 and 5.2, respectively.

The majority of sensitive information in NUBes-PHI are temporal expressions—dates (`Dat`) and times (`Tim`)—, followed by mentions of healthcare facilities (`Fac`) and the age of patients (`Age`). Mentions of people are not that frequent, with doctor names (`Doc`) occurring much more often than patient names (`Pat`). The least frequent sensitive information types, which account for $\sim$10% of the remaining annotations, consist of the sex of patients (`Sex`), patient professions (`Job`), and information about relatives of patients (`Kin`); locations (`Loc`)

Table 5.1: Size of the NUBes-PHI corpus

	Train	Dev	Test
# sentences	23,079	2,565	6,411
# tokens	379,401	41,936	107,024
Vocabulary	25,304	7,483	12,750
# annotations	5,570	623	1,579

Table 5.2: Sensitive data type distribution over dataset splits in the NUBES-PHI corpus

	Train		Dev		Test		All	
	#	%	#	%	#	%	#	per doc
Date (`Dat`)	2,169	38.87	251	40.29	660	41.80	3,076	0.45 ± 1.07
Healthcare facility (`Fac`)	1,012	18.17	105	16.85	275	17.42	1,392	0.20 ± 0.55
Patient's age (`Age`)	701	12.59	77	12.36	200	12.67	978	0.14 ± 0.35
Time (`Tim`)	608	10.92	63	10.11	155	9.82	826	0.12 ± 0.43
Doctor's name (`Doc`)	486	8.73	44	7.06	134	8.49	664	0.09 ± 0.35
Patient's sex (`Sex`)	270	4.85	35	5.62	71	4.50	376	0.05 ± 0.23
Patient's relative (`Kin`)	158	2.84	20	3.21	44	2.79	222	0.03 ± 0.25
Location (`Loc`)	71	1.27	10	1.61	19	1.20	100	0.01 ± 0.14
Patient's name (`Pat`)	48	0.86	5	0.80	11	0.70	64	0.01 ± 0.15
Patient's profession (`Job`)	31	0.56	3	0.48	9	0.57	43	0.01 ± 0.09
Contact information (`Con`)	8	0.14	2	0.32	0	0.00	10	0.00 ± 0.05
Other (`Oth`)	12	0.22	8	1.28	1	0.06	21	0.00 ± 0.07
Total	5,570		623		1,579		7,772	1.14 ± 1.91

other than healthcare facilities; and contact information (`Con`), such as phone numbers and e-mail addresses. Finally, the category other (`Oth`) includes, for instance, mentions to agencies unrelated to healthcare and whether the patient is right- or left-handed. It occurs just 21 times. The distribution of sensitive data over medical specialities and record sections can be consulted in Appendix C.

5.2.1.2 NUBES-PHI and MEDDOCAN

The MEDDOCAN corpus (Tables 4.1 and 4.2 in the previous chapter) and NUBES-PHI differ primarily in the frequency and distribution of the sensitive data they contain. While the corpora are similar in size (NUBES-PHI 632K vs MEDDOCAN 528K tokens), MEDDOCAN contains almost thrice the annotations (7,772 vs 22,795). This is mainly because NUBES-PHI documents do not contain semi-structured sections with metadata like those of MEDDOCAN do.

Furthermore, the sensitive data types considered in MEDDOCAN differ in part from those in NUBES-PHI. Specifically, MEDDOCAN contains finer-grained labels overall. Nevertheless, an approximate mapping between the two sets can be established, as declared in Table 5.3, which will be helpful throughout the chapter. A notable difference is that NUBES-PHI does not contain identification numbers (`Ide`), therefore, no such category was included in the annotation guidelines. In sharp contrast, MEDDOCAN distinguishes 5 identifiers: patient's ID (`Pid`), license ID (`Lid`), insurance ID (`Iid`), episode ID (`Eid`), and doctor's ID (`Did`). Finally, MEDDOCAN's annotation policy explicitly bans the annotation of time mentions, while they are annotated in NUBES-PHI (as `Tim`). For practical purposes, we map NUBES-PHI's `Dat` and `Tim` to MEDDOCAN's `Dat`.

Table 5.3: Equivalences established between MEDDOCAN and NUBes-PHI sensitive data categories in order to facilitate corpus comparison and zero-short experiments. The central column indicates the name given to each equivalence.

MEDDOCAN				NUBes-PHI
Dat	=	Dat	=	Dat + Tim
Hos + Ins + Cli	=	Fac	=	Fac
Age	=	Age	=	Age
Doc	=	Doc	=	Doc
Sex	=	Sex	=	Sex
Kin	=	Kin	=	Kin
Ter + Str + Ctr	=	Loc	=	Loc
Pat	=	Pat	=	Pat
Job	=	Job	=	Job
Ema + Pho + Fax	=	Con	=	Con
Oth	=	Oth	=	Oth
Pid + Lid + Iid + Eid + Did	=	Ide		

In Figure 5.1, we show the average frequency per token of each sensitive data type in NUBes-PHI and the MEDDOCAN corpus. It can be observed that the distribution does not follow the same trend in one corpus and the other. Most strikingly, NUBes-PHI documents do not contain mentions of locations (`Loc`) as much, nor do they include explicitly patient names (`Pat`) or contact information (`Con`), even less so, as mentioned earlier, identification numbers (`Ide`).

5.2.1.3 Data representation

The NUBes-PHI corpus comes sentence-splitted and tokenised. The labelling scheme chosen for this corpus was BIO: Beginning (`B-`), Inner (`I-`), Outside (`O`). We repeat the examples for MEDDOCAN E1 and E2 encoded with the BIO scheme and the NUBes-PHI's tagset:

E1

Domicilio	O
:	O
Avenida	B-Loc
de	I-Loc
concha	I-Loc
espina	I-Loc
16	I-Loc
,	I-Loc
2,1	I-Loc
.	O

E2

El	O
número	O
de	O
móvil	O
de	O
su	O
esposa	B-Kin
es	O
el	O
633	B-Con
349	I-Con
565	I-Con
.	O

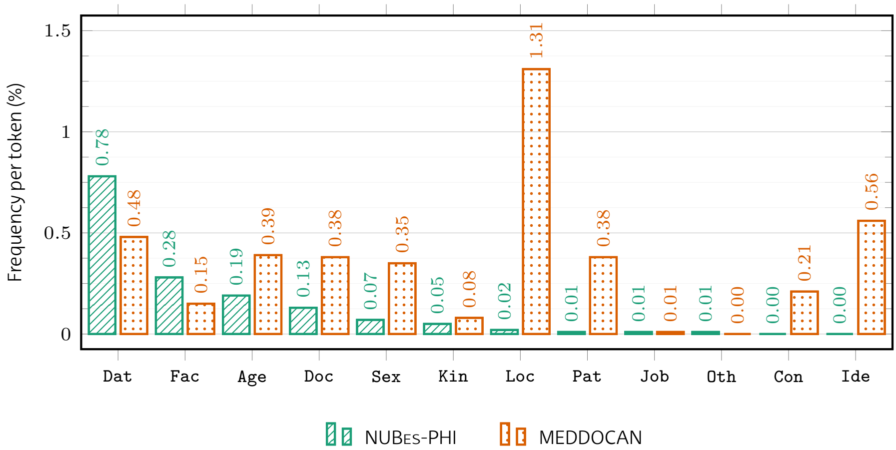


Figure 5.1: Comparison between sensitive data type frequencies in the MEDDOCAN and NUBes-PHI corpora. See data type groupings and equivalences in Table 5.3.

The features used to represent each instance for the Conditional Random Field (CRF) and NCRF++ systems is the same as that described for MEDDOCAN (See Section 4.2.1.2), with the following exceptions:

- **Brown clusters**: new clusters were computed with the training set of the NUBes-PHI corpus. The procedure and tools to compute them were the same as for the MEDDOCAN challenge.
- **Position** and **Header**: while these features made sense in MEDDOCAN due to the highly structured nature of the synthetic documents, they were not used in this chapter as they do not describe any salient characteristic of the NUBes-PHI corpus.

5.2.1.4 Output handling

As in MEDDOCAN, predicted labels were post-processed to ensure that the results were well-formed in terms of the tagging scheme, which in this case was the BIO scheme.

5.2.2 Systems

In this chapter, we train and evaluate in NUBes-PHI the same systems applied to MEDDOCAN in Chapter 4. For convenience, we list them succinctly here, and refer the reader to the corresponding section in the previous chapter for details:

- **spaCy**: spaCy's NER implementation, consisting of a transition system over Convolutional Neural Networks (CNN). Read more in Section 4.2.2.1.
- **CRF**: a Conditional Random Field (CRF) classifier, the only shallow algorithm tested. Read more in Section 4.2.2.2.
- **NCRF++**: a character CNN, followed by a word bidirectional LSTM (biLSTM) and an output CRF classifier. Read more in Section 4.2.2.3.
- **BERT**: a Multilingual BERT encoder with a token classification head on top. Read more in Section 4.2.2.4.

In addition, as the NUBes-PHI corpus is private and these are the first experiments reported with it, we also implement a **baseline** system, in order to establish the difficulty of the task in this corpus. To that end, a sensitive data recogniser and classifier has been developed that consists of regular-expressions and dictionary lookups. For each category to detect a specific method has been implemented. For instance, the `Dat`, `Age`, `Tim` and `Doc` detectors are based on regular expressions; `Fac`, `Sex`, `Kin`, `Loc`, `Pat` and `Job` are looked up in dictionaries. The dictionaries are hand-crafted from the training data available, except for `Pat`, for which the possible candidates considered are the 100 most frequent female and male names in Spain according to the National Statistics Institute [42].

5.2.3 Evaluation

In this chapter, we follow the same experimental design as that described for the MEDDOCAN challenge (Section 4.2.3). It distinguishes two scenarios:

- **Detection**: measures how well the systems are at recognising sensitive data spans. Performance is measured in terms of precision (P), recall (R) and F_1-score (F_1). There are two versions of these metrics: **merged** and **strict**.
- **Classification**: measures how well the systems are at recognising and categorising sensitive data spans. Performance is measured in terms of P, R, F_1 and leak (Lk).

A subject worth being studied is the need of labelled data. Manually labelled data is an scarce and expensive resource, which is difficult to come by for some application domains or languages. In this line, we performed two experiments sets in addition to testing the systems trained on NUBes-PHI:

First, we evaluate the systems trained on MEDDOCAN in a zero-shot fashion. It must be noted that, as explained earlier, the tagset handled by the MEDDOCAN models is different to that defined for NUBes-PHI. In order to evaluate the predictions of the MEDDOCAN models in NUBes-PHI, we post-processed the predictions applying the conversion map presented in Table 5.3.

Second, we study the dependency of each system on the available amount of training data by training all the compared models using decreasing amounts of data—from 100% of the available training instances to just 1%. The same data subsets have been used to train all the systems. Due to the knowledge transferred from the pre-trained BERT model, the BERT-based model is expected to be more robust to data scarcity than those that start their training from scratch.

5.3 Results

5.3.1 In-domain results

Table 5.4: Results of sensitive data detection and classification in the NUBes-PHI corpus. The lower section of the table reports zero-shot results (zs) of models trained in the MEDDOCAN corpus. Best and second-best results are highlighted in boldface and underlined respectively.

	Merged Detection			**Strict Detection**			**Classification**			
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**Lk**
baseline	0.441	0.308	0.363	0.427	0.301	0.353	0.414	0.292	0.342	0.174
spaCy	<u>0.923</u>	0.896	<u>0.909</u>	<u>0.921</u>	0.891	<u>0.906</u>	<u>0.910</u>	0.881	<u>0.895</u>	0.029
CRF	**0.925**	0.881	0.903	**0.922**	0.877	0.899	**0.912**	0.868	0.890	0.032
NCRF++	0.898	<u>0.912</u>	0.905	0.893	<u>0.903</u>	0.898	0.879	<u>0.889</u>	0.884	<u>0.027</u>
BERT	0.908	**0.941**	**0.924**	0.894	**0.932**	**0.913**	0.884	**0.921**	**0.902**	**0.019**
spaCy$_{zs}$	0.550	0.134	0.215	0.545	0.132	0.213	0.534	0.130	0.209	0.214
CRF$_{zs}$	0.335	0.073	0.120	0.329	0.072	0.118	0.321	0.070	0.115	0.228
NCRF++$_{zs}$	0.593	0.183	0.280	0.583	0.179	0.274	0.560	0.172	0.263	0.203
BERT$_{zs}$	0.673	0.534	0.595	0.654	0.522	0.580	0.627	0.500	0.556	0.123

Table 5.4 shows the results of the conducted experiments in NUBes-PHI for all the compared systems. The baseline system gives us insight about how challenging the data is: with simple regular expressions and gazetteers, a precision of 0.441 is obtained in the easiest evaluation scenario; the recall, which directly depends on the coverage provided by the rules and resources, is even lower—0.308. These results suggest that the task is unlikely to be solved without the generalisation capabilities of Machine Learning (ML) and Deep Learning (DL).

Regarding the models fine-tuned on NUBes-PHI, a similar behaviour to that noted in the MEDDOCAN data can be observed: BERT surpasses the rest of the systems due to the remarkable advantage of 3 recall points over the second-best model, NCRF++, across all the evaluation scenarios. Also as in MEDDOCAN, the highest precision overall is achieved by CRF. A fact worth highlighting is

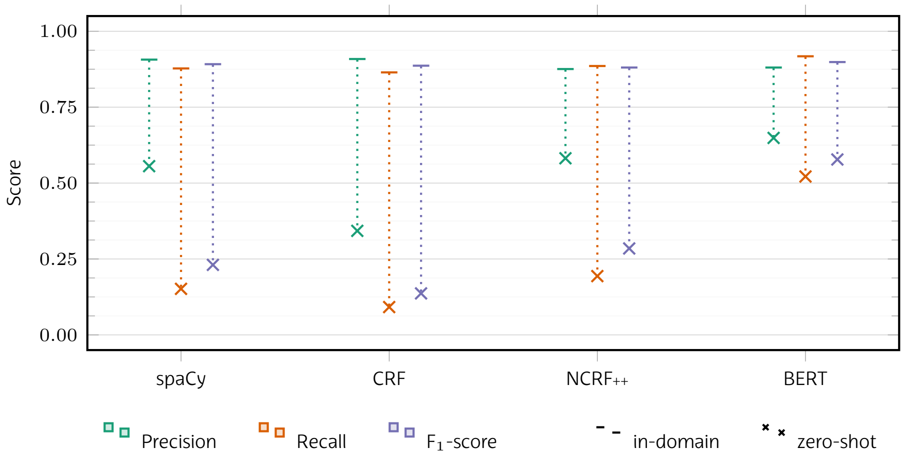


Figure 5.2: Results on the classification task of in-domain trained/fine-tuned models (upper marks) vs MEDDOCAN model zero-shot predictions (lower marks)

that, according to these results, and unlike in MEDDOCAN, BERT achieves a precision lower than the rest of the systems (i.e., it makes more false positive predictions). This, among other topics, is examined in the Error analysis section.

Interestingly, the scores are lower than those obtained in MEDDOCAN, although not so much as one would expect given the repetitiveness of MEDDOCAN and the sparsity of NUBES-PHI. Results worsen by 5 to 7 F_1-score score points across the board, with BERT achieving 0.902 in the strict classification scenario in contrast to 0.967 in MEDDOCAN (Table 4.6 in Chapter 4). These results are in line with the analysis made by Lange et al. (2019), who evaluated their MEDDOCAN systems exclusively in the MEDDOCAN document sections consisting of the actual clinical cases, a subcorpus more similar to NUBES-PHI. They report to have obtained results of ~0.900 F_1-score.

5.3.2 Zero-shot results with MEDDOCAN models

As for the zero-shot evaluations (lower part of Table 5.4), none of the systems except BERT surpass the baseline in terms of F_1-score. The CRF model struggles most of all with the change of target domain, obtaining an F_1-score of 0.120 in the easiest evaluation scenario (i.e., merged detection). As Figure 5.2 shows, the drop in performance is most marked in recall metrics: 0.073, 0.134 and 0.183 for CRF, spaCy and NCRF++, respectively in merged detection. In contrast, BERT shows a recall of 0.534, evidencing once more its greater generalisation capabilities.

5.3.3 Training curves

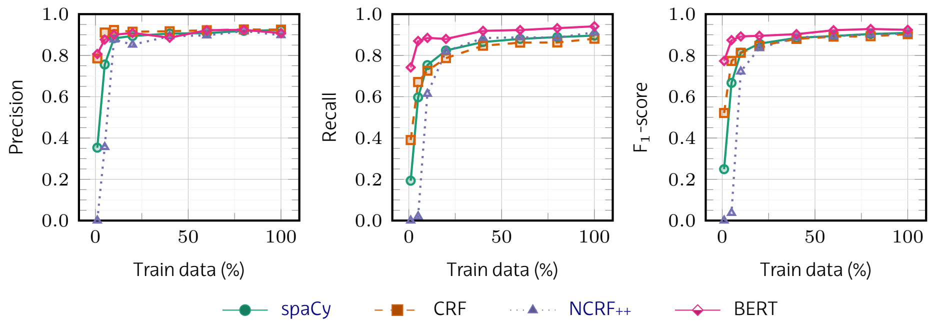


Figure 5.3: Performance curves with increasing amounts of training data on the sensitive span detection task in the NUBes-PHI corpus

Figure 5.3 shows the impact of decreasing the amount of training data in the (merged) detection scenario. It shows the difference in precision, recall, and F_1-score with respect to that obtained using 100% of the training data. A general downward trend can be observed, as one would expect: less training data leads to less accurate predictions. However, as expected, BERT is the most robust to the reduction of training data, showing a steadily low performance loss. With only 1% of the dataset (i.e., 230 training instances), it only suffers a striking 15-point F_1-score loss, in contrast to the 65, 38 and 90 points lost by the spaCy, CRF and NCRF++ models, respectively. This steep performance drop stems to a larger extent from recall decline, which is not that marked in the case of BERT. Admittedly, the hyperparameters of spaCy and NCRF++ have not been adapted to each subset size, but neither have they been in the case of BERT.

5.3.4 Error analysis

We next focus on the models with best and worst recall, namely, BERT and CRF. Their confusion matrices in the classification scenario are shown in Tables 5.5 and 5.6 respectively (see the confusion matrices of spaCy and NCRF++ in Appendix D). As can be seen, the fine-tuned BERT (Table 5.6b) has less difficulty in predicting correctly less frequent categories, such as `Loc`, `Job`, and `Pat`. One of the most common mistakes according to the confusion matrices is classifying hospital names as location (`Loc`) instead of the more accurate hospital (`Hos`); this is hardly a harmful error, given that a hospital is actually a location. Last, the category other (`Oth`) is completely leaked by all the compared systems, most likely due to its almost total lack of support in the training dataset.

Table 5.5: Confusion matrices of CRF for the classification task on NUBes-PHI. The matrices have been computed with token-level predictions without taking the BIO tags into account.

(a) Zero-shot (model trained on the MEDDOCAN corpus)

														predicted
		N	Dat	Fac	Age	Tim	Doc	Sex	Kin	Loc	Pat	Job	Oth	O
	Dat	1,479	17.38	00.00	00.54	00.00	00.00	00.00	00.00	00.27	00.20	00.00	00.00	81.61
	Fac	557	00.00	08.62	00.00	00.00	05.03	00.00	00.00	00.72	00.36	00.00	00.00	85.28
	Age	574	00.00	00.00	48.43	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	51.57
	Tim	407	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.25	00.00	00.00	00.00	99.75
	Doc	401	00.00	00.25	00.00	00.00	07.48	00.00	00.00	01.25	00.75	00.00	00.00	90.27
	Sex	71	00.00	00.00	00.00	00.00	00.00	66.20	00.00	00.00	00.00	00.00	00.00	33.80
	Kin	44	00.00	00.00	00.00	00.00	04.55	00.00	31.82	00.00	00.00	00.00	00.00	63.64
	Loc	26	00.00	00.00	00.00	00.00	03.85	00.00	00.00	23.08	00.00	00.00	00.00	73.08
	Pat	14	00.00	00.00	00.00	00.00	28.57	00.00	00.00	00.00	00.00	00.00	00.00	71.43
	Job	17	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
	Oth	1	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
true	O	103K	00.01	00.00	00.00	00.00	00.05	00.00	00.04	00.41	00.03	00.00	00.00	99.43

(b) Model trained on NUBes-PHI

														predicted
		N	Dat	Fac	Age	Tim	Doc	Sex	Kin	Loc	Pat	Job	Oth	O
	Dat	1,479	91.14	00.00	01.35	00.68	00.14	00.00	00.00	00.07	00.00	00.00	00.00	06.63
	Fac	557	00.00	88.51	00.00	00.00	00.18	00.00	00.00	00.54	00.00	00.00	00.00	10.77
	Age	574	00.00	00.00	96.34	00.35	00.00	00.00	00.00	00.00	00.00	00.00	00.00	03.31
	Tim	407	00.98	00.00	00.00	94.10	00.00	00.00	00.00	00.00	00.00	00.00	00.00	04.91
	Doc	401	00.00	00.75	00.00	00.00	94.76	00.00	00.00	00.00	00.00	00.00	00.00	04.49
	Sex	71	00.00	00.00	00.00	00.00	00.00	100	00.00	00.00	00.00	00.00	00.00	00.00
	Kin	44	00.00	00.00	00.00	00.00	00.00	00.00	93.18	00.00	00.00	00.00	00.00	06.82
	Loc	26	00.00	19.23	00.00	00.00	00.00	00.00	00.00	30.77	00.00	00.00	00.00	50.00
	Pat	14	00.00	00.00	00.00	00.00	21.43	00.00	00.00	00.00	42.86	00.00	00.00	35.71
	Job	17	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	11.76	00.00	88.24
	Oth	1	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
true	O	103K	00.06	00.03	00.00	00.01	00.00	00.00	00.01	00.00	00.00	00.00	00.00	99.88

Table 5.6: Confusion matrices of BERT for the classification task on NUBES-PHI. The matrices have been computed with token-level predictions without taking the BIO tags into account.

(a) Zero-shot (model fine-tuned on the MEDDOCAN corpus)

													predicted
		N	Dat	Fac	Age	Tim	Doc	Sex	Kin	Loc	Pat	Job	Oth	O
	Dat	1,479	85.94	00.00	00.81	00.00	00.00	00.00	00.00	00.34	00.00	00.00	00.00	12.91
	Fac	557	00.00	40.57	00.00	00.00	00.00	00.00	00.00	02.69	00.00	00.00	00.00	56.73
	Age	574	00.00	00.00	64.29	00.00	00.17	00.00	00.00	00.17	00.00	00.00	00.35	35.02
	Tim	407	08.11	00.00	00.00	00.00	00.00	00.00	00.00	00.25	00.00	00.00	00.00	91.65
	Doc	401	00.00	00.50	00.00	00.00	04.49	00.00	00.50	01.75	12.47	00.00	00.00	80.30
	Sex	71	00.00	00.00	00.00	00.00	00.00	100	00.00	00.00	00.00	00.00	00.00	00.00
	Kin	44	00.00	00.00	00.00	00.00	00.00	04.55	77.27	00.00	00.00	00.00	00.00	18.18
	Loc	26	00.00	00.00	00.00	00.00	00.00	00.00	00.00	57.69	00.00	00.00	00.00	42.31
	Pat	14	00.00	00.00	00.00	00.00	14.29	00.00	35.71	00.00	42.86	00.00	00.00	07.14
	Job	17	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
	Oth	1	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
true	O	103K	00.06	00.02	00.01	00.00	00.00	00.01	00.09	00.02	00.02	00.00	00.00	99.76

(b) Model fine-tuned on NUBES-PHI

													predicted
		N	Dat	Fac	Age	Tim	Doc	Sex	Kin	Loc	Pat	Job	Oth	O
	Dat	1,479	95.61	00.00	01.35	00.81	00.14	00.00	00.00	00.00	00.00	00.00	00.00	02.10
	Fac	557	00.00	96.05	00.00	00.00	00.54	00.00	00.00	00.72	00.00	00.00	00.00	02.69
	Age	574	00.00	00.00	99.30	00.35	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.35
	Tim	407	00.74	00.00	00.00	98.03	00.00	00.00	00.00	00.00	00.00	00.00	00.00	01.23
	Doc	401	00.00	00.50	00.00	00.00	99.25	00.00	00.00	00.00	00.00	00.00	00.00	00.25
	Sex	71	00.00	00.00	00.00	00.00	00.00	100	00.00	00.00	00.00	00.00	00.00	00.00
	Kin	44	00.00	00.00	00.00	00.00	00.00	00.00	97.73	00.00	00.00	00.00	00.00	02.27
	Loc	26	00.00	23.08	00.00	00.00	00.00	00.00	00.00	50.00	00.00	00.00	00.00	26.92
	Pat	14	00.00	00.00	00.00	00.00	07.14	00.00	00.00	00.00	85.71	00.00	00.00	07.14
	Job	17	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	29.41	00.00	70.59
	Oth	1	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
true	O	103K	00.06	00.05	00.00	00.02	00.01	00.01	00.01	00.00	00.00	00.00	00.00	99.83

Upon manual inspection of the errors committed by our BERT-based model, we discovered that it has a slight tendency towards producing ill-formed BIO sequences, as in Examples E3 and E4 (page 74; incorrect predictions are marked with an asterisk and separated from the true label with a backslash). We could expect that complementing the BERT-based model with a CRF layer on top would help enforce the emission of valid sequences, alleviating this kind of errors and further improving its results. Yet, as mentioned in the previous chapter, a BERT-based system with CRF (Mao et al., 2019) fell behind our simpler BERT implementation in the MEDDOCAN challenge.

E3

Acudirá	O
a	O
la	B-
Clínica	*B- / I-
Marseille	I-

E4

control	O
(	O
15	B-
y	O
22	*I- / B-
de	I-
junio	I-
)	O

Finally, the manual error analysis also uncovered several human errors (HE) in the annotation of NUBes-PHI, which contributed falsely towards the few false positive (FP) errors committed by the systems. As mentioned earlier, the goal of having created the NUBes-PHI corpus in the first place was to be able to publish the NUBes corpus by substituting the detected sensitive information with fake data (more on this topic in Chapter 10). Thus, we processed the whole NUBes-PHI corpus with our models, including the training and development partitions, merged the alleged FP predictions of the systems—except NCRF++, which had not been trained at this point—and reviewed them one by one in search of HEs, so as to minimise potential leaks of sensitive data in the published corpus. The result of this analysis is shown in Table 5.7.

This process helped us detect 141 sensitive data items overlooked in the original human annotation, which make 1.8% of the total sensitive data items annotated and substituted in the final version of NUBes. As can be seen, the BERT and spaCy models were most helpful in this regard, who together detected 137 of the 141 HEs—although BERT committed most true FP errors as well. Of the 141 HEs, 39%, 20% and 15% were date, healthcare facility and time mentions, respectively. The remainder ∼25% belonged to the less frequent categories.

5.4 Conclusions

In this chapter, we extend the work carried out for the Medical Document Anonymization (MEDDOCAN) challenge, described in Chapter 4. We concluded

Table 5.7: Alleged false positive (FP) errors and uncovered human errors (HE) after their revision

Predicted by					
BERT	**spaCy**	**CRF**	**Alleged FP**	**of which HE**	**False FP %**
✓			171	55	32
	✓		59	21	36
		✓	11	0	0
Total by any 1 system			*241*	*76*	*32*
✓	✓		43	33	77
✓		✓	3	3	100
	✓	✓	3	1	33
Total by any 2 systems			*49*	*37*	*76*
✓	✓	✓	40	28	70
Total by all 3 systems			*40*	*28*	*70*
Total			*330*	*141*	*43*

that chapter by raising the concern that, although MEDDOCAN seemed to be solved in practice, it was sound to suspect that the excellent results achieved by our systems and the competitors might be somewhat distorted by the repetitiveness of the synthetic corpus. Thus, this chapter has replicated the experimental setup of MEDDOCAN in a corpus of real health records.

We showed that, overall, the results worsen 5 to 7 F_1-score point across the board in comparison to the MEDDOCAN evaluation. Other than that, the results show a similar trend to that identified in the MEDDOCAN challenge: the BERT-based model outperforms the other systems without requiring any adaptation or domain-specific feature engineering, just by being trained on the provided labelled data. Interestingly, this model obtains a remarkably higher recall than the other systems. High recall is a desirable outcome because, when anonymising sensitive documents, accidentally leaking sensitive data is likely to be more dangerous than over-obfuscating non-sensitive text.

Further, we have conducted an additional experiment on this dataset by progressively reducing the training data for all the compared systems. The BERT-based model shows the highest robustness to training-data scarcity, losing only 15 points of F1-score when trained on 230 instances instead of 21,371. These results indicate that the transfer-learning achieved through the pre-trained Multilingual BERT model not only helps obtain better results, but also lowers the need of manually labelled data for this application domain. These observations are in line with the literature that uses BERT for other tasks.

Another experiment set consisted of zero-shot evaluations of the MEDDOCAN models in the NUBES-PHI corpus. Here as well, BERT proved to be superior with a recall of 0.534 in the detection scenario—the second-best recall in the same scenario was 0.183 by NCRF++.

Although a recall of 0.534 is far from being applicable in production scenarios, this is not to say that the MEDDOCAN corpus may not be found beneficial when exploited in other setups than that described here. To begin with, we have shown that NUBes-PHI and the MEDDOCAN corpus differ so much that they could even be considered to constitute different domains. And whereas NUBes-PHI is not a synthetic corpus, unlike MEDDOCAN, it cannot be considered the true representative of the average EHR document in Spain either. In fact, Pérez-Díez et al. (2021) describe a corpus of radiology reports whose documents look much more alike those in the MEDDOCAN corpus than NUBes-PHI. Further, it might be the case that exploiting MEDDOCAN alongside NUBes-PHI helps improve the reported results. We leave these experiments as future work.

Finally, the models trained for these experiments served to detect errors in the original human annotation of NUBes-PHI. After manually reviewing a set of alleged 330 false positive errors, 141 turned out to be correct detections of sensitive data. 137 of these human errors were contributed by the BERT and/or spaCy models. The final, corrected version of NUBes-PHI is the basis for the NUBes corpus, the collection of health records manually annotated with negation and uncertainty that is presented in Chapter 10.

PART III

TERM IDENTIFICATION

Chapter 6

Term identification: background and literature review

6.1 Definition and motivation

Given the vast amount of text data that is produced on a daily basis both in the academia and every health care centre worldwide, biomedical Information Extraction (IE) has become increasingly relevant to the Natural Language Processing (NLP) community in recent years, as it can help lighten the burden of researchers and clinicians alike by facilitating the discovery and usage of biomedical knowledge.

Biomedical **term identification** (also known as "term normalisation", "term disambiguation", "term linking", or "semantic annotation", to name a few) is an essential step in the automatic extraction of this valuable knowledge: recognising key terms mentioned in texts and linking them to the entry in an ontology or controlled vocabulary that represents the concept denoted by the term. Figures 6.1 and 6.2 illustrate the task in Spanish and English, respectively.

Each coloured span is a recognised biomedical term. In traditional terminology, a **term** is an expression that has a particular meaning in a language for specific purposes. For instance, clinical terms are expressions that denote disorders, clinical procedures, symptoms, body structures, and so on. The category of each term is given by the background colour in the figure: disorders in red, living beings in green, medical procedures in yellow, chemicals and drugs in blue, and physiological processes in orange. Finally, an example of term identification is given for the term "Aztreonam": it denotes the concept C0004521 [43] in the Unified Medical Language System (UMLS) Metathesaurus (Lindberg et al., 1993), a large biomedical terminological resource.

Term identification may be addressed end-to-end, i.e., by jointly recognising and identifying terms, or may be applied in already recognised terms as a downstream step—in which case the task is more likely to be called "term disambiguation".

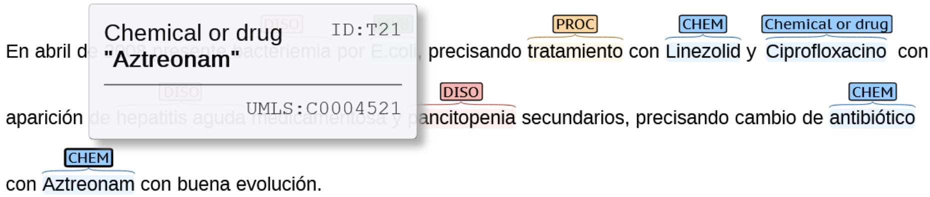


Figure 6.1: Example of term identification with UMLS in Spanish text (see translation in Figure 6.2; visualisation rendered with brat [Stenetorp et al., 2012])

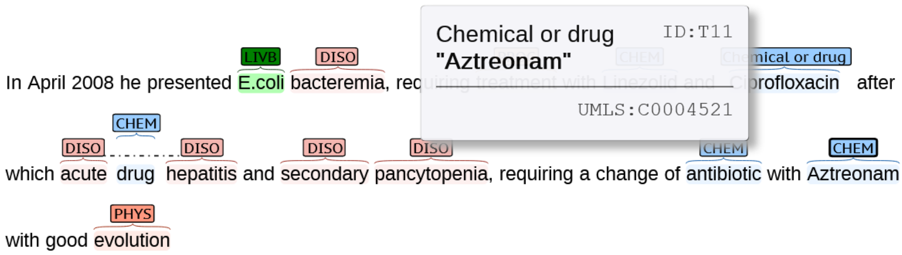


Figure 6.2: Example of term identification with UMLS in English text

6.2 Related resources

The UMLS Metathesaurus (Lindberg et al., 1993), created and maintained quarterly by the U.S. National Library of Medicine (NLM), brings together biomedical vocabulary sources or terminologies of different languages. The entries in the vocabularies are arranged by concept or meaning. It maps one terminology to another, in addition to keeping the original relations stated in the source terminologies themselves. Thus, the Metathesaurus can be viewed as a comprehensive thesaurus or ontology of biomedical concepts. Each concept is categorised into one or more of the 133 **semantic types** of the UMLS Semantic Network (McCray et al., 1995). These types, in turn, are aggregated into 15 broader **semantic groups** (McCray et al., 2001).

The next chapters exploit the 2016AA Full Release Metathesaurus [44] as the reference knowledge base to perform term identification. This release gathers 196 terminology sources in 25 different languages, amounting to 3,250,226 concepts and 10,586,865 terms in total. The great bulk of concepts are provided by three English terminologies and their translations to Spanish: Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), the Medical Subject Headings®

(MeSH), and the Medical Dictionary of Regulatory Activities (MedDRA). The complete English subset covers almost the complete Metathesaurus (3,250,158 concepts); in contrast, the Spanish subset, while being the second largest subset, accounts only for 14% of the Metathesaurus concepts (451,296).

Currently, there exist 3 public corpora of texts in Spanish that are annotated with UMLS concepts:

Mantra GSC (Kors et al., 2015) The Mantra Gold Standard Corpus (Mantra GSC) is a collection of parallel biomedical corpora in English, French, German, Spanish, and Dutch that has been manually annotated with concepts of the UMLS Metathesaurus. The Spanish portion consists of 100 scientific publication titles and 100 drug labels, for a total of 639 manually identified terms. This corpus is the basis for the experiments of Chapter 8.

CT-EBM-SP The Clinical Trials for Evidence-Based Medicine in Spanish corpus (Campillos-Llanos et al., 2021) is a collection of 1,200 texts about clinical trials annotated with entities from certain UMLS semantic groups. Further, out of the 46,698 annotated entities, at least 33,391 are manually identified with one or more UMLS concepts. In total, this corpus contains annotations for ~5,000 unique UMLS concepts, which makes it at the moment the biggest of its kind for the Spanish language. It is publicly available online [20].

E3C (Magnini et al., 2021a,b) The European Clinical Case Corpus (E3C) is a collection of clinical cases in 5 languages, namely, Italian, English, French, Spanish, and Basque. Among other information, this corpus contains annotations of disorders, which have been identified with a UMLS concept following the ShARe annotation guidelines Elhadad et al. (2012). The Spanish portion of the corpus consists of 1,400 clinical cases, annotated with 2,582 identified disorders (938 unique). It is publicly available at the European Language Grid catalogue [45].

6.3 State of the Art

The automatic identification of biomedical terminology in scientific texts is an active research area but most of the recent works are targeted at the English language. This is due, in part, to the greater availability of biomedical resources—such as scientific articles, vocabularies and ontologies—in English. In this scenario, MetaMap (Aronson, 2001, 2006), cTakes (Savova et al., 2010) and NCBO Annotator (Dai et al., 2008) are well-known tools for the semantic annotation of biomedical text. Metamap is probably the better-known tool. It is "knowledge intensive" as it relies heavily on the SPECIALIST Lexicon, a large syntactic lexicon of biomedical and general English. cTakes recognises biomedical concepts in texts and relates them to their UMLS concept. And the NCBO Annotator, developed by the National Center for Biomedical Ontology (NCBO), is a web service that provides links between the text of biomedical literature and the knowledge embedded in the BioPortal ontologies and the UMLS Metathesaurus.

In the last years, new works have emerged to face this challenging task, allowing the advance of the state of the art. Nunes et al. (2013) developed BeCAS, a biomedical concept annotation system, which uses dictionary-matching techniques to recognise diverse types of concepts (including species, anatomical concepts, microRNAs, enzymes, chemicals, drugs, diseases, metabolic pathways, cellular components, biological processes and molecular functions) from multiple sources, including UMLS, NCBI BioSystems (Geer et al., 2010), LexEBI (Rebholz-Schuhmann et al., 2013b), ChEBI (Hastings et al., 2016), miRBase (Griffiths-Jones, 2004) and the Gene Ontology (Gene Ontology Consortium, 2004). It provides a web API for biomedical concept identification.

NOBLE Coder (Tseytlin et al., 2016) is another open-source system for biomedical text annotation in English. It can be configured through a graphical interface to work with different vocabularies, even with customised terminologies, allowing to select one or more branches of a set of vocabularies and/or filtering vocabularies by semantic types.

Recently, Soysal et al. (2017) implemented CLAMP, a pipeline composed of multiple modules for the analysis and the extraction of information contained in clinical text. It includes a named entity recogniser to detect biomedical terminology. Then, an UMLS encoder links each term with the corresponding concept in the UMLS Metathesaurus.

In the case of non-English biomedical text, term identification becomes even more difficult mainly by a shortage of biomedical resources. In this scenario, we present the most relevant works for the Spanish language. Carrero et al. (2008a,b) presented one of the first works in using a combination of automatic translation and a term identifier for English (MetaMap) in order to annotate biomedical entities in Spanish texts with their corresponding UMLS concepts.

Later, Castro et al. (2010) developed an automatic system for the recognition of SNOMED CT concepts by computing a similarity function between sentences in clinical notes and then term normalisation is based on the results obtained by querying an Apache Lucene [46] index of SNOMED CT and re-ranking the candidates with a function of their own. They obtained an average F_1-score of 0.11 on their own corpus of 100 manually annotated documents. Furthermore, Berlanga et al. (2010) introduced the notion of *concept retrieval*, which was based on applying information retrieval methods in order to obtain UMLS concepts relevant to a text and later use them to properly annotate matching text spans.

The systems developed in the context of the 2013 CLEF-ER challenge for biomedical entity recognition in parallel multilingual corpora (Rebholz-Schuhmann et al., 2013a) provide some of the first prototypes for the annotation of biomedical texts in languages other than English. Among the participating systems there were some targeted at Spanish including the ones proposed by Attardi et al. (2013) and Bodnari et al. (2013), which exploited word alignment information by statistical translation and parallel corpus, respectively, in order to transfer annotations from English to Spanish. Specifically, Attardi et al. (2013) translated an English corpus with biomedical entity annotations to Spanish, including the transfer of annotations. Then, a Named Entity Recognition (NER) module was trained in the translated Spanish corpus in order to recognise biomedical entities in unseen Spanish text. Otherwise, Bodnari et al. (2013) manually annotated biomedical entities in English text from a parallel corpus and were transferred to Spanish (and French) text in order to train a NER for each language. These works were not evaluated against a golden corpus.

In the same year, Oronoz et al. (2013) presented FreelingMed, an extension of the Freeling Spanish analyser (Carreras et al., 2004) to recognise biomedical entities extracted from available knowledge resources (lists of medical abbreviations and drug names, as well as the SNOMED CT thesaurus). Oronoz et al. (2013) evaluated their proposal with their own corpus of medical reports annotated by health professionals with diseases, medications and other substances, obtaining 0.90 F1 score with approximate boundary matching for the term recognition task.

More recently, Roller et al. (2018) presented a sequential cross-lingual candidate search for biomedical term normalisation. The main component of their approach is a character-based neural translation model trained on UMLS for multiple languages, such as Spanish, French, Dutch and German. Roller et al. (2018) achieved an F_1-score of 0.69 on the task of normalisation of oracle terms in the Spanish Medline sub-corpus of the Mantra GSC. Slightly better results were just obtained by Yuan et al. (2022) in the same task with CODER, a more intricate system that exploits cross-lingual term and graph embeddings. It must be noted, however, that these works have oracle terms as starting point.

To this day, biomedical semantic annotation in non-English text is still one of the most challenging research topics in biomedical NLP. In this work, we describe a system for term recognition and identification based on the UMLS (Chapter 7) that does not require supervision, and which we evaluate exhaustively against the Mantra GSC (Chapter 8).

Chapter 7

Term identification: the UMLSmapper prototype

7.1 Introduction

This chapter presents UMLSmapper, a lexically motivated module that performs term recognition and normalisation with the UMLS Metathesaurus.

In contrast to most other chapters in this work, this one is purely a description of a system from a technological perspective. The performance of the system is evaluated in Chapter 8 (next). The remainder of this chapter is structured as follows: Section 7.2 offers an overview of the system, with an account of its general workflow and key implementation details; it also discusses briefly the limitations of the proposed approach. Section 7.3 describes the terminology and knowledge resources exploited by the system. Section 7.4 explains each technological module of the system individually. Finally, Section 7.5 concludes the chapter.

7.2 System overview

7.2.1 Implementation details

The entire program has been written in Java 8 and packaged as a Docker image. It deploys various third-party libraries and tools, among which we must highlight:

- Apache Lucene™ [46] for fast consultation of the UMLS Metathesaurus.
- IXA-Pipes (Agerri et al., 2014) [47], a linguistic analysis toolkit. It is deployed as a Docker container with which UMLSmapper interacts via HTTP.
- UKB (Agirre et al., 2009) [48], a collection of programs to perform unsupervised word sense disambiguation based on a given knowledge base. It is deployed as a TCP server.

UMLSmapper itself is run as a REST web service, with which clients interact through Hypertext Transfer Protocol (HTTP) requests. In the most common use case, it receives plain text and returns a JSON file with morpho-syntactic information and the result of the normalisation attempt.

The core engine of the program can be described as a pipeline of modules, each responsible for a logical step of the process. Further, UMLSmapper is in all respects a lexically/knowledge-driven solution; it relies heavily on several terminological resources, mostly derived from the UMLS Metathesaurus, without which UMLSmapper is but an empty shell. At the same time, UMLSmapper may in principle work for any language well-enough represented in the Metathesaurus, as long as basic NLP tools (i.e., tokenisation, PoS) exist for that language. At the moment of writing this work, UMLSmapper has been tested and used in Spanish and English texts. These tests are the subject of Chapter 8.

The program is highly configurable, as will become evident throughout the following sections. It can be configured globally in a PROPERTIES file, and it also accepts one-time settings with each request to the web service. Said web service's public API is described in detail in the online documentation [5]. Figure 7.1 illustrates the general architecture of the program. Next, we review briefly the general workflow of the core pipeline, and then examine each resource and module individually.

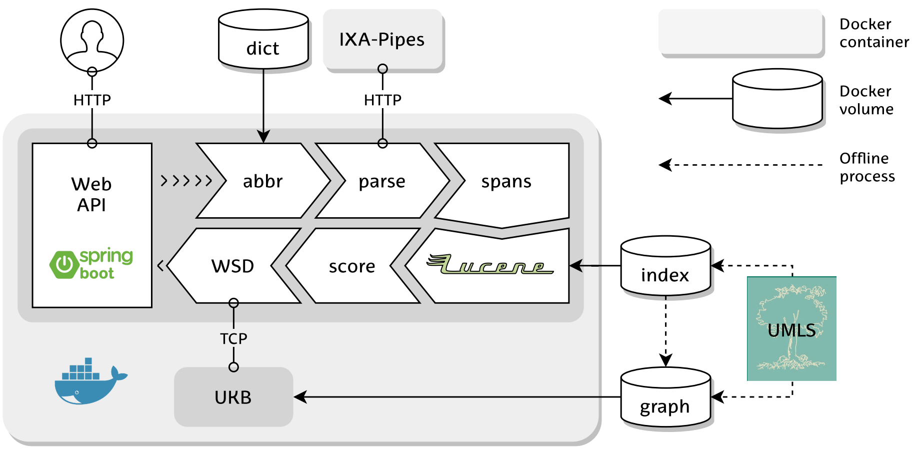


Figure 7.1: Diagram of UMLSmapper's components and its key dependencies

7.2.2 General workflow

First, the text received may be analysed in search of abbreviations, acronyms and initialisms, which are expanded to their corresponding full expressions. Next, the system carries out low-level linguistic processing of the expanded text: tokenisation, Part of Speech (PoS) tagging, and, depending on the configuration chosen, constituent parsing. The linguistic information obtained serves then as basis to generate text spans or sequences of tokens candidate for being mapped to a medical concept.

Alternatively, the user of UMLSmapper may choose to perform these steps with third-party tools and provide to the program a text already analysed and marked for the spans to be mapped (e.g., with a medical NER tool) in the format required by UMLSmapper (see the online documentation [5]).

After, the system makes per given span an initial suggestion of links with UMLS Metathesaurus. It does so using Apache Lucene™ to retrieve UMLS lexicalisations similar to the spans. Next, the retrieved links are ranked according to a certain scoring function, and a threshold is applied to discard candidates with too low a score. Finally, the match candidate with highest score is chosen as final link for each span, if any candidate still remains. It is possible that several candidates obtain top scores; these cases may be resolved by Word Sense Disambiguation (WSD) or other simpler strategies.

Of note, not all the suggested spans are processed; UMLSmapper arranges the spans in descending order of length, and does not evaluate a given span if another, longer span that subsumes it produced a good-enough link. For example, the span 'extremidades' (*extremities*) would not be processed if 'extremidades inferiores' (*lower extremities*) were already linked. Following this logic, spans that overlap can be annotated, but not spans that are nested within another.

7.2.3 Limitations

UMLSmapper's main selling points—namely, that it works virtually out-of-the-box with no need of annotated data and that it adapts easily to specific biomedical domains—are facilitated by the simplicity of its approach, i.e., the lexically motivated search of terms over a vast terminology source that is the UMLS Metathesaurus. Naturally, this simple approach imposes at the same time several limitations to what UMLSmapper can and cannot do.

On the one hand, UMLSmapper will never generate a link between a text span and a Concept Unique Identifier (CUI) if none of the lexicalisations of the latter are similar in form to the text span. That is, UMLSmapper's strategy for tackling **synonymy** or lexical variability is almost completely limited to relying on the coverage provided by the Metathesaurus. This limitation may lead to false negative errors.

On the other hand, UMLSmapper will always generate a link between a text span and a CUI if any of the lexicalisations of the latter are similar enough in form to the text span, regardless of semantics. That is, UMLSmapper does not analyse the meaning of the text spans in context, so if it makes a lexical match, the link is taken as valid. The strategy for tackling **polysemy** is reliant on the Metathesaurus and WSD techniques, but there is no policy in place for the cases when the specific, intended meaning of a text span is not captured by any CUI at all. This limitation may lead to false positive errors.

These limitations will be discussed in the error analysis of Chapter 8.

7.3 Resources

UMLSmapper exploits two big terminological and knowledge resources that must be prepared as a configuration step prior to using the tool. The key resource is an index of the concepts to map and their possible lexicalisations. In addition, UMLSmapper needs a graph that describes the relations between the concepts in the index. Optionally, UMLSmapper may exploit a third resource, consisting of a dictionary of abbreviations, acronyms and initialism, and their corresponding long forms. Next, we describe each of these resources in detail.

7.3.1 Metathesaurus index

The UMLS Metathesaurus is indexed with Apache Lucene™ in order to be able to produce subset views of the Metathesaurus according to convenient criteria (e.g., language, terminology source, semantic types, and so on) and, most importantly, to make time-efficient fuzzy queries of lexicalisations.

The index is derived from the Metathesaurus—or parts of it, as needed—in Rich Release Format (RRF) format; specifically, we use the information contained in the files MRCONSO and MRSTY [49]. From the given input, the program filters out the lexicalisations that do not meet the following criteria:

a) the lexicalisation comes from the terminology Logical Observation Identifiers Names and Codes (LOINC),
b) it is longer than 15 tokens,
c) it consists of a single character,
d) it consists of just numbers, or
e) it consists of only stopwords.

Then, each remaining MRCONSO entry is converted to a Lucene document with the structure described in Table 7.1. Each entry in the index associates a lexicalisation to its concept, vocabulary source, and semantic type, among others. A normalised version of the original lexicalisation is also indexed. Normalisation consists in removing spurious parenthetical material, undoing transpositions, and erasing stopwords. These changes are illustrated in Examples E1 (original lexicalisation) to E4 (final normalised lexicalisation):

E1 en blanco, cara que mira fijo durante sonambulismo (hallazgo)
blank, staring face whilst sleep walking (finding)

E2 en blanco, cara que mira fijo durante sonambulismo
blank, staring face whilst sleep walking

E3 cara que mira fijo durante sonambulismo en blanco
staring blank face whilst sleep walking

E4 cara mira fijo sonambulismo blanco
staring blank face sleep walking

Table 7.1: Apache Lucene document schema for UMLSmapper

Field	Description	Example
`cui`	Concept Unique Identifier (CUI)	`C0424280`
`lat`	Language of the lexicalisation	`SPA`
`sab`	Abbreviated name of the source	`SCTSPA`
`suppress`	Whether the lexicalisation is suppressible due to "ambiguity in meaning or lack of face validity" [50] (`O` = obsolete)	`O`
`str`	Lexicalisation of the concept	`en blanco, cara que mira fijo durante sonambulismo (hallazgo)`
`strnorm`	Normalised lexicalisation	`cara mira fijo sonambulismo blanco`
`sty`	Abbreviated name of the semantic type	`fndg`
`stypath`	Path in the semantic type tree from root—entity (`enty`) or event (`evnt`)—to `sty`	`/enty/cnce/fndg`

At runtime, the index is queried with the normalised versions of the phrases extracted from the input text and returns entries with lexicalisations similar to those phrases. Each entry retrieved is a candidate concept mapping for the corresponding trigger phrase. This process is described in depth in Section 7.4.4.

7.3.2 UKB graph and dictionary

UKB is a collection of programs to perform unsupervised WSD based on a given knowledge base in the form of a graph, where the vertices are concepts, and the edges are relations between those concepts. In turn, each concept is associated with one or more lexicalisation through a so-called dictionary.

The UKB graph and dictionary for UMLSmapper are constructed from the aforementioned Lucene index and the Metathesaurus' MRREL file [49]. This file describes relationships between concepts in the Metathesaurus. In general, they connect closely related concepts, such as those that "share some common property or are related by definition". Table 7.2 quantifies and illustrates the different relationship types included in the Metathesaurus.

The UKB graph constructed from MRREL includes all the relations that have as origin and target concepts included in our UMLS index. For each relation, we indicate the source CUI, target CUI, direction, and type of the relation.

Table 7.2: Frequency and examples of relationships in MRREL.RRF (release 2016AA)

Label	Description	Example	Frequency
SIB	is sibling of	fisioterapeuta SIB masajista	29,035,314
RO	is related to (not synonym)	ventriculograma RO ventrículo	17,833,705
SY	is synonym of	dermatitis SY sarpullido	5,648,988
PAR	is hypernym of	tegumento PAR uñas	5,320,020
CHD	is hyponym of	sinovitis CHD artropatia	5,320,020
RQ	is related to (maybe synonym)	vómitos RQ diaforesis	2,412,372
RN	is closely related to	vegetarianismo RN régimen	1,866,725
RB	is broadly related to	soledad RB nostalgia	1,866,725
QB	can be qualified by	fatiga QB estabilizado	610,433
AQ	is allowed qualifier of	mejorado AQ ansiedad	610,433
RL	is similar or "alike"	discromia RL vitíligo	62,672

7.3.3 Dictionary of short forms

An optional input preprocessing step UMLSmapper performs is the detection and resolution of abbreviated forms. At the moment, the process of resolution consists simply in looking up the detected short form in a dictionary, where each short form is associated to its long form only if the short form is typically unambiguous in the medical field. This dictionary was curated by Montoya (2017) from Yetano Laguna et al. (2003) and the manual annotation of health records in Spanish by several physicians, for a total of 2,312 short-long form entries. A sample of the dictionary is shown in Table 7.3.

7.4 Modules

As illustrated earlier, UMLSmapper consists of a set of technological modules, some of which are optional, and that are executed in a pipeline fashion. In this section, we explain what each module does and how, their inputs and outputs, and available configuration options.

Table 7.3: Most frequent unambiguous short forms collected by Montoya (2017)

Short form	Long form (es)	Long form (en)
Rx	radiografía	radiography
TAC	tomografía axial computarizada	computed tomography scan
AC	auscultación cardiaca	cardiac auscultation
x´	por minuto	per minute
mmHg	milímetros de mercurio	millimetre of mercury
mm	milímetro	millimetre
EEII	extremidades inferiores	lower limbs
TA	tensión arterial	blood pressure
ECG	electrocardiograma	electrocardiogram
O2	oxígeno	oxygen

7.4.1 Abbreviation and acronym handling

Input	User provided plain text
Output	Same text after short forms substitution
Options	• Strategy to detect short forms: rules, a classifier or none (i.e., skip this step)

The processing starts with an optional step: abbreviation and acronym recognition and resolution. UMLSmapper comes with two strategies to detect short forms: a rule-based algorithm or a Random Forest classifier. The latter (Cuadros et al., 2018) was learned from the training and development sets provided at the 2nd Edition of the Biomedical Abbreviation Recognition and Resolution Workshop (Intxaurrondo et al., 2018). As explained before (in Section 7.3.3), the resolution step consists in looking up the detected short forms in a dictionary of short forms and corresponding expansions. For example, given the following input text:

E5 Refiere dolor intermtente en EEII (sic)
[The patient] complains of intermittent pain in LEs

the output of this module is:

E6 Refiere dolor intermtente en extremidades inferiores
[The patient] complains of intermittent pain in lower extremities

7.4.2 Basic linguistic analysis

Input	Plain text
Output	Segmentation and morpho-syntactic information
Options	• Language of the input text: Spanish (es) or English (en)

```xml
<?xml version="1.0" encoding="UTF-8"?>
<NAF xml:lang="es" version="v1.naf">
  <nafHeader>
    ...
  </nafHeader>
  <text>
    <wf id="w1" offset="0" length="7" sent="1" para="1">Refiere</wf>
    <wf id="w2" offset="8" length="5" sent="1" para="1">dolor</wf>
    <wf id="w3" offset="14" length="11" sent="1" para="1">intermtente</wf>
    <wf id="w4" offset="26" length="2" sent="1" para="1">en</wf>
    <wf id="w5" offset="29" length="4" sent="1" para="1">EEII</wf>
  </text>
  <terms>
    <term id="t1" type="open" lemma="referir" pos="V" morphofeat="VMIP3S0">
      <span>
        <target id="w1" />
      </span>
    </term>
    <term id="t2" type="open" lemma="dolor" pos="N" morphofeat="NCMS000">
      <span>
        <target id="w2" />
      </span>
    </term>
    <term id="t3" type="open" lemma="intermtente" pos="G" morphofeat="AQ0CS0">
      <span>
        <target id="w3" />
      </span>
    </term>
    <term id="t4" type="close" lemma="en" pos="P" morphofeat="SPS00">
      <span>
        <target id="w4" />
      </span>
    </term>
    <term id="t5" type="close" lemma="extremidades_inferiores" pos="R"
         morphofeat="NP00000">
      <span>
        <target id="w5" />
      </span>
      <externalReferences>
        <externalRef resource="Yetano.2003" reference="extremidades inferiores" />
      </externalReferences>
    </term>
  </terms>
</NAF>
```

Figure 7.2: Output of the IXA-Pipes tokenizer and PoS tagger, enriched by UMLSmapper with short form annotations, for the sentence E6.

This module's task is to perform tokenization, part-of-speech tagging and constituent parsing on the input text. To do so, it can consume a web API of any third-party tool that provides the analysis in NLP Annotation Format (NAF).

The standard UMLSmapper configuration exploits IXA-Pipes (Agerri et al., 2014). An example of its output is given in Figure 7.2. Note that the analysis is performed on the original input text, and that the information regarding short forms is introduced as external references of term objects.

7.4.3 Candidate span generation

Input	Segmentation and morpho-syntactic information
Output	Text spans and their variants
Options	• Strategy to generate spans: rules over syntactic tree or ngram extractor • Maximum length of extracted spans

The objective of this module is to generate spans candidate of being linked to UMLS Metathesaurus concepts. Spans are extracted either by *a)* calculating n-grams that do not start or end with a stopword; or, *b)* applying rules to the constituent trees of the sentences, obtained also with IXA-Pipes.

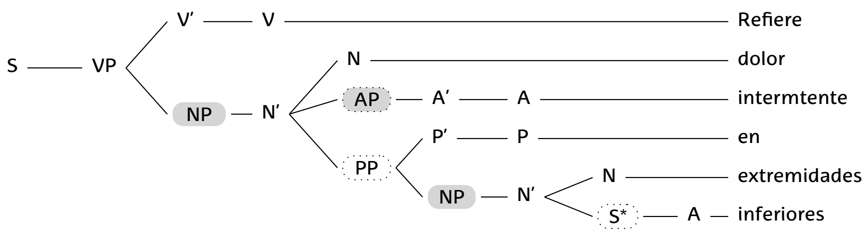


Figure 7.3: Constituent tree produced by IXA-Pipes for example E6 (note that the original node labels have been substituted for simpler, better-known labels; IXA-Pipes outputs AnCora's rich tagset (Taulé et al., 2008) [51])

The latter strategy consists in extracting from the constituent trees of each sentence all the possible phrases headed by nouns (N) or adjectives (A). Such subtree root nodes are shaded in grey in Figure 7.3, the constituent tree produced by IXA-Pipes for sentence E6. Each phrase tree can then produce one or more spans, depending on whether the phrase head is accompanied by modifiers. That is, the algorithm will compute the Cartesian product of the modifiers—from the span that includes all the modifiers to the span that has none (i.e., that includes just the head of the phrase). In practice, modifiers are taken to be phrases or clauses c-commanded by Ns or As (dotted in Figure 7.3).

For instance, the dominating noun phrase (NP), where ‘intermtente’ (sic) and ‘en extremidades inferiores’ are modifiers of the nucleus ‘dolor’, would produce the following 4 spans: *a)* ‘dolor intermtente en extremidades inferiores’, *b)* ‘dolor intermtente’, *c)* ‘dolor en extremidades inferiores’, and simply *d)* ‘dolor’. Similarly, the NP within the prepositional phrase (PP) would yield ‘extremidades inferiores’ and ‘extremidades’. Of note, ‘inferiores’ in ‘extremidades inferiores’ has incorrectly been parsed as a relative clause (marked with an *); had it been correctly parsed as an adjective phrase (AP), ‘inferiores’ would also be extracted as a candidate span.

Further, the module computes lemmatized variants of each span, in an attempt to maximize the recall of the next module (e.g., ‘extremidades’ yields the variant ‘extremidad’).

7.4.4 Candidate match retrieval

Input	Text span and generated variants
Output	Candidate links for the span to the UMLS Metathesaurus
Options	• Maximum number of retrieval hits • CUI blacklist • Language blacklist or whitelist • Source terminology blacklist or whitelist • Semantic type blacklist or whitelist • Suppressible or obsolete CUI acceptability

This module suggests candidate CUI links for each of the text spans generated by the prior module or the spans provided directly by the user. In practice, it constructs Lucene queries from those spans to retrieve similar CUI lexicalisations from the Metathesaurus index presented in Section 7.3.1.

The module accepts several whitelists and blacklists (see above), allowing for easy customisations of the knowledge base instead of having to compute a new index for each problem that requires focusing on specific subsets of the UMLS Metathesaurus. Let us consider query E7; it limits the search to documents that

a) contain ‘dolor’ and ‘intermtente’ in the normalised lexicalisation, each within an allowed Levenshtein (1966) edit distance of 2, and in any order of appearance,

b) belong to the source terminologies SNOMED CT, MeSH or MedDRA (original English terminologies or their translations to Spanish),

c) are not suppressible nor obsolete, and

d) do not belong to the given set of semantic types nor their hypernyms (activity [`acty`], behaviour [`bhvr`] and so on).

Note that, while the queries are built programmatically with Lucene's Java API, here we show human-readable representations in Lucene's parser syntax [52]:

E7
```
+strnorm:dolor~2
+strnorm:intermtente~2
#sab:"(SCTSPA MSHSPA MDRSPA SNOMED_US MSH MDR)"
-suppress:"(E Y O)"
-stypath:"(acty bhvr ... shro)"
```

The following example applies the same constraints, but the search concerns the span 'extremidades inferiores' (and lemmatised variants):

E8
```
+spanOr([strnorm:extremidades~2, strnorm:extremidad~2)])
+spanOr([strnorm:inferiores~2, strnorm:inferior~2)])
#sab:"(SCTSPA MSHSPA MDRSPA SNOMED_US MSH MDR)"
-suppress:"(E Y O)"
-stypath:"(acty bhvr ... shro)"
```

The results of these queries are shown in Tables 7.4 and 7.5, respectively. LSF (Lucene Scoring Function) indicates the score given by Lucene to each hit. Notice how Lucene assigns a much higher score to 'flebografía de extremidad inferior por RM' when queried with 'extremidades inferiores' than to 'dolor intermitente' when queried with 'dolor intermtente' (sic). Lucene's score does not measure the lexical similarity between the indexed entries and the query; it measures the relevance of an indexed entry with respect to the query and in contrast to the rest of the entries in the index [53].

Table 7.4: Documents retrieved from the Metathesaurus index with query E7 ('dolor intermtente')

cui	str	LSF
C1282310	dolor intermitente	17.533

7.4.5 Scoring and thresholding

Input	One or more candidate matches for the same text span
Output	Ranked and filtered matches
Options	• Function to score matches (see below) • Threshold, i.e., minimum score below which hits are discarded

As Tables 7.4 and 7.5 illustrate, the LSF score is not a reliable estimator of which retrieval hit matches best the queried span, in the sense that we handle here. This module assigns new scores to the candidates using a function other than LSF, and filters out candidates by applying a minimum-score threshold.

Table 7.5: Documents retrieved from the Metathesaurus index with query E8 ('extremidades inferiores')

cui	str	LSF
C1720201	extremidad inferior o ambas extremidades inferiores	590.054
C0023216	extremidad inferior	560.878
C0023216	Extremidad Inferior	560.878
C0023216	Extremidades Inferiores	560.878
C0230411	superficie anterior de la extremidad inferior	547.716
C0230411	estructura de la cara anterior de la extremidad inferior	512.850
C1562943	estructura de la pelvis y/o las extremidades inferiores	508.224
C1633984	flebografía de extremidad inferior por RM	508.224
C1640384	ecoflebografía de extremidades inferiores	508.224

The prototype has two alternatives to LSF: the function by Castro et al. (2010), CSF, and a variant of it, hereafter CSF'. CSF is given by:

$$CSF = \frac{overlapTokens(q, r)^2}{tokens(q) \cdot tokens(r)} \tag{7.1}$$

where $overlapTokens$ is the length in tokens of the overlap between the query, q, and the normalised lexicalisation of the retrieved hit, r. Because this function only counts as overlaps tokens that match exactly in q and r, it penalises severely the hits that might be a small edit distance away from the query—a possibility that we introduce on purpose with the lemmatisation and the fuzzy queries—.

The variant function CSF' intends to soften this penalty by counting substrings instead of tokens:

$$CSF' = \frac{overlapSubstrings(q, r)^2}{characters(q) \cdot characters(r)} \tag{7.2}$$

$overlapSubstrings$ extracts the longest common substrings between q and r and returns the length in characters of their concatenation. Tables 7.6 and 7.7 show the CSF and CSF' scores for the hits listed in Tables 7.4 and 7.5, respectively.

Table 7.6: Table 7.4 documents re-scored with CSF and CSF'

cui	str	LSF	CSF	CSF'
C1282310	dolor intermitente	17.533	0.250	0.837

After re-ranking the hits, the module applies a threshold given by the user in order to discard candidates with scores lower than desired. As a result, three scenarios are possible: that none of the candidates passes the filter, that only one

Table 7.7: Table 7.5 documents re-scored with CSF and CSF'

cui	str	LSF	CSF	CSF'
C1720201	extremidad inferior o ambas extremidades inferiores	590.054	0.400	0.469
C0023216	extremidad inferior	560.878	0.000	0.826
C0023216	Extremidad Inferior	560.878	0.000	0.826
C0023216	Extremidades Inferiores	560.878	1.000	1.000
C0230411	superficie anterior de la extremidad inferior	547.716	0.000	0.402
C0230411	estructura de la cara anterior de la extremidad inferior	512.850	0.000	0.357
C1562943	estructura de la pelvis y/o las extremidades inferiores	508.224	0.400	0.535
C1633984	flebografía de extremidad inferior por RM	508.224	0.000	0.462
C1640384	ecoflebografía de extremidades inferiores	508.224	0.667	0.554

passes the filter, or that more than one pass it. The final match of a span is the candidate with highest score, if there still are any. If more than one candidate has the top score, the next module (Section 7.4.6) is invoked to choose the final match.

Let us consider a threshold of 0.7 in the above examples. The span 'dolor intermtente' would not be linked at all when using CSF, as the score assigned to the document with CUI C1282310 is lower than the threshold; with CSF', the hit passes the filter so the span would receive this link. As for the span 'extremidades inferiores', the document with CUI C0023216 receives a perfect score regardless of the scoring function; hence, this would be the final match for the span.

7.4.6 Disambiguation

Input	Two or more equally ranked candidate matches for the same text span
Output	Final match for the text span
Options	• Disambiguation strategy: UKB, first, skip or none (i.e., skip this step)

This module is only invoked when a span has more than one top-scored mapping candidate. Notice that not only ambiguous lexicalisations trigger this situation—which they do, inevitably; because of the scoring functions explained in the previous section, different lexicalisations can also receive the same score. That is, two sources of ambiguity come into play: the first is given by the Metathesaurus, when it assigns several CUI (i.e., meanings) to one lexicalisation. This is proper ambiguity in a linguistic sense. The second is produced at runtime and depends on the scoring function used: it is possible that distinct lexicalisations (each mapped to a different CUI) receive the same score. All the same, the user must choose how the system should behave in these situations. UMLSmapper offers 4 possibilities:

- Choose one candidate performing WSD with UKB (Agirre et al., 2009).
- Simply choose the first candidate.
- Skip this module, i.e., return all the top-scoring candidates.
- Reject ambiguous candidates, i.e., do not return any candidate at all.

The algorithm behind UKB is Personalized PageRank (Haveliwala, 2002). A possible application would be, as in Agirre et al. (2010), to first map all the non-ambiguous spans in the text and then use those as context to assign a CUI to the ambiguous ones.

Here we explore a somewhat different approach. Initializing the graph is an expensive process, given its massive size (which will become clear in the next chapter). Thus, we want to do it just once and as early in the processing chain as possible. The context here consists simply of the tokens in the text, without stopwords; the system is able to provide this information as early as the basic linguistic analysis is done. When the disambiguation module is put to work, it just chooses the CUI with highest activation among the mapping candidates in the PageRank vector.

7.5 Conclusions

This chapter presented UMLSmapper, a prototype to perform unsupervised biomedical term identification with the UMLS Metathesaurus. The system is prepared to do end-to-end term identification (i.e., recognise terms and identify them in the same step) or it may receive text annotated with the terms to be normalised. In principle, the prototype may be used to process text in any language well-enough covered in the Metathesaurus, as long as basic NLP tools are available for that language.

The system is lexically motivated. In few words, it consists of a pipeline that extracts text spans candidate to be mapped, consults an Apache Lucene™ index of the Metathesaurus to retrieve relevant lexicalisations, and ranks them according to lexical similarity. When more than one candidate obtain top scores, the user may choose to apply UKB, a program for WSD, in order to choose the most semantically relevant. When processing text in Spanish, the user may also choose to carry out a pre-processing step, consisting in the automatic detection of abbreviated forms and their expansion to long forms.

In the next chapter, we evaluate this prototype on the task of end-to-end biomedical term recognition using the Mantra Gold Standard Corpus (Mantra GSC) (Kors et al., 2015), which comprises short texts in English and Spanish manually annotated with UMLS Metathesaurus CUI. Our results are analysed thoroughly and compared to two other systems. UMLSmapper is also used in Chapter 12 to prepare a corpus of medical assertion classification.

Chapter 8

Term identification: experiments with the Mantra GSC

8.1 Introduction

In this chapter we evaluate several approaches, including the system UMLSmapperpresented in Chapter 7, to identify biomedical terminology in text written in Spanish and English. The compared systems exploit symbolic or hybrid Natural Language Processing (NLP) techniques to map to the texts a specific subset of the Unified Medical Language System (UMLS) Metathesaurus.

These systems perform term recognition and identification in a single step with no supervision; they do so solely by exploiting the lexical and semantic information contained in the Metathesaurus. That is, the decision of what constitutes a term and what does not is not outsourced to an automatic medical entity recogniser, but is resolved drawing on the knowledge base itself, the Metathesaurus, with which the system is trying to produce links. This type of systems is needed in situations where training entity recognisers is not a viable option or existing recognisers are not well suited to the particular problem at hand. Further, these systems can be easily adapted to different application domains by subsetting or extending the Metathesaurus as needed.

In this chapter, then, we evaluate UMLSmapper alongside two such systems: MetaMap (Aronson, 2001, 2006)—a well-known rule-based system for term identification in English—and Transfer (Accuosto et al., 2018)—a pipeline that uses automatic translation to perform term identification in languages other than English. Furthermore, we test several combinations of UMLSmapper and Transfer. Said systems are assessed against the Mantra Gold Standard Corpus (Mantra GSC) (Kors et al., 2015), a corpus of scientific article excerpts and drug labels manually annotated with UMLS Concept Unique Identifier (CUI)s.

The chapter is organised as follows: in Section 8.2 we present the data used throughout the chapter (namely, the Mantra GSC and part of the UMLS Metathesaurus), all the compared systems and their combinations, as well as the evaluation framework. Section 8.3 reports the obtained results in the Spanish and English data of the Mantra GSC and provides a thorough error analysis of UMLSmapper. Finally, 8.4 summarises the conclusions extracted from the work described in the chapter.

8.2 Materials and methods

8.2.1 Data

The evaluation described in this chapter uses the Mantra GSC (Kors et al., 2015) as testing corpus. Next, we describe this corpus and the subset of the UMLS Metathesaurus with which it was annotated and that we, in turn, take as reference to configure the selected systems.

8.2.1.1 The Mantra GSC

The Mantra GSC is a collection of parallel biomedical corpora in English, French, German, Spanish, and Dutch that has been manually annotated with concepts of the UMLS Metathesaurus to test concept identification systems.

As per the published description of the corpus (Kors et al., 2015), the Mantra GSC annotation policy limits the annotations to concepts of the UMLS Metathesaurus that meet the following two criteria:

- the concept belongs to the terminologies Medical Subject Headings® (MeSH), Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), and/or the Medical Dictionary of Regulatory Activities (MedDRA);
- the concept belongs to one or more of these semantic groups (McCray et al., 2001; Bodenreider et al., 2003): anatomy (`anat`), chemicals and drugs (`chem`), devices (`devi`), disorders (`diso`), geographic areas (`geog`), living beings (`livb`), objects (`objc`), phenomena (`phen`), physiology (`phys`), and procedures (`proc`).

In the following section, we describe thoroughly this subset of the UMLS Metathesaurus (henceforth referred to as the *Mantra terminology* per Kors et al. (2015)), as it is relevant to the configuration of the systems tested in this chapter, and it also helps understand the difficulty of the problem.

Table 8.1 shows the size of the corpus subset that is used in this work—namely, the Spanish (es) and English (en) samples. This subset consists of 100 parallel text samples for two different genres: scientific abstract titles from Medline, and drug labels from the European Medicines Agency (EMEA). A total number of

639 and 648 annotations can be found, respectively, in the Spanish and English samples, which in turn point to 550 and 559 CUIs of the UMLS Metathesaurus. Note that the systems evaluated do not need to be trained, so the whole corpus is used for testing throughout the chapter.

Table 8.1: Size of Mantra GSC Spanish (es) and English (en) data sets. Tokens are counted after whitespace tokenisation.

	Medline		EMEA	
	es	**en**	**es**	**en**
# documents	100	100	100	100
# tokens	1,087	989	1,984	1,738
# annotations	278	285	361	363
discontinuous	5	7	12	10
ambiguous	40	41	61	60
suppressible	1	2	2	4
missourced	0	0	5	6
unique concepts	285	288	295	301

Of these annotations, 17 in each language are discontinuous (i.e., the concepts are expressed in disjoint text spans) and 101 in each language—more than 18% of the total annotations—are "ambiguous" (i.e., the text spans are linked to more than one CUI). This is due to the human annotators not being able to resolve the "semantic difference between the suggested concepts" (Kors et al., 2015, p. 950); that is, having multiple annotations for the same text span does not indicate that the meaning of the target phrase itself is ambiguous, but that there are several entries in the UMLS Metathesaurus that seemingly denote the same concept.

It is also interesting to note that, according to the 2016AA UMLS release, a few of the annotations point to suppressible concepts or can only be found in UMLS sources that are not supposed to be included in the Mantra terminology (labelled as "missourced" in Table 8.1). These facts suggest that Mantra GSC annotations are based on a UMLS release older than 2016AA, the one used to configure the systems evaluated in the chapter.

8.2.1.2 The Mantra terminology

Figure 8.1 shows the distribution of CUIs over terminology sources in the English and Spanish Mantra terminology. Some interesting observations can be made:

- Most of the concepts (44.59%) can only be found in the original (i.e., English) version of MeSH.
- The Spanish translation of MeSH is a small proper subset of the original counterpart—it covers less than a tenth part of MeSH.

- The second largest subset (40.74%) is composed of concepts in the intersection of only SNOMED CT and its translation to Spanish.
- The Spanish translation of SNOMED CT is almost completely contained in the English version, except for 46 concepts.
- MedDRA and its translation to Spanish overlap completely; that is, the whole MedDRA has been translated to Spanish.

Overall, there are 327,160 and 38 concepts that can only be accessed through English or Spanish terms, respectively, while 372,672 concepts are common to both English and Spanish. That is, the conceptual coverage of the Spanish Mantra terminology with respect to the English is of 53.25%. The whole Spanish and English Mantra terminology contains 699,770 concepts and 2,993,323 terms (1,938,466 in English and 1,094,413 in Spanish)[1].

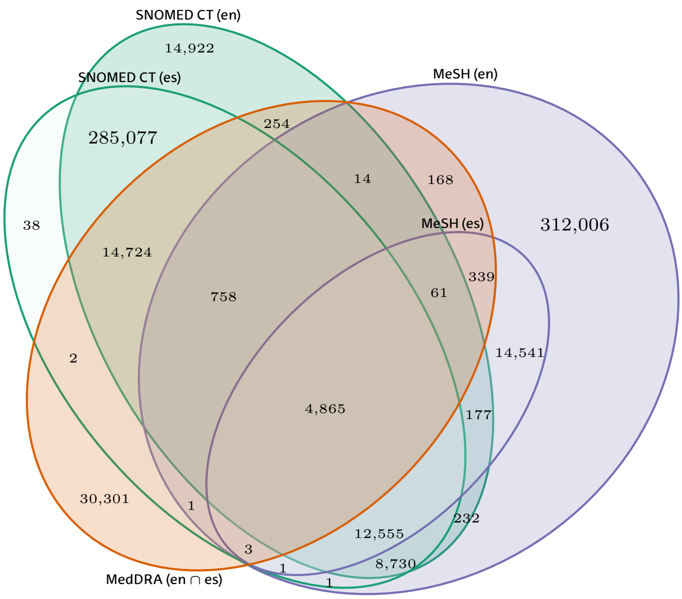


Figure 8.1: Size of the Mantra terminology by vocabulary source (not in scale).

[1]Kors et al. (2015, p. 949) report that "[t]he Mantra terminology includes 591 918 concepts with a total of 3 238 015 terms, most of which are in English (2 039 988), followed by Spanish (785 083)." We have been unable to replicate these numbers; the reasons might include the difference in the UMLS version, a different method to count concepts and terms, or that there were other criteria in creating the Mantra terminology that they did not report in the article.

Table 8.2: Distribution of SNOMED CT ∪ MeSH ∪ MedDRA concepts in Spanish (es) and English (en) over the 10 Mantra-accepted semantic groups, and their proportion.

	es	en	es/en (%)
chemicals and drugs (`chem`)	44,521	347,581	12.81
disorders (`diso`)	155,222	169,850	91.39
procedures (`proc`)	70,597	75,798	93.14
living beings (`livb`)	41,465	42,770	96.95
anatomy (`anat`)	30,831	31,470	97.97
devices (`devi`)	13,255	14,229	93.15
object (`objc`)	5,388	5,980	90.10
physiology (`phys`)	5,335	5,689	93.78
phenomena (`phen`)	4,919	5,251	93.67
geographic areas (`geog`)	1,028	1,059	97.07
`chem` ∩ `objc`	45	51	88.24
`chem` ∩ `phen`	4	4	100.00
Total	372,610	699,732	53.25

Regarding semantic groups, most of the 10 Mantra-accepted semantic groups are well covered in Spanish (see Table 8.2), except for chemicals and drugs, of which only 12.81% of the concepts in the English subset have at least one Spanish term associated. More than 90% of the missing concepts is accounted for by the following 4 semantic types (UMLS Type Unique Identifier (TUI), given between parenthesis): organic chemical (T109), amino acid, peptide, or protein (T116), clinical drug (T200), and nucleic acid, nucleoside, or nucleotid (T114). Furthermore, 99.66% of the missing chemicals and drugs belong to MeSH.

8.2.2 Systems

The experiments conducted in this chapter involve three systems that perform term normalisation of biomedical texts through symbolic or hybrid NLP pipelines: *a)* MetaMap, *b)* a system that exploits machine translation, and *c)* UMLSmapper. We also explore combinations of the latter two. Furthermore, MetaMap and UMLSmapper have two variants each: one for processing text in English and another for Spanish.

Foe the systems to be compared under the same conditions, all of them exploit the same knowledge base, which comprises a total of 675,670 CUIs: all CUIs accessible through Spanish lexicalisations in the Mantra terminology, plus all the chemicals and drugs in the English Mantra terminology. The inclusion of the English chemicals and drugs was motivated by the poor coverage of this semantic group in the Spanish terminology (see Table 8.2).

8.2.2.1 MetaMap

The baseline for the experiments is established by MetaMap (Aronson, 2001, 2006) 2016v2 [54], a well-known program developed at the National Library of Medicine (NLM) for the specific purpose of projecting the UMLS Metathesaurus onto biomedical text. It was primarily developed to process text written in English, although it can be easily customised to exploit any custom knowledge base—albeit with an expected performance loss due to the modules for lexico-morphological analysis, in which MetaMap relies heavily, not being prepared for languages other than English, among other limitations. That is, MetaMap is expected to be a stable competitive baseline in the English evaluations, while lagging behind in the Spanish evaluations.

For the evaluations over the Spanish portion of the Mantra GSC, the MetaMap Data File Builder [55] was used to compile the custom knowledge base of the aforementioned 675,670 concepts and corresponding lexicalisations. It must be noted that MetaMap can only read ASCII encoded files. Thus, both the terms indexed and the test input texts had to be converted to ASCII. We used the Linux command `iconv -f utf-8 -t ascii//TRANSLIT`, which replaces non-ASCII characters with their transliterations (e.g., it converts "publicaciones científicas en español" to "publicaciones cientificas en espanol").

As for execution details, MetaMap was launched with default arguments except the following:

- `-y`: perform Word Sense Disambiguation (WSD)
- `-V`: use the custom knowledge base
- `-R`: constrain the annotations to sources in the Mantra terminology
- `-k`: constrain the annotations to semantic types in the Mantra terminology

8.2.2.2 UMLSmapper

UMLSmapper has been introduced in Chapter 7. In short, it approaches the problem of term normalisation through an information retrieval mechanism to identify terms based on a linguistic analysis and a disambiguation procedure. In contrast to Transfer (next system), it does so natively in the language of the input texts—English or Spanish.

The UMLSmapper variant for Spanish uses the Spanish tokenisation and Part of Speech (PoS) tagging models distributed with the IXA-pipeline (Agerri et al., 2014), along with the abbreviation detection and resolution module introduced earlier (Section 7.4.1 of Chapter 7). The variant for English uses the analogous IXA-pipeline models for English and does not have a module specific for handling abbreviations. This is the only disadvantage over the Spanish variant. The knowledge graph for the WSD module built on the UKB program (Agirre et al., 2009), common to both variants, has 675,670 edges and 4,669,477 relations among them. The rest of the configuration parameters are shown in Table 8.3; they were chosen empirically in early experiments with UMLSmapper (Perez et al., 2018).

Table 8.3: UMLSmapper configuration

Parameter	Value
Abbreviation and acronym detection (Section 7.4.1)	
Strategy	Random Forest classifier
Candidate span generation (Section 7.4.3)	
Strategy	ngram extractor
Maximum length	5 tokens
Candidate match retrieval (Section 7.4.4)	
Maximum number of hits	100
CUI blacklist	C0032863, C0557651
Term blacklist	'ii', 'hace'
Scoring and thresholding (Section 7.4.5)	
Scoring function	Castro et al. (2010)
Threshold	0.7
Disambiguation (Section 7.4.6)	
Strategy	UKB

8.2.2.3 Transfer pipeline

The transfer pipeline (Accuosto et al., 2018; Perez et al., 2020) (henceforth, Transfer) automatically translates the input texts into English and uses MetaMap at its full potential to produce the UMLS annotations on the translated text. Then, it transfers the obtained annotations back to the original text. It could be said to be a step forward in the work proposed by Carrero et al. (2008a,b).

In short, the process of annotation transfer consists in assigning the annotation (i.e., the CUI) to the span in the original text that gives maximum cosine similarity with any of the lexicalisations of said CUI. The cosine similarity is computed over biomedical Spanish fastText embeddings (Bojanowski et al., 2017) pre-trained for this purpose.

In contrast to UMLSmapper, this pipeline does not require lexical resources in the language of the input texts because MetaMap does all the heavy lifting in this regard. Still, Transfer requires an automatic translation model to English that is suited for the biomedical domain and the desired origin language. In this work, the reported results are obtained using a Neural MT (NMT) Spanish-English model trained on the UFAL medical corpus [56] and the data released for the WMT2016 biomedical translation task (Bojar et al., 2016).

8.2.2.4 Combination of Transfer and UMLSmapper

Because of the fundamental differences between UMLSmapper and Transfer, they are expected to succeed and fail in different types of annotations. Thus, combining the two pipelines may prove beneficial. In this chapter, we also evaluate three combinations of Transfer and UMLSmapper, which differ in the way that overlapping predictions are handled:

- **Joint (+)**: Annotates the union of spans with the union of the CUIs.
- **Joint (T)**: Takes as valid the prediction made by Transfer.
- **Joint (U)**: Takes as valid the prediction made by UMLSmapper.

Let us illustrate the output of these combinations with an example. The true annotations of the text 'Headaches can occur with normal human immunoglobulin' (Example E1) link two text spans to a different concept each, here labelled A and D for simplicity:

E1 Con la **inmunoglobulina**$_A$ humana normal pueden producirse **cefaleas**$_D$

Examples E2 and E3 show the predictions of Transfer and UMLSmapper for the same text respectively:

E2 Con la **inmunoglobulina humana normal**$_B$ pueden producirse cefaleas

E3 Con la **inmunoglobulina humana**$_C$ normal pueden producirse **cefaleas**$_D$

Neither manages to predict correctly the span nor the CUI of 'inmunoglobulina'. Transfer misses the term 'cefaleas', while UMLSmapper manages to annotate it correctly in span and CUI. Then, the combinations Joint (+), Joint (T) and Joint (U) would produce the following annotations (Examples E4, E5 and E6 respectively):

E4 Con la **inmunoglobulina humana normal**$_{(B,C)}$ pueden producirse **cefaleas**$_D$

E5 Con la **inmunoglobulina humana normal**$_B$ pueden producirse **cefaleas**$_D$

E6 Con la **inmunoglobulina humana**$_C$ normal pueden producirse **cefaleas**$_D$

8.2.3 Evaluation

The main evaluation scenario of this chapter is **term normalisation** (a.k.a., *term identification*, or *term recognition and disambiguation*, among others). This scenario measures how good systems are at detecting relevant biomedical terms and assigning to them a Concept Unique Identifier (CUI) of the UMLS Metathesaurus. The systems presented earlier are tested against the English and Spanish datasets of the Mantra GSC. Their performance is measured in precision (P), recall (R) and F_1-score (F_1), whose definitions we repeat here for convenience:

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F_1 = 2 \cdot \frac{P \cdot R}{P + R} \qquad (6.1\ (=4.1))$$

In the context of this chapter, true positives (TP), false positives (FP) and false negatives (FN) are counted as follows:

- TP: number of predictions that match in span boundaries and CUI with a gold annotation.
- FP: number of predictions that do not match in span boundaries with any gold annotation or that have a different CUI to the gold annotation they match with.
- FN: number of gold annotations that do not match in span boundaries with any prediction or that have a different CUI to the prediction they match with.

The reported P, R and F_1 are micro-averages (μ). Further, we report two metric variants:

- **Strict**: requires the boundary matches to be exact.
- **Relaxed**: accepts as correct matches predictions that do not have exactly the same boundaries as a gold annotation but that overlap with one.

All of these definitions apply to discontinuous gold annotations as well, even if none of the systems assessed, except MetaMap, is able to produce discontinuous predictions. As for ambiguous gold annotations, a prediction is only required to guess one of the gold CUIs in order to be counted as a true positive, on account of the suggested gold CUIs being interchangeable rather than complementary, as explained in Section 8.2.1.1.

In addition, we report overlap percentages (OP) (Accuosto et al., 2018) alongside the relaxed measurements. This metric indicates how similar the predicted spans are to the gold standard, as the relaxed measurements allow for inexact matches. The overlap percentage of two annotations a and b is calculated as the relation between the length of the overlapping span and the length of the longest annotation:

$$OP(a, b) = 100 \cdot \frac{len(overlap(a, b))}{max(len(a), len(b))} \tag{8.1}$$

We report macro-average OP.

As complementary measurements to help explain the performance of the systems, we also compute how well the systems do at less demanding scenarios: **term recognition** and **term classification**. In the former, we are concerned with the correctness of the annotated spans, i.e., the CUIs are ignored when counting true and false predictions. In the latter, we look at the semantic groups to which the gold and predicted CUIs belong. To put it simply, the label space is reduced from 675,670 to 10 in term classification (10 semantic groups) and to just 1 in term recognition.

Finally, the evaluation ends with a comprehensive error analysis of UMLSmapper in the Spanish test data.

8.3 Results

8.3.1 Term identification in Spanish

UMLSmapper achieves a global F_1-score of 0.626 in the strict term identification scenario. Table 8.4 shows the results broken down by semantic groups. As can be seen, the results vary greatly from one semantic group to another as well as from one sub-dataset to another. At the same time, some of the semantic groups are more poorly represented than others. Hence, it is not possible to make generalised, categorical statements about the performance of UMLSmapper over semantic groups. Looking at this dataset in particular, we can simply say that UMLSmapper has achieved the best scores for chemicals and drugs, living beings and geographic areas (the latter has just 10 examples in total); the worst results were obtained for objects (11 examples), devices (6) and physiology (30).

Let us compare UMLSmapper's results with the other presented systems. Strict and relaxed scores for term identification are shown in Table 8.5, where we also include the reported results of Roller et al. (2018) and Yuan et al. (2022) as reference, who apply more advanced techniques but assume oracle terms in their evaluations. Regarding Medline and considering non-combination systems, all systems improve the baseline, MetaMap, by more than 0.9 F_1-score points. The pipelines based in transfer are remarkably precise (0.720 and 0.767 on strict and relaxed evaluations, respectively) compared to UMLSmapper and the baseline, but they do not improve the baseline's recall at all. Overall, UMLSmapper achieves the best F_1-score (0.630 and 0.634). It exceeds the other systems in terms of recall particularly, while lifting precision as well with respect to the baseline. As for EMEA, a similar pattern as in Medline can be observed, except that the best F_1-score when span overlaps are allowed is achieved by Transfer. This is due to the outstandingly high precision, which outdoes the better recall obtained by UMLSmapper.

Table 8.4: Results of strict term identification by UMLSmapper on the Spanish Mantra GSC over UMLS Metathesaurus semantic groups. # is the number of gold annotations.

	Medline				EMEA				All			
	#	P	R	F_1	#	P	R	F_1	#	P	R	F_1
diso	100	0.732	0.710	0.721	111	0.656	0.568	0.609	211	0.694	0.635	0.663
chem	27	0.583	0.519	0.549	93	0.811	0.828	0.819	120	0.765	0.758	0.762
proc	57	0.545	0.421	0.475	58	0.404	0.397	0.400	115	0.465	0.409	0.435
livb	37	0.683	0.757	0.718	45	0.630	0.644	0.637	85	0.655	0.695	0.675
anat	26	0.739	0.654	0.694	20	0.400	0.600	0.480	46	0.547	0.630	0.586
phys	12	0.417	0.417	0.417	19	0.667	0.526	0.588	31	0.556	0.484	0.517
phen	6	0.571	0.667	0.615	7	0.625	0.714	0.667	13	0.600	0.692	0.643
objc	3	0.000	0.000	0.000	6	0.235	0.667	0.348	9	0.190	0.444	0.267
geog	7	0.667	0.857	0.750	0	0.000	0.000	0.000	7	0.545	0.857	0.667
devi	3	0.250	0.333	0.286	3	0.200	0.333	0.250	6	0.222	0.333	0.267
μ	278	0.645	0.615	0.630	361	0.615	0.632	0.623	639	0.627	0.624	0.626

Combining the pipelines yields slightly better results than using them in isolation, the improvement being more pronounced in the case of EMEA. Specifically, recall does raise with respect to UMLSmapper—the best evaluated system in this regard—, but precision is almost always worse. Among the three combinations, Joint (+) and Joint (T) seem to work best, except in strict Medline, where Joint (U) works better than Joint (T). Given that Transfer is more precise than UMLSmapper (as Figure 8.2 illustrates), it makes sense that the combinations that prefer Transfer's predictions in case of conflict tend to yield better results.

Regarding the performance at the different annotation levels, as Figure 8.3 shows, the losses from the easiest task (namely, term recognition) to the most difficult (term identification) are small—$\sim$6 F_1-score percentage points. That is, if a system recognises correctly a term, the link to the UMLS Metathesaurus suggested for that term is most likely correct as well. This is true for all the systems. One could think, then, that a better term recogniser would lift this upper bound. However, none of the systems evaluated here (all of which use as core engines MetaMap, UMLSmapper, or both) resolve these tasks sequentially: first recognise a term, then categorise it into coarse-grained categories, and finally predict an identity. It is rather the other way around: a term is only recognised insofar as it meets certain criteria to be assigned a particular identity; otherwise, it is simply not recognised at all. Hence the behaviour depicted in Figure 8.3.

8.3.2 Term identification in English

Table 8.6 reports the results of the experiments in the English dataset of the Mantra GSC. Here, we have included a second version of MetaMap, which consists

Table 8.5: Results of term identification on the Spanish Mantra GSC. The results of BTM (Roller et al., 2018) and CODER (Yuan et al., 2022), in italics, assume oracle terms.

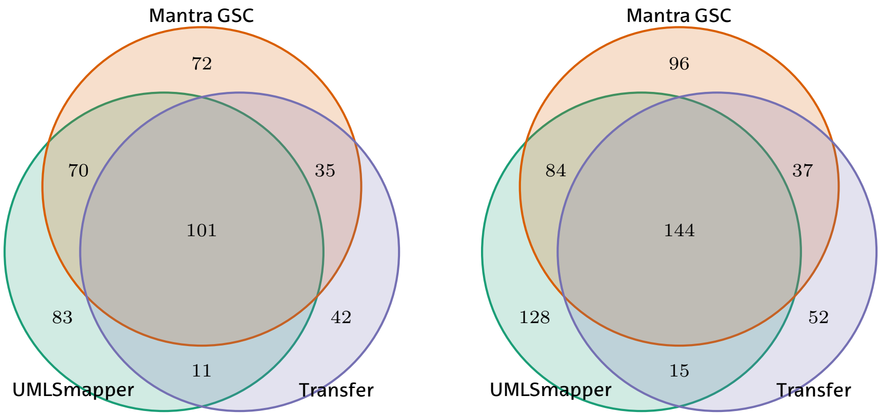
		Medline				**EMEA**			
		P	**R**	$\mathbf{F_1}$	**OP**	**P**	**R**	$\mathbf{F_1}$	**OP**
Strict	MetaMap	0.486	0.496	0.491		0.405	0.443	0.423	
	Transfer	**0.720**	0.489	0.582		**0.730**	0.501	0.594	
	UMLSmapper	0.645	0.615	0.630		0.615	0.632	0.623	
	Joint (+)	0.598	**0.678**	**0.636**		0.584	**0.701**	**0.637**	
	Joint (T)	0.620	0.612	0.616		0.624	0.662	0.642	
	Joint (U)	0.627	0.640	0.633		0.596	0.637	0.616	
	BTM	*0.781*	*0.619*	*0.691*					
	CODER			*0.704*				*0.681*	
Relaxed	MetaMap	0.511	0.522	0.516	86.02	0.430	0.471	0.450	85.35
	Transfer	**0.767**	0.522	0.621	87.88	**0.810**	0.557	0.660	**90.72**
	UMLSmapper	0.649	0.619	0.634	**90.61**	0.636	0.654	0.645	88.02
	Joint (+)	0.629	**0.712**	**0.668**	88.78	0.640	**0.767**	0.698	88.32
	Joint (T)	0.657	0.647	0.652	88.07	0.679	0.720	**0.699**	90.19
	Joint (U)	0.634	0.647	0.641	90.61	0.622	0.665	0.643	87.92

of the original, out-of-the-box MetaMap without modifications to the knowledge base. That is, this MetaMap variant (identified as ⊟) is not limited to annotating concepts of the Mantra terminology. For comparison purposes, we also include in the experimentation an analogous UMLSmapper variant.

UMLSmapper has obtained an overall F_1-score of 0.674, surpassing MetaMap across the board, both in the restricted and the unrestricted (⊟) frameworks, as well as the strict and relaxed evaluations. It must be pointed out that the evaluation dataset consists of grammatical, standard and formal biomedical text; it might be the case that in less controlled text genres, such as health records,

Table 8.6: Results of term identification on the English Mantra GSC.

		Medline				**EMEA**			
		P	**R**	$\mathbf{F_1}$	**OP**	**P**	**R**	$\mathbf{F_1}$	**OP**
Strict	MetaMap	0.628	0.628	0.628		0.600	0.653	0.625	
	MetaMap (⊟)	0.355	0.572	0.438		0.268	0.576	0.365	
	UMLSmapper	**0.701**	**0.660**	**0.680**		**0.651**	**0.689**	**0.669**	
	UMLSmapper (⊟)	0.526	0.681	0.593		0.444	0.702	0.544	
Relaxed	MetaMap	0.663	0.663	0.663	89.71	0.613	0.667	0.639	**92.23**
	MetaMap (⊟)	0.379	0.611	0.468	87.59	0.274	0.590	0.374	89.09
	UMLSmapper	**0.705**	0.663	**0.684**	91.10	**0.654**	0.691	**0.672**	91.45
	UMLSmapper (⊟)	0.537	**0.695**	0.606	**91.77**	0.448	**0.708**	0.549	91.41



(a) Medline (b) EMEA

Figure 8.2: Overlap of gold annotations (Mantra GSC) and predictions made by UMLSmapper and Transfer in the strict term identification scenario (not in scale).

the lexico-morphological engine of MetaMap grants it a greater advantage over UMLSmapper. Currently there is no corpus publicly available to test this setup.

As for the differences between the Mantra-specific and unrestricted variants (⊟), the unrestricted variants suffer an expected loss of precision due to having augmented the knowledge bases beyond the Mantra terminology. Recall values are not that affected in comparison, and even improve in the case of UMLSmapper.

It is also interesting to note that both systems perform slightly better compared to the results in the Spanish dataset (Table 8.5). Taking into account that the two datasets (i.e., English and Spanish) are parallel, their level of difficulty can be safely assumed to be similar. If anything, the English dataset could be said to be more challenging as it contains a few more annotations and more unique concepts (see Table 8.1). Still, the systems perform consistently better in English than in Spanish. This is not surprising in the case of MetaMap, because its original intended usage was for this language in particular. In the case of UMLSmapper, the improvement can only be explained by the richer lexical coverage of the knowledge base in English, as explained in Section 8.2.1.1.

8.3.3 Error analysis

This section provides a manual error analysis of UMLSmapper on the entire Spanish Mantra GSC. In sum, UMLSmapper has made 240 false negative and 237 false positive errors. Table 8.7 relates the types of errors identified and their frequency. Each type of error is explained below.

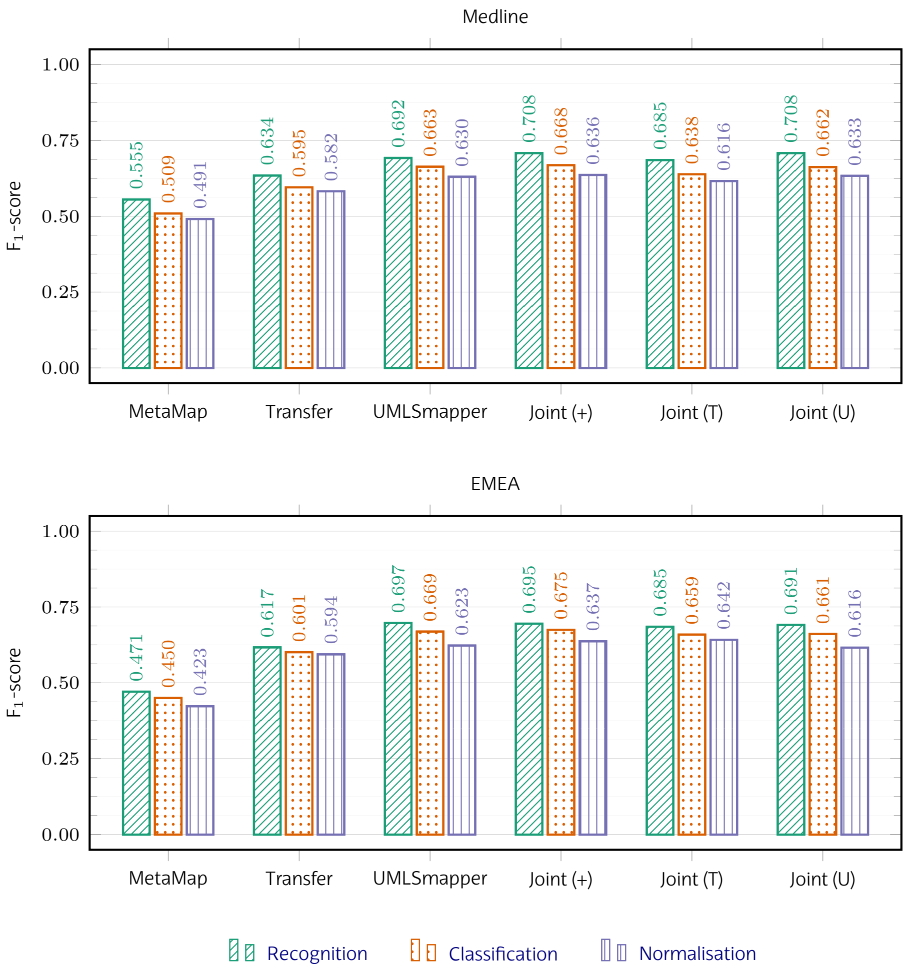


Figure 8.3: Strict F_1-score results for term recognition, classification and normalisation on the Spanish Mantra GSC

Table 8.7: Classification of errors and their distribution

		%
False positives	Terms included in the UMLS but senses missing	40.5
	Missed multi-word annotations, annotated shorter spans	32.1
	Discrepancies with gold standard	19.8
	WSD errors	5.5
	Other	2.1
False negatives	Lexical variability issues	41.7
	Made multi-word annotations containing the gold span	12.5
	Over- or underspecification	10.4
	Discrepancies with gold standard	7.5
	Discontinuous gold annotations	7.1
	Other	6.2
	Exact lexical match with incorrect CUI	5.4
	WSD errors	5.0

Most of the **false positives** are errors made by the system due to relying completely on pure lexical match with the knowledge base, while the knowledge base does not capture all the possible meanings of the terms it contains. Thus, we annotate concepts that are not actually denoted in the texts. Consider the following example: the word "organismo" has at least two meanings: *a)* organism, living being; and *b)* organisation, institution. While the former meaning is captured in the UMLS (as the concept C0029235 [57]), the latter is not. Consequently, whenever the word "organismo" is used in the dataset, UMLSmapper annotates it as C0029235 regardless of the actual intended meaning.

The next most common spurious predictions were made as a consequence of missing a multi-word gold annotation and having made shorter spanned predictions contained within the boundaries of the gold span (e.g., annotating "Staphylococcus aureus" instead of the expected "Staphylococcus aureus meticilin resistente"). Of the 76 errors of this type, we consider 67 are given correct CUIs.

Next in frequency, we fail to understand why 19.8% of the false positives are not annotated in gold standard corpus, i.e., we believe that the predictions are correct and that they are missing in the corpus. For instance, given the sentence "Diarrea crónica 'naturalmente' identificable en la anamnesis." (*Chronic diarrhea 'naturally' recognizable in the anamnesis.*); "anamnesis" is not annotated in the gold standard corpus, although there exists a concept in the Mantra terminology, C0199182 [58], which we believe denotes exactly that.

13 of the 237 false positives stem from UKB errors. UKB is only invoked when several top-ranked CUIs compete to become final annotations for a phrase. This happens 175 times in total on the whole dataset, of which in 134 the term is correctly recognised. In 13 of those 134 cases, UKB assigns an incorrect CUI to

the phrase. That is, UKB has made a correct guess 90% of the times it has been invoked. The remaining marginal type of spurious annotations are explained by incorrect brief form expansions, faulty stopword treatment, and/or inaccurate sentence boundary detection.

Regarding **false negative** or missing predictions (240 in total), we find more variability in the typology of errors. 100 are due to lexical variability: the UMLS Metathesaurus does not capture all the existing synonyms, singulars and plurals, morphological derivations, and so on, and we do not treat this problem other than with lemmatisation and the expansion of abbreviated forms.

Some annotations are missed because of having made predictions that involve more tokens than the gold annotations. Consider the following example: UMLSmapper maps the concept C1708335 [59]—healthy participant or subject—to the phrase "voluntarios sanos" (*healthy volunteers*), while the gold standard only annotates "voluntarios". Of these type of errors, we consider that 20 are given incorrect CUIs, but 10 could be considered correct.

Another 25 gold annotations are missed because the gold CUI denotes concepts more specific or less specific than the actual words annotated do when taken literally, and world knowledge or common sense is needed to resolve the gap between the two. For instance, in the sentence "Valoración de la capacidad de esfuerzo en la EPOC" (*Assessment of effort capacity in COPD*), human annotators assign C0015264 [60]—physical effort—to the span "esfuerzo" (*effort*), because they know that COPD has nothing to do, say, with mental effort, and the sentence only makes sense if the word "effort" does denote physical effort. However, the Spanish lexicalisations of C0015264 explicitly mention physical effort, so the lexical match with "esfuerzo" does not go through.

In 45 cases, the term is correctly recognised but a CUI is given to the term that does not coincide with the gold annotation. Of these 45, we judge that 18 times the CUI proposed is correct, and thus it should be ambiguous—or more ambiguous, if it already is—(these 18 annotations contribute to the 47 controversial false positives mentioned earlier). 13 other errors are due to incorrect disambiguation (the same as explained earlier).

17 annotations are discontinuous, as described in Section 8.2.1.1; UMLSmapper does not make discontinuous annotations with the present configuration. Thus, these annotations add to the missed predictions inevitably.

Finally, the remaining false negatives are due to incorrect tokenisation or lemmatisation of the input text, or because of the system's configuration: the gold concept is not in the knowledge base, the gold annotation is longer than the maximum annotation allowed in UMLSmapper, and so on.

8.4 Conclusion

In this chapter, we have evaluated UMLSmapper in a gold standard corpus of biomedical text annotated with UMLS entities and CUIs: the Mantra GSC (Kors et al., 2015). We have focused on the parallel Spanish-English subset, comprised of scientific paper titles and drug labels. The annotations cover a specific subset of the Metathesaurus, consisting of the three most important terminological sources—namely, SNOMED CT, MeSH and MedDRA—and 10 semantic groups.

UMLSmapper has obtained an overall F_1-score of 0.626 and 0.674, in Spanish and English respectively, in the most demanding evaluation scenario: strict term identification. This scenario requires predictions to match exactly in span boundaries and linked CUIs with gold annotations. UMLSmapper has shown balanced precision and recall metrics, with better precision than recall in the article titles sub-corpus and the other way round in the drug labels. The results varied greatly when broken down by semantic group, although no conclusion can be extracted in this respect due to the scarce representation of most of the groups in the corpus.

A manual error analysis of the predictions has shown that the main source of missing as well as spurious predictions is the dependency of the tool on a rich lexical and semantic coverage by the knowledge base. On the one hand, meanings of polysemous expressions missing in the knowledge base may lead to false positive predictions, as the tool may link a span to one of the other registered meanings. On the other hand, a poor coverage of the lexicalisations of the concepts in the knowledge base leads to false negative predictions, because the tool relies on approximate lexical match with the knowledge base to recognise terms. Of note, the disambiguation module built on UKB (Agirre et al., 2009, 2010) has shown an accuracy of 90%, and just 13 incorrect predictions out of a total of 636 predictions can be traced down to this module.

UMLSmapper has been compared to two other systems: MetaMap (Aronson, 2001, 2006) and Transfer (Accuosto et al., 2018; Perez et al., 2020). In the experiments involving the Spanish data, MetaMap has served as a naive baseline: we simply compiled a new MetaMap knowledge base with the Spanish lexicalisations of the Mantra terminology, even though MetaMap's mapping engine draws heavily upon rules and heuristics implemented for the English language. Unsurprisingly, UMLSmapper has surpassed this baseline. In the experiments with English data, we consider MetaMap a competitive baseline, which UMLSmapper has also managed to improve—although by a much narrower margin. We concede that in less standard text, such as health records, MetaMap may prove a better alternative thanks to its more powerful lexical engine.

Transfer is a pipeline conceived to process texts in one language with a term identification engine for another. The pipeline first translates the input text to the engines' language, annotates the translation with said engine, then transfers the annotations in the translated text to the original text using semantic similarity techniques. The implementation reported here uses a NMT Spanish-English model, MetaMap, and biomedical Spanish fastText embeddings (that is, it was only evaluated in the Spanish data). This pipeline showed worse F_1-scores than UMLSmapper. Specifically, it showed greater imbalance between precision and recall: it yielded by far the best precision of all the compared systems but ranked among the worst in terms of recall.

Finally, we also evaluated the combination of UMLSmapper and Transfer, seeing as the two pipelines make complementary predictions. The evaluated combinations manage to improve the results of the individual pipelines, specifically by rising the recall metrics. Furthermore, the most competitive combinations are those that favour Transfer's predictions in case of conflict, which is expected because Transfer has a higher correct prediction rate.

Analysing the results at different levels of difficulty—namely, term recognition, term classification and term identification—, we saw that all the compared systems have a similar behaviour: whenever a term is correctly recognised, it is almost always correctly identified (thus, correctly classified as well). Thinking about the results in these terms, the upper bound of the tools seems to be in the term recognition. However, we also explained that this chain of thoughts does not apply to UMLSmapper nor to the other compared systems because in neither case is the process of term recognition independent from that of term identification.

PART IV

NEGATION AND UNCERTAINTY DETECTION

Chapter 9

Negation and speculation: background and literature review

9.1 Definition and motivation

Negation is the universal linguistic phenomenon whereby the polarity of statements or clauses is reversed. In the English language, it is most evidently realised by the words ‘no’ and ‘not’, but also ‘never’, and even the prefixes ‘a-’ and ‘in-’, for example. **Speculation** has to do with modality. In this work and the related studies, it is an umbrella term that refers broadly to linguistic phenomena related to hedging, evidentiality, uncertainty, and factuality (Morante et al., 2012c). To put it simply, we construe speculation as explicit language that signals a speaker is unsure whether a statement is true or lacks evidence to commit fully to it.

Properly detecting and handling these phenomena is crucial because they are ubiquitous and have a direct, strong impact on the quality and usability of clinical solutions based on NLP. Doctors and nurses write about negative findings and hypothesised explanations as much as positive observations. An incorrect interpretation of this data by an automated clinical support program might simply lead to harmful medical decisions.

Automatic negation and speculation processing is a well-established research topic, particularly for English, as show several survey articles on the matter (Jiménez-Zafra et al., 2018b; Cruz Díaz et al., 2019; Jiménez-Zafra et al., 2020b; Morante et al., 2021). The processing of negation in Spanish text has gained attention too in recent years, encouraged by the NEGES (Negation in Spanish) workshops (Jiménez-Zafra et al., 2018a, 2019) and the publications of several freely available corpora, which we present succinctly below. Notably, the automatic processing of speculation, a fuzzier and inconspicuous phenomenon than negation, is yet to be thoroughly addressed in Spanish text.

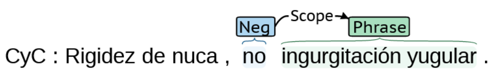
It must be noted that the processing of negation (more than speculation) is also of interest in Natural Language Processing (NLP) research areas other than the clinical, especially in relation to sentiment analysis, where negative expressions may reverse or reinforce the polarity of a text. Actually, Barnes et al. (2021) demonstrate that explicitly training a model with negation as an auxiliary task helps improve the main task of sentiment analysis.

The NLP community has proposed multiple models to represent the problem of negation and speculation detection:

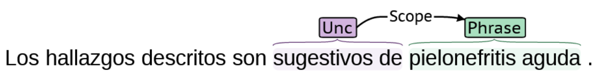
- On the one hand, there is the task of **detecting cues and scopes**, the constituent parts of negation and speculation, as pictured in Figure 9.1. **Cues** (also known as **markers** or **triggers**) are words or phrases that express negation or speculation. **Scopes** are the clauses affected by a cue, that is, whose propositional values are somehow modified or reversed. Some works focus exclusively on finding the scopes of given pre-annotated cues; this task is known as negation and/or speculation **scope resolution**. The detection of cues and/or scopes is usually addressed as a sequence labelling problem.

- The second common way of modelling negation and uncertainty detection in the biomedical field is as a text classification task known as **assertion classification**. In this case, the text to analyse is pre-annotated with medical entities, whose **assertion category**—present, absent, or possible—needs to be automatically determined. The sentences of Figure 9.1 are depicted in Figure 9.2 framed as entity assertion annotations.

Astride the previous two, a few works study the recognition of negated medical entities, i.e., they explore sequence labelling approaches to target exclusively negated medical entities.



(a) A negation cue and its scope. Translation: "H[ead] & N[eck]: stiff neck, no jugular vein distention".

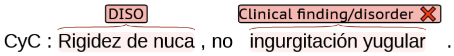


(b) An uncertainty cue and its scope. Translation: "The findings described are suggestive of acute pyelonephritis.".

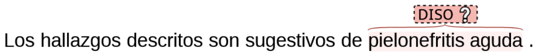
Tumoraciones faciales en paciente transplantada hepatica

(c) Example without negation nor uncertainty cues. Translation: "Facial tumors in liver transplant patient".

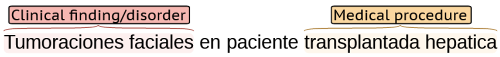
Figure 9.1: Annotations of negation and uncertainty cues and scopes.



(a) Medical entities annotated as absent (red cross). `DISO` stands for "clinical finding/disorder". Translation: "H[ead] & N[eck]: stiff neck, no jugular vein distention".



(b) A medical entity annotated as possible (white question mark and dashed border. Translation: "The findings described are suggestive of acute pyelonephritis.".



(c) Present medical entities. Translation: "Facial tumors in liver transplant patient".

Figure 9.2: Annotations of medical entities and their assertion category.

9.2 Related resources

In what follows, we present briefly the corpora of Spanish text annotated with negation and/or speculation information, with special attention to corpora of the biomedical domain. They are presented in ascendant chronological order of publication. Multiple review articles can be found in the literature on this topic (Cruz Díaz et al., 2019; Jiménez-Zafra et al., 2020b; Morante et al., 2021), to which we refer the reader interested in other languages or domains.

The presented corpora differ in text genre and domain, and conform to divergent guidelines for string-level annotations. In this respect, we must mention the effort of the NEGES organisers towards providing a unifying framework for the annotation of negation in Spanish through Task 3 of the 2018 workshop edition (Jiménez-Zafra et al., 2019).

UAM Spanish Treebank (Moreno et al., 2003; Moreno Sandoval et al., 2013) The first ever Spanish corpus annotated for negation consists of 1,500 sentences of the news domain and the corresponding syntactic trees after the PENN treebank model (Marcus et al., 1994). In 2013, it was enriched with annotations of negation cues and scopes based on BioScope guidelines (Szarvas et al., 2008; Vincze et al., 2008). The corpus is freely available under a non-commercial license [61].

IxaMed-GS (Oronoz et al., 2015) The IxaMed Gold Standard corpus consists of 75 health reports from outpatient consultations. Although the primary focus of this work is on adverse drug reaction (ADR) events, the annotations include information about negation and speculation as well. Specifically, they encode this information as attributes of the annotated entities. In this sense, the annotations are akin to those shown in Figure 9.2 for assertion classification. The corpus is not public due to confidentiality issues.

UHU-HUVR (Cruz Díaz et al., 2017) This corpus of 604 clinical reports from a Spanish hospital was manually annotated with negation cues, their linguistic scope, and clinically relevant events (the latter based on the THYME guidelines [Styler IV et al., 2014]). It was the first Spanish corpus to include affixal negation annotations. At present, the corpus is not publicly available in spite of the author's alleged intention to make it so.

IULA-SCRC (Marimon et al., 2017a) The IULA Spanish Clinical Record Corpus is the first clinical corpus annotated with negation-related information to be publicly available [62]. The corpus consists of 3,194 sentences of which 1,093 contain negation. The annotations consist of negation markers and their scope, inside which relevant medical entities are also highlighted, among other data. The annotation policy is loosely based on the BioScope (Szarvas et al., 2008; Vincze et al., 2008) and ConanDoyle-neg (Morante et al., 2012b) guidelines.

Cotik et al. (2017) This corpus consists of 513 ultrasound reports manually annotated with a diverse set of entity types and relations. Among the entities we find negation and speculation cues, which are linked to the entities of type 'finding' they have scope over. The authors do not acknowledge having based their annotation guidelines in any other previous work. This corpus is private due to the sensitivity of the data.

SFU Review$_{\mathrm{SP}}$-NEG (Martí et al., 2016; Jiménez-Zafra et al., 2018c) This corpus stems from the Spanish portion of the SFU Review corpus (Taboada et al., 2006), which comprises 400 product reviews across 8 domains. The manually annotated negation structures consists of cues, scopes, and events. The SFU Review$_{\mathrm{SP}}$-NEG corpus was used in the 2018 edition of the NEGES workshop (Jiménez-Zafra et al., 2019) for the task on automatically detecting negation cues. It is available for non-commercial purposes [63].

NewsCom (Taulé et al., 2021) The NewsCom corpus consists of 2,955 comments posted in response to news articles from online newspapers. The NewsCom guidelines extend those of SFU Review$_{\mathrm{SP}}$-NEG (Martí et al., 2016; Jiménez-Zafra et al., 2018c) to include criteria for the annotation of the focus of negation. This work is in fact the first to include the annotation of foci in Spanish text. It contains 2,965 negative structures with their corresponding negation cue, scope, and focus. The corpus is available upon request [64].

T-MexNeg (Bel-Enguix et al., 2021) This corpus consists of 13,704 tweets written in Mexican Spanish, out of which 4,895 contain negation structures. The annotation guidelines, adapted from those of SFU Review$_{SP}$-NEG (Martí et al., 2016; Jiménez-Zafra et al., 2018c) to better conform to the Twitter text genre, identify three main negation components: cues, scopes, and events. The corpus is available as a GitLab repository [65].

E3C (Magnini et al., 2021a,b) The European Clinical Case Corpus (E3C) is a collection of clinical cases in 5 languages, namely, Italian, English, French, Spanish, and Basque. The authors propose an adaptation of the THYME guidelines (Styler IV et al., 2014), where negation and speculation information is added as attributes of events, similarly to IxaMed-GS (Oronoz et al., 2015). The Spanish portion of the corpus consists of 1,400 clinical cases. It is publicly available at the European Language Grid catalogue [45].

Also relevant is the work by Campillos Llanos et al. (2017), who analyse a corpus of 354,677 emergency admission notes in search of negation contexts by applying hand-crafted patterns. It is to date the biggest corpus considered in such a study. On the downside, the automated annotation of the corpus through patterns (a thorough manual analysis being impracticable) poses the risk of missing the long tail of negation contexts. This corpus is also not publicly available.

Table 9.1 offers a comparative view of the corpora from the clinical domain. As can be seen, just two of them are available, of which only the smallest (i.e., EC3 [Magnini et al., 2021a,b]) considers speculation. It does so at the entity and event level. The other available corpus is IULA-SCRC (Marimon et al., 2017a), which is thrice the size of E3C, although it only annotates negation, in this case, following the cue-scope model.

One of the contributions of this thesis is the NUBes corpus, a collection of sentences extracted from health records and manually annotated with negation and speculation cues and scopes. The corpus is introduced in Chapter 10. It is currently the biggest corpus of the clinical domain annotated thus that is publicly available. It must be noted that the work reported in this thesis regarding NUBes predates some of the studies described here, such as the latest corpora and the publications of the results of the NEGES workshops.

9.3 State of the Art

Regarding work devoted to the automatic processing of negation and speculation in Spanish, we find approaches based on hand-crafted heuristics, shallow machine learning and, more recently, deep learning. Table 9.2 offers a summary of this work, which we elaborate below; of note, the table also exposes how fragmented

Table 9.1: Spanish biomedical corpora with annotations of negation and/or speculation, adapted from Jiménez-Zafra et al. (2018b) and Martí et al. (2018). The upper table section describes the corpora qualitatively, in terms of the types of annotations they contain; the middle table section describes the corpora quantitatively. [1]27.58% of the diseases annotated are negated. [2]1.90% of the diseases annotated are speculative. [3]513 radiology reports. [4]56% of the findings are negated.

	IxaMed-GSC	**UHU-HUVR**	**IULA-SCRC**	**Cotik et al. (2017)**	**E3C**
Negation cue		✓	✓	✓	
Speculation cue				✓	
Scope		✓	✓		
Entity	✓	✓	✓	✓	✓
Event					✓
Total sentences	5,410	8,412	3,194	?[3]	1,134
with negation (%)	?[1]	2,298 (27.32)	1,093 (34.22)	?[4]	240 (21.16)
with speculation (%)	?[2]	-	-	?	114 (10.05)
Available at	-	-	[62]	-	[45]

this research field is, the only comparable results being those pertaining to the NEGES workshops (Jiménez-Zafra et al., 2018a, 2019) or having been authored by the same researcher team.

The earliest related studies (Costumero et al., 2014; Cotik et al., 2015; Stricker et al., 2015; Santiso et al., 2017; Solarte-Pabón et al., 2020) consist of adaptations and/or extensions of NegEx (Chapman et al., 2001) to the Spanish language. NegEx is an algorithm originally based on English lexicons that categorises pre-annotated medical entities as present or absent given the contexts the entities occur in. These Spanish adaptations obtain F_1-scores (F_1) 0.64 to 0.78 in diverse corpora and evaluation methodologies.

Koza et al. (2019) worked on the recognition of negated medical findings in radiological reports by means of rules based on morpho-syntactic and semantic information. They report an F_1 of 0.98 on an evaluation against their own private corpus but acknowledge that the test data set lacks variability in the negation structures it includes.

The task of recognising negated findings has also been undertaken by Santiso et al. (2019, 2020), but with machine learning techniques. They approach the problem as a sequence labelling task. They first assess Conditional Random Fields (CRF) (Lafferty et al., 2001) over symbolic features and features derived from word embeddings, achieving 0.82 and 0.75 span-level F_1 (partial match) in IULA-SCRC (Marimon et al., 2017a) and their private corpus IxaMed-GS (Oronoz et al., 2015), respectively. Next, they implement a Recurrent Neural Network (RNN) featuring character embeddings, bidirectional LSTMs (biLSTM) layers and a CRF classifier, surpassing their previous results on IxaMed-GS.

Table 9.2: Literature review on negation and uncertainty detection in Spanish text. *SEM 2012 F1 is the evaluation metric proposed by Morante et al. (2012a) for the *SEM 2012 shared task on resolving the scope and focus and negation. ZS stands for zero-shot performance. Notice that scores are only comparable if they result from the same evaluation corpus, task and metric. An extensive discussion of the different evaluation metrics can be consulted in Sineva et al. (2021).

Reference	Task	Approach	Metric	Score
Tested on SFU Review$_{SP}$-NEG (Jiménez-Zafra et al., 2018c)				
Loharja et al. (2018)	NEG cue detection	CRF	*SEM-2012 F_1	0.86
Fabregat et al. (2018a)	NEG cue detection	biLSTM	*SEM-2012 F_1	0.68
Fabregat et al. (2019b)	NEG cue detection	biLSTM	*SEM-2012 F_1	0.83
Beltrán et al. (2019)	NEG cue detection	CRF	*SEM-2012 F_1	0.84
Domínguez-Mas et al. (2019)	NEG cue detection	CRF	*SEM-2012 F_1	0.81
Giudice (2019)	NEG cue detection	bi-GRU	*SEM-2012 F_1	0.23
Jiménez-Zafra et al. (2020a)	NEG cue detection	CRF	*SEM-2012 F_1	0.87
"	NEG scope resolution	CRF	*SEM-2012 F_1	0.81
Shaitarova et al. (2020)	NEG scope resolution	Transformer$_{ZS}$	token F_1	0.78
Shaitarova et al. (2021)	NEG scope resolution	Transformer$_{ZS}$	token F_1	0.79
Rivera Zavala et al. (2020)	NEG cue+scope detection	Transformer	*SEM-2012 F_1	0.88
Tested on IULA-SCRC (Marimon et al., 2017a)				
Hartmann et al. (2021)	NEG scope resolution	Transformer$_{ZS}$	*SEM-2012 F_1	0.94
Solarte-Pabón et al. (2020)	NEG cue+scope detection	Rules	sentence F_1	0.92
Rivera Zavala et al. (2020)	NEG cue+scope detection	biLSTM+CRF	CoNLL-2010 F_1	0.85
Santiso et al. (2019)	negated entity detection	CRF	inexact span F_1	0.82
Solarte Pabón et al. (2022)	NEG scope detection	Transformer	token$_{wBIO}$ F_1	0.88
Tested on NUBES (Chapter 10)				
Lima-López et al. (2020a)	NEG cue detection	biLSTM+CRF	token F_1	0.96
"	UNC cue detection	biLSTM+CRF	token F_1	0.85
"	NEG scope detection	biLSTM+CRF	token F_1	0.91
"	UNC scope detection	biLSTM+CRF	token F_1	0.79
Hartmann et al. (2021)	NEG scope resolution	Transformer$_{ZS}$	*SEM-2012 F_1	0.90
Solarte Pabón et al. (2022)	NEG cue detection	Transformer	token$_{wBIO}$ F_1	0.95
"	UNC cue detection	Transformer	token$_{wBIO}$ F_1	0.84
"	NEG scope detection	Transformer	token$_{wBIO}$ F_1	0.88
"	UNC scope detection	Transformer	token$_{wBIO}$ F_1	0.72
Tested on private corpora				
Costumero et al. (2014)	assertion classification	Rules	F_1	0.74
Stricker et al. (2015)	assertion classification	Rules	F_1	0.67
Koza et al. (2019)	negated entity detection	Rules	sentence F_1	0.98
Santiso et al. (2017)	negated entity detection	CRF+Rules	inexact span F_1	0.74
Santiso et al. (2019)	negated entity detection	CRF	inexact span F_1	0.75
Santiso et al. (2020)	negated entity detection	biLSTM+CRF	inexact span F_1	0.82
Solarte Pabón et al. (2022)	NEG cue detection	Transformer$_{ZS}$	token$_{wBIO}$ F_1	0.90
"	UNC cue detection	Transformer$_{ZS}$	token$_{wBIO}$ F_1	0.81
"	NEG scope detection	Transformer$_{ZS}$	token$_{wBIO}$ F_1	0.84
"	UNC scope detection	Transformer$_{ZS}$	token$_{wBIO}$ F_1	0.74

Systems based on CRFs and biLSTMs were also the most popular among the participants of the shared task about negation cue detection in the NEGES workshops (Fabregat et al., 2018a; Loharja et al., 2018; Beltrán et al., 2019; Domínguez-Mas et al., 2019; Fabregat et al., 2019b; Giudice, 2019). The corpus provided in both workshop editions to train and test the competing systems was SFU Review$_{\mathrm{SP}}$-NEG (Jiménez-Zafra et al., 2018c). The best overall results (0.86 span-level F_1) were obtained by Loharja et al. (2018) with a CRF classifier over lexical and morphological features.

The organisers of NEGES implemented another CRF classifier and managed to improve the state of the art on negation cue detection in SFU Review$_{\mathrm{SP}}$-NEG with an F_1 of 0.87 (Jiménez-Zafra et al., 2020a). This is also the first work in the literature that tackles the problem of negation scope resolution along with cue detection in Spanish text. Specifically, they follow a 2-stage setup with two separate classifiers, where the first detects cues, whose scopes are determined by the second. The classifier of scopes yields F_1s of 0.81 and 0.73 with gold and predicted cues as input, respectively.

In view of the across-the-board success of the neural network Transformer architecture (Vaswani et al., 2017) and the availability of pre-trained neural language models steadily and rapidly increasing in number, the focus of works about negation detection has lately shifted towards studying these models' behaviour and advantages.

Rivera Zavala et al. (2020) compare a RNN-based classifier and a Transformer-based classifier in the task of negation cue detection and scope resolution in the corpora IULA and SFU Review$_{\mathrm{SP}}$-NEG. The RNN classifier combines character, word and sense embeddings as input to a biLSTM network, whose output is fed to a CRFs classifier. The Bidirectional Encoder Representations from Transformers (BERT)-based system follows the conventional setup of a pre-trained language model (Multilingual BERT o mBERT [23]) with a softmax output layer. Both systems tackle the problem of cue and scope detection jointly. They achieve 0.81 and 0.85 token-level F_1 with BERT and the RNN, respectively, in the IULA-SCRC corpus. In SFU Review$_{\mathrm{SP}}$-NEG, the results are 0.92 and 0.88.

Shaitarova et al. (2020, 2021) explore the transferability of negation scope resolution models between the languages English, French, Spanish and Russian. Their work is built on NegBERT (Khandelwal et al., 2020), a system originally built for English that performs negation cue detection and scope resolution in a 2-stage fashion using BERT. These works adapt NegBERT to the cross-lingual setting by replacing BERT with Multilingual BERT (mBERT) and XLM-RobERTa (Conneau et al., 2020). They achieve token-level F_1s $\sim$0.78 when zero-shot testing English and French models on the SFU Review$_{\mathrm{SP}}$-NEG corpus, with XLM-RobERTa outperforming mBERT by a narrow margin.

Hartmann et al. (2021) also study zero-shot cross-lingual transfer approaches for negation scope resolution. Specifically, they explore how to best exploit disparate available datasets (in their work, multiple datasets in English) to overcome the lack of training data on the target languages (here, Spanish). They propose the application of a Multi-Task Deep Neural Network (MT-DNN) (X. Liu et al., 2019), where each dataset available for training is treated as an independent task. This approach is compared to the simple concatenation of the training datasets, which they find works slightly better overall when evaluated in IULA-SCRC (Marimon et al., 2017a) and NUBes (Chapter 10), among others. They report *SEM 2012 scope token F_1s (Morante et al., 2012a) of 0.94 and 0.90 in these datasets, respectively.

Notably, the processing of speculation, a task considerably more difficult than the detection of negation cues and scopes, is yet to be thoroughly addressed in Spanish text (clinical or otherwise). Lima-López et al. (2020a) report the first exploratory experiments with the NUBes corpus using the biLSTM + CRF architecture over a rich set of morpho-syntactic and lexical features. This work has recently been extended to incorporate the first published experiments with a Transformer-based model on the NUBes corpus (Solarte Pabón et al., 2022), achieving similar results to Lima-López et al. (2020a). In Chapters 11 and 12, we carry out a battery of experiments with NUBes and Transformer models, among others, managing to surpass all previous scores.

Chapter 10

NUBes: A corpus of negation and uncertainty in Spanish clinical texts

10.1 Introduction

This chapter describes the NUBes corpus (from Negation and Uncertainty annotations in Biomedical texts in Spanish), a new collection of health record excerpts enriched with negation and uncertainty annotations. To date, NUBes is one of the largest available corpora of clinical reports in Spanish annotated with negation, and the first to include the annotation of speculation cues and scopes.

In a nutshell, **cues** (also called **markers** or **triggers**) are words or phrases that express negation or speculation; **scopes** are the phrases or clauses affected by a cue, that is, whose propositional values are somehow modified. For a higher level of granularity, there are other elements that can be annotated, such as the element within the scope most clearly affected by the cue—usually a medical entity—, or the element that reinforces or diminishes the meaning of the cue, called a **polarity item**. A typical annotation that includes all these elements is shown in Figure 10.1:

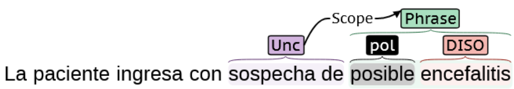


Figure 10.1: A sentence annotated with an uncertainty cue and a scope with a polarity item and a medical entity of type Disorder. Translation: "The patient is admitted under suspicion of possible encephalitis".

The chapter is structured as follows: Section 10.2 starts by describing the origin and pre-processing of the raw data with which NUBes was created. Next, it explains the methodology followed to write the annotation policy and to annotate the corpus. Finally, it discusses the limitations of this work and the corpus itself. Section 10.3 first presents the final annotation guidelines of NUBes, then reports the inter-annotator agreement, and provides a quantitative description of NUBes. It also discusses the differences of NUBes with related corpora. Finally, Section 10.4 concludes the chapter and establishes the links with the following chapters.

10.2 Materials and Methods

10.2.1 Data

NUBes consists of health records provided by a Spanish private hospital. Specifically, we extracted plain text from 7 sections consisting of free narrative—namely, Chief Complaint (CC), History of Present Illness (HPI), Physical Examination (PE), Diagnostic Tests (DXT), Patient History (hx), Progress Notes (PNo), and Treatment Notes (TNo)—, and split them into sentences with spaCy [37]. Then, documents were sampled into batches of around 3,000 sentences, by iteratively picking documents from random medical specialities and sections.

Further, NUBes had to be anonymised as a requirement to its publication. Succinctly, the anonymisation process consisted of 3 phases:

1. Manual annotation of sensitive information, such as names, dates, locations, contact details, and so on. The result of this phase was the corpus NUBes-PHI described in Chapter 5, Experiments with health records.
2. Manual revision of the alleged false positive errors committed by 3 systems when applied to the whole NUBes-PHI, having themselves been trained on NUBes-PHI, as explained in Section 5.3.4 of said chapter. This revision uncovered a few additional sensitive data items missed by the human annotators of NUBes-PHI.
3. Semi-automatic replacement of the identified sensitive data with similar phrases. We exploited methods based on rules and dictionaries designed for this purpose (Lima-López et al., 2020b).

Thus, the content's readability was preserved while being suitable for sharing.

In total, 10 batches have been anonymised and annotated with negation and uncertainty, amounting to 7,019 documents and 29,682 sentences. Of note, the public version of NUBes was shuffled at sentence level in order to hinder even further potential de-anonymisation efforts.

10.2.2 Methodology

An initial draft of our guidelines was produced by extending IULA-SCRC's (Marimon et al., 2017a) to include uncertainty. After annotating IULA-SCRC with this initial draft, we decided to make further changes with respect to negation by annotating

a) negations inside indirect speech (e.g., ‘The patient denies’);
b) verbs that convey a change of state (e.g., ‘remove’); and,
c) morphological negation (e.g., ‘incoherent’).

Other minor changes to the guidelines had to be made in order to accommodate uncertainty annotations. These differences with IULA-SCRC and other related corpora are further described in Section 10.3.4.

After producing the second draft, two linguists worked independently on the first batch of documents of the NUBes corpus. Their results were compared and multiple questions and disagreements that arose were discussed. The team also consulted a medical expert who aided them with some difficult scenarios, which are also examined in Section 10.3.4. These discussions contributed greatly towards producing the final version of the guidelines.

Then, the two linguists annotated the same batch adhering to the final guidelines. Next, a third annotator resolved the differences between the previous two in the first batch in order to create a Gold Standard. Finally, the remaining 9 batches were annotated by one linguist. The current NUBes release includes, then, one batch reviewed by three people and nine batches produced by a single annotator.

All the annotation work was done with BRAT (Stenetorp et al., 2012). To speed up the process, an automatic cue annotator service was developed for BRAT that detects a list of the most frequent cues. On average, we invested around eight hours of annotation work for each batch of ∼3,000 sentences.

10.2.3 Limitations

The most notable limitation of NUBes is the above-mentioned fact that ∼90% of the corpus has been annotated and reviewed by a single person. While the inter-annotator agreement rates on ∼3,000 sentences—reported below in Section 10.3.2—indicate our guidelines are clear and unambiguous-enough given the complexity of the task, we are aware that a corpus annotated to a large extent by one person does not meet the requirements to be considered a Gold Standard Corpus by the standards of the NLP community. Still, we defend that NUBes is a valuable contribution as the first and—at the moment of writing—only corpus annotated with negation and uncertainty phenomena in Spanish clinical text, helping further the researcher in this field while the quality of NUBes is improved and/or better corpora are published by other researchers in the future. In addition, to the best our knowledge, it is currently the biggest freely available corpus of real health record excerpts in Spanish, which we consider in itself quite valuable a contribution.

Another area for improvement involves the pre-processing of the data. Clinical text is known to be problematic even for the most basic NLP tasks, namely, sentence splitting and tokenisation (Cruz Díaz et al., 2015). While we have not performed a systematic evaluation of the existing splitters and tokenisers for the Spanish language and the clinical domain, none of such tools tested in informal evaluations stood out as producing consistently fewer or less serious errors than the others. We decided to use spaCy to split and tokenise NUBES for the sake of convenience, in spite of the result not being fully satisfactory.

Finally, we must acknowledge a limitation in the scope of the corpus itself. While negation is a binary operator, uncertainty is most certainly not; it is a continuum from utter conviction to pure speculation. At the moment, uncertainty annotations in NUBES do not include information about the level of confidence.

10.3 Results

10.3.1 NUBES annotation guidelines

This section reports the final annotation guidelines of NUBES. They define three main elements of interest: negation cues, uncertainty cues and their scope. Moreover, polarity items and entities are also annotated as part of the scope.

10.3.1.1 A note on formatting

The guidelines include a large number of examples to illustrate each of the rules and exceptions of the policy. For the sake of practicality, the examples are not presented in figures, as in the introduction, but in plain text. Here we introduce the formatting of the examples; the concepts listed below will be defined in the corresponding sections:

- **Boldface**: Negation or uncertainty marker
- *Italics*: Scope of a marker
- Solid underline: Medical entity
- {Curly brackets}: Polarity item
- Dotted underline: Scope of a marker located within another scope

The example of the introduction, repeated as Figure 10.2 for convenience, would be formatted as shown in example E1:

E1 La paciente ingresa con **sospecha de** *{posible} encefalitis*
The patient is admitted under suspicion of possible encephalitis

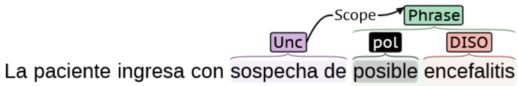


Figure 10.2: A sentence annotated with an uncertainty cue and a scope with a polarity item and an entity of type "disorder"

10.3.1.2 Negation cues

We define negation cues as elements that modify the truth value of a clause or specify the absence of an entity. Three different types of cues can be distinguished: syntactic, lexical and morphological.

10.3.1.2.1 Syntactic negation cue (`NSyn`) These are mostly function words or adverbs that can accompany multiple syntactic constructions or occur on their own. It is the simplest type of negation, as well as the most common, as it covers words such as 'no' (*no*) and 'sin' (*without*):

E2 **No** *ha tomado analgesia* (sic)
[The patient] has not taken any pain medication

E3 6.- *Drenaje*: **no**
6.- Drainage: no

E4 Fiebre de 38,5 **sin** *foco*
38.5 degrees fever without a focus

E5 **Nunca** *ha precisado valoración psiquiátrica*.
[The patient] has never required psychiatric assessment.

10.3.1.2.2 Lexical negation cue (`NLex`) They are content words or multi-word expressions that convey negation depending on the context, including verbs, adjectives or noun phrases. These cues are harder to detect as the way in which they negate a phrase is usually subtler than that of syntactic cues. Some examples are 'suspender' (*suspend*), 'incapacidad para' (*inability to*) or 'descartar' (*discard*):

E6 **Desestiman** *actualmente la realización de endoscopia*
At present they dismiss conducting an endoscopy

Phrases headed by negative determiners are also considered lexical cues:

E7 **Ninguna de ellas** *de evolución aguda-subaguda*
None of them of acute-subacute course

While far less common still, dashes can be used to indicate negative results of tests, which we also include in this category:

E8 Tira reactiva de orina: leucocitos (+), *eritrocitos* (**-**)
Urine strip test: leukocytes (+), erythrocytes (-)

10.3.1.2.3 Morphological negation cue (NMph) Morphological negation refers to negation by means of affixes. Since NUBEs is a medical texts corpus, we decided to limit the annotation of these cues to words that explicitly state the absence of symptoms ('afebril' [*afebrile*]) or that could be seen as negating a symptom or state ('deshidratado' [*dehidrated*]). Words that do not fulfil those conditions or that are part of a condition name are not annotated.

E9 **Afebril** al ingreso
Afebrile at admission

Furthermore, a word in question should be classified as a morphological negation cue only if the word can be paraphrased as a negated sentence that would be annotated under those conditions. For example, 'insuficiencia' (*failure*), as in Example E10, is not annotated because '?no suficiencia' or '?falta de suficiencia' are not grammatical or natural expressions in Spanish.

E10 Presentó descompensacion de su insuficiencia cardiaca (sic)
[The patient] showed decompensation of their heart failure

The intuition behind this rule is that 'insuficiencia' itself conveys a complete, independent idea in this context—meaning "diminished capacity" of the heart—, rather than being the negated counterpart of another concept upon which it depends to be assigned a meaning (as in the pairs 'capable'/'incapable', 'symptomatic'/'asymptomatic', 'oriented'/'disoriented', and so on).

10.3.1.2.4 Negation cue exceptions It is worth noting that not all occurrences of words that express negation are annotated as cues. Four main exceptions exist:

1. Concerning the adjective 'negativo' (*negative*), it is not a negation cue if it is part of name, e.g., 'bacterias Gram negativas' (*Gram-negative bacteria*).
2. Seeming negation cues might be used to modify the meaning of degree and frequency adverbs, as in 'no siempre' (*not always*). As the negation of a universal quantifier—e.g., 'not always'—is logically equivalent to the existential quantifier—e.g., 'sometimes'—, we do not consider these cases to constitute negation cues. Similar rationale applies to expressions like 'casi sin' (*almost no*).
3. Similarly, seeming negation cues can be part of uncertainty cues, like in 'no claro' (*not clear*). We elaborate on this exception in the next section.
4. In general, conditional constructions (E11), volition verbs (E12) and final adjuncts (E13) should not be considered for negation cues, as they describe wishes or events that might or might not happen in the future.

   E11 Si fiebre alta que no cede [...]
   If [they have] high fever that doesn't drop [...]

E12 Refiere molestias y quiere quitárselo
[The patient] says it hurts and wants it removed

E13 Varón de 68 años, remitido desde su C.Salud, para descartar TVP
68-year-old male sent by their local clinic to discard DVT

10.3.1.3 Uncertainty cues

Similarly to negation, uncertainty cues can be broken down into two groups: syntactic cues and lexical cues.

10.3.1.3.1 Syntactic uncertainty cue (`USyn`) The only instances of this class are the coordinating conjunction 'o' (*or*) and the preposition 'versus'. 'o' should not be annotated in the context of enumerations or when introducing paraphrases (E14), but when used to introduce alternative explanations, as in E15:

E14 En las intercrisis refiere sensación contínua de mareo o inestabilidad
[The patient] mentions continuous dizziness or instability

E15 *Una complicación postCNG* **o** *una patología de origen digestivo*
A post-coronary angiography complication or a pathology of digestive origin

10.3.1.3.2 Lexical uncertainty cue (`ULex`) As with lexical negation, these are content words that express uncertainty depending on the context. Some of the most used cues are 'probable', 'posible' and 'sospecha de' (see E16 and E17). Verbs in the conditional tense or subjunctive mood also treated as uncertainty cues, including those that usually act as negation cues, as in example E18.

E16 **Sospecha de** *dehiscencia de suturas*
Suspicion of wound dehiscence

E17 **Se pensó en** *un origen funcional de ambos síntomas*
A functional origin of both symptoms was considered

E18 **Descartaría** *{de forma razonable} una arteritis [...] como causa de la clínica*
It would reasonably rule out [...] arteritis as the origin of the symptoms

As with negation, certain punctuation marks are sometimes (rarely) used to indicate uncertainty. In this case, we are concerned with the question marks '¿' and/or '?':

E19 **¿***Ca in situ***?**.
Ca in situ?

Medical jargon deserves special attention in this section, as reports abound with phrasings with very specific meanings that might surprise the non-expert annotator. The most compelling cases include (the exclamation mark indicates that the sentence is not standard Castilian Spanish):

- ‘orientar’, lit. *navigate* or *aim to*, here *indicate* or *point to*:

  E20 !Todo ello **orienta** *junto con la clínica a un cuadro suboclusivo*
  All this, along with the symptoms, points to/?orients to a subocclusion case

- ‘impresionar’, lit. *move*, *affect*; here, *strike as*, *look like* from ‘dar la impresión’:

  E21 ![...] **impresionando** *el cuadro de síndrome confusional*
  [...] the case striking as/?impressing as a confusional state

- ‘asociarse’, lit. *join* or *merge*; vague umbrella expression related to the ideas of co-occurrence or addition (it does not imply uncertainty):

  E22 !Tras limpieza quirúrgica se asocia al tto con antifúngicos
  After surgical cleaning, [the patient] is given/?is associated to antifungal treatment

Another aspect to take into consideration is the interaction between negative and uncertainty cues in the same sentence. Seemingly negative cues may express uncertainty depending on the context they appear in. For example, a negated negative cue might be used to express uncertainty (E23), while words that express confidence are also classified as uncertainty when they are negated (E24 and E25).

E23 **No se descarta** *{definitivamente} sangrado activo*
Active bleeding is not definitively ruled out

E24 **No claro** *transtorno sensitivo*
No clear sensitive disorder

E25 **Sin** *signos* **claros** *de isquemia aguda*
No clear signs of severe ischemia

Finally, subordinate interrogative clauses licensed by—possibly negated—verbs of knowing, thinking and believing also express doubt or hypothetical ideation, as in Examples E26 and E27:

E26 **No pudiendo precisar si** *ha presentado* **o no** *pérdida de conciencia*
[The patient] is not able to specify whether they lost consciousness or not

E27 Sugerimos una valoración psiquiatrica, **por si** *el origen del cuadro {pudiera} [...]*
We suggest a psychiatric evaluation, in case the origin of symptoms could [...]

10.3.1.3.3 Uncertainty cue exceptions As with negation cues, there exists occurrences of words or phrases typically annotated as uncertainty cues that should not be labelled under certain circumstances. The main exception is where the uncertainty is cancelled by a negation cue. For example, in E28, ‘sugestiva de’ stops indicating that the speaker is unsure of what they say when it is negated by ‘no’. In such cases, the uncertainty cue is not annotated, just the negation cue:

E28 **No** *clínica sugestiva de aura migrañosa*
No symptoms suggestive of migraine aura

This case contrasts with the following example (and E38 below), where the negative word ‘no’ does not cancel the uncertainty conveyed by ‘parecer’ (*to seem*), thus the two words are jointly annotated as an uncertainty cue:

E29 **No parece** *haber tenido TCE*
Does not appear to have had TBI

10.3.1.4 Scopes

Generally speaking, the scope is the part of the sentence that is affected by a negation or uncertainty cue; more specifically, cues have scope over the constituents of the sentence whose status being false or uncertain is sufficient to establish the truth of the sentence (see Huddleston et al. (2002), among others, for a comprehensive explanation regarding negation—here, we stretched their definition to include scopes of uncertainty cues).

Here, we follow IULA-SCRC’s definition of the scope as “the maximal syntactic unit that is affected by the marker” (Marimon et al., 2017a, p. 46) ignoring the subject (only included when in post-verbal position). Also following IULA-SCRC, cues are not part of scopes, as has been illustrated in all the examples above.

In what follows, we present several phenomena related to the scopes. First, we introduce two types of scopes that deviate from the canonical shape of scopes: discontinuous scopes and embedded scopes. Then, we introduce two new annotation categories that are only annotated within scopes: entities and negative polarity-sensitive items (NPI).

10.3.1.4.1 Discontinuous scopes The scope of a cue can sometimes be discontinuous. That is, a cue can affect multiple text spans that are separated by unaffected material. The most frequent structures that trigger discontinuous scopes are the following:

a) The cue occurs between the head and the complements or modifiers of the phrase or clause it affects, causing the cue to be surrounded by its scope (see also E25):

   E30 *Relación* **probable** *con incipientes cambios por otitis media crónica*
   Probable relation to early changes caused by chronic otitis media

b) The cue affects anaphoric expressions. In E31, ‘inhalers’ is the antecedent of the anaphor ‘them’, which is in the scope of the negation cue ‘not’; thus ‘inhalers’ too is annotated as being part of the scope:

E31 Refiere su Médico de Cabecera que le pautó *inhaladores* pero **no** *los tolera*
Her family doctor refers that she gave him inhalers but he does not tolerate them

Similary, the antecedent of the relative pronoun 'which' in E32, 'micturition symptoms', is part of the scope of the cue 'not':

E32 Refiere *clínica miccional* [...] que **no** *consultó {ni} trató*
[The patient] refers to micturition symptoms which they did not consult nor treat.

In Example E33, the verb 'repeats' is omitted in the non-initial coordinated clause, forming a gapped coordination; thus, the first mention of the omitted material is annotated as being part of the scope:

E33 *Repite* palabras sencillas pero **no** *frases*
[The patient] repeats simple words but not sentences

c) The cue is or contains a correlative conjunction, in which case both the cue and the scope are discontinuous (see also E26):

E34 Valorar **si** *precisa* **o no** *tratamiento antibiótico*
Assess whether or not [the patient] needs antibiotic treatment

10.3.1.4.2 Embedded scopes Up to this point, the examples given have only included one—continuous or discontinuous—cue and its scope. It is possible, however, to have a cue-scope pair embedded within another scope, as illustrated by Examples E35 and E36. The dotted underlines are the embedded scopes:

E35 **Sospecha de** *{posible} HSA* ***no*** *apreciada en el TAC*
Suspicion of a possible subarachnoid hemorrhage not detected in the CT

E36 **Imposibilidad para** *una bipedestación* ***sin*** *ayuda*
Inability to stand without help

In these cases, the two cues are semantically independent from each other and are annotated as such. Notice, however, that at least 3 special cases have been described throughout the previous sections where seemingly co-occurring cues are not annotated as two independent cues with independent scopes:

1. Negated certainty may indicate uncertainty (see Example E23);
2. the co-occurrence of negation and uncertainty may express just uncertainty (E29 and E38) or annul it (E28); and,
3. non-initial instances of cues of the same type as the initial cue may be treated as negative polarity-sensitive items (see Section 10.3.1.4.4).

10.3.1.4.3 Entities Scopes may contain mentions to medical entities that could be of interest for applications or application functionalities developed with NUBes, as entities constitute information units more easily understood and manageable by computers than scopes. Structurally, entities are light nominal phrases within scopes (underlined):

E37 **No** *se aprecian lesiones estructurales*
No structural lesions are observed

When a sentence contains coordinated phrases, each of them is annotated as an individual entity within a longer scope, as in Example E38. However, the whole constituent is annotated as entity when it is the modifiers or complements of the nominal head that are coordinated (E39):

E38 **Sin aparente** *TCE {ni} focalidad*
With no apparent TBI or [neurological] focus

E39 **No** *clínica digestiva {ni} miccional*
No digestive nor voiding symptoms

Only the most relevant entity (or coordinated entities) is annotated, that which conveys new information. While it might be tempting to think of these entities as the *foci* of negation or uncertainty, we refrain from using the term in this work, because *a)* foci come in many forms and shapes while, as mentioned earlier, entities are generally light nominal phrases; and, most importantly, *b)* it is not always possible to infer the intended focus of the speaker from written utterances.

For instance, E40 contains 2 medical entities that could theoretically play focus of the sentence, namely, 'metastatic lesions' and 'adrenal glands'; the focus might even be the heavier phrase 'metastatic lesions in adrenal glands' (too heavy perhaps to be considered an entity):

E40 **Sospecha de** *lesiones metastásicas en glándulas suprarrenales*
Suspected metastatic lesions in adrenal glands

Such examples must be interpreted and assessed in context. In E40, we would annotate 'metastatic lesions' as an entity instead of 'adrenal glands' because, intuitively, it is understood that the clinically most relevant, new and impactful information is that "the patient may have metastasic lesions (in their adrenal glands)" rather than "the metastasic lesions that the patient may have would be located in their adrenal glands". While intuitive, entities are admittedly the most difficult annotated pieces of information for which to provide rigorous criteria. Nevertheless, they are secondary to negation or uncertainty cues and scopes, which is what NUBes is primarily about.

Entities are labelled with a set of categories adapted from IULA-SCRC's interpretation of the SNOMED CT classification: Medical Findings and Disorders, Medical Procedures, Chemicals and Body Substances, Body Structure, Other—for other types of medical concepts; not in IULA-SCRC—and Phrase—used for entities not specific to the medical field. If the entity and the scope within which it lies match in span (as in E24 and E39, among others), the most specific label is used for the whole scope. Otherwise (e.g., E18, E28 and E37), the entity or entities are embedded within a Phrase scope.

10.3.1.4.4 Negative polarity-sensitive items (NPI) NPIs are lexical elements that are only licensed under specific conditions, negation being the quintessential licensor as the name 'negative polarity-sensitive item' suggests. In NUBES, the most frequent NPIs are pronouns or negative determiners, such as 'alguna' or 'ninguna' (*any*):

E41 **Niega** *dolor a {ningún} nivel*
[The patient] denies pain at any level

In examples like E41, NPIs seem to reinforce the expressive power of the negation cues that license them. From this perspective, we also label as polarity items cues of the same category that appear in the same sentence if they were used to reinforce the initial cue, as in the following example, even though they are not actual NPIs in a strict sense[1]:

E42 **Parece** *detectarse un {posible} deterioro cognitivo de {posible} origen vascular*
A possible cognitive impairment of possible vascular origin has seemingly been detected

10.3.2 Inter-annotator agreement

We report agreement measured as Cohen's kappa coefficient (κ) (Cohen, 1960) and agreement percentage (%). κ is defined as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{f_o - f_e}{N - f_e} \tag{10.1}$$

where p_o (resp. f_o) is the proportion (resp. frequency) of units in which the annotators agree—i.e., the *observed agreement*—and p_e (resp. f_e) is the proportion (resp. frequency) in which agreement is expected by chance—i.e., the *chance agreement*. Chance agreement is the sum of the joint probabilities of the marginal proportions. N is the total number of units annotated. In our case, $N = 43,060$, the tokens of the first batch. Agreement percentage (%) is simply p_o presented as a percentage.

Intuitively, κ tells how much the annotators agree beyond the expected agreement if annotations were random. There is no universally accepted interpretation

[1]None of the words labelled as polarity items in Example E42 requires licensing from the cue. Consider the sentence where the initial cue has been removed and is still perfectly grammatical:

- Se detecta un posible deterioro cognitivo de posible origen vascular
  A possible cognitive impairment of possible vascular origin has been detected

The point is that, overall, these words serve to strengthen the conveyed level of uncertainty instead of constituting independent cue-scope pairs; thus, we annotate them with the same label as the NPIs for the sake of simplicity, given that they produce a similar semantic effect.

of κ as to what is considered high or low agreement. Landis et al. (1977) proposed the interpretation shown in Table 10.1, which is widely cited, but has no evidential grounding.

We computed our inter-annotator agreement twice on the first batch of the corpus, before and after the discussion that led to the final guideline annotations. As Table 10.2 shows, agreement improved after the discussion, particularly for cues. The low agreement in polarity items is explained by the fact that they occur very few times (15) and the number of possible tags is also small (2; a token is either part of a polarity item or it is not), which distorts the κ measurement.

Table 10.1: Cohen's kappa coefficient (κ) interpretation by Landis et al. (1977)

Value	Meaning
$\kappa < 0.00$	No agreement
$0.00 \leq \kappa \leq 0.20$	Slight agreement
$0.21 \leq \kappa \leq 0.40$	Fair agreement
$0.41 \leq \kappa \leq 0.60$	Moderate agreement
$0.61 \leq \kappa \leq 0.80$	Substantial agreement
$0.81 \leq \kappa \leq 1.00$	Almost perfect or perfect agreement

Table 10.2: κ and agreement percentage (%) between 2 annotators on the first batch (2,971 sentences). N is the number of categories considered. The best results are highlighted in bold.

		Round 1		Round 2	
	N	κ	%	κ	%
Negation cue	4	0.85	99.43	**0.93**	**99.74**
Uncertainty cue	3	0.75	99.74	**0.84**	**99.84**
Scope	6	0.75	96.57	**0.80**	**97.17**
Entity	6	0.74	97.47	**0.80**	**98.04**
NPI	2	0.45	99.94	**0.50**	**99.95**
All	14	0.78	95.96	**0.83**	**96.83**

10.3.3 The NUBes corpus

NUBes consists of 29,682 sentences, out of which 24.59% include negation and 7.51% include uncertainty (see Table 10.3). In many of the sentences there is more than one cue. Further, while it is more common for the two phenomena to occur independently, they appear together in a small percentage of sentences (2.26%). Discontinuous cues and scopes seem to be much more frequent for uncertainty than for negation. Concerning the different cues that appear in the corpus, 345 unique negation and 297 unique uncertainty cues have been annotated. The most frequent cues by type are listed in Tables 10.4 and 10.5. Appendix C shows the distribution by medical speciality and Electronic Health Record (EHR) section.

Table 10.3: Quantitative description of NUBEs

	Negation	Speculation	Total
Sentences			29,682
Tokens			518,068
Vocabulary size			31,698
Sentences affected	7,298 (24.59%)	2,229 (7.51%)	8,855 (29.83%)
Average cues per affected sentence	1.29 ± 0.70	1.11 ± 0.37	1.35 ± 0.75
Total cues	9,431	2,480	11,911
Unique cues	345	297	634
Discontinuous cues	0	95	95
Average scope length in tokens	4.01 ± 3.59	5.27 ± 4.97	4.30 ± 3.98
Discontinuous scopes	219	123	342

Table 10.4: Top 5 negation cues by type (lemmatised and normalised)

NSyn		NLex		NMph	
Cue	**#**	**Cue**	**#**	**Cue**	**#**
no (*no, not*)	4,058	negativo (*negative*)	305	asintomático (*asymptomatic*)	443
sin (*without*)	2,518	retirar (*remove*)	290	afebril (*afrebile*)	252
tampoco (*neither*)	40	suspender (*withhold*)	180	desorientado (*disoriented*)	72
nunca (*never*)	5	negar (*deny*)	87	inespecífico (*non-specific*)	63
excepto (*except*)	4	descartar (*rule out*)	76	inestabilidad (*instability*)	24

Table 10.5: Top 5 speculation cues by type (lemmatised and normalised)

USyn		ULex	
Cue	**#**	**Cue**	**#**
versus, vs	15	probable	364
o (*or*)	4	compatible con (*compatible with*)	255
		posible (*possible*)	216
		parecer (*to seem*)	156
		sospecha de (*suspicion of*)	143

10.3.4 Differences with related corpora

The most basic step in the process of creating the corpus consisted in attempting to reach an agreement on what the terms negation and uncertainty encompass.

An overview of the existing literature both in English and Spanish, revealed the there is no one main, agreed-upon definition of these phenomena, not only across the disciplines of theoretical and computational linguistics, but even across corpus descriptions generated within the NLP community. The main differences between them have to do with what is accepted as negation and the way in which elements such as scopes are annotated.

We ultimately considered that our definition of negation should encompass every word that implies an entity is not occurring or has not occurred—either at all ('imposibilidad para' [*impossibility to*]) or anymore ('retirada de' [*removal of*], 'suspender' [*withhold*]). Marimon et al. (2017a) among others argue that they did not take into account these type of cues because they express a "change of state" (Marimon et al., 2017a) or, in the case of 'negar' (*deny*), that it "is considered, in factual terms, an statement of what someone says". From the point of view of the applicability of the corpus, we still considered interesting to annotate negation and uncertainty in reported speech.

Another debatable example is the postnominal adjective 'negativo' (*negative*). The authors of UHU-HUVR (Cruz Díaz et al., 2017) only annotate this word for test results whenever the name of the test and that of the condition is the same, as it means that the patient does not have said condition; otherwise, it means that the test has taken place and that it is the results that are negative. This contrast is shown respectively in examples (E43) and (E44), taken from UHU-HUVR.

E43 Serología materna: [Toxoplasma]: Negativo
Maternal serology: Toxoplasma: Negative

E44 Técnicas de Z-N (normal y largo) negativo
Negative Z-N stain (normal and long)

In NUBES, the latter case (E44) is also annotated as it still accommodates into our definition of negation.

Finally, some of the instances that are categorised as negation by other corpora were annotated as uncertainty in NUBES due to the inclusion of this phenomenon. For example, given the sequence 'sin clara' (*no clear*), IULA-SCRC annotates 'sin' as a cue and 'clara' as part of the scope. In NUBES, 'sin clara' as a whole is considered an uncertainty cue, as illustrated several times throughout the guidelines.

All things considered, it must be noted that each set of guidelines is the product of a long, challenging debate not free of hesitation—even after a consensus is reached among the authors—, and highly influenced by the applications that the authors might have in mind for the corpus.

10.4 Conclusions

In this chapter we have presented the NUBES corpus, a new collection of biomedical texts in Spanish annotated for negation and uncertainty. It is publicly available in a GitHub repository [6]. To the best of our knowledge, NUBES is the largest public corpus of clinical reports in Spanish annotated with negation and the first one that includes the annotation of speculation cues, scopes, and entities. Table 10.6 offers a comparison of NUBES with related existing corpora in quantifiable terms.

Table 10.6: Spanish biomedical corpora with annotations of negation and/or speculation, including NUBES, adapted from Jiménez-Zafra et al. (2018b) and Martí et al. (2018). The upper table section describes the corpora qualitatively, in terms of the types of annotations they contain; the middle table section describes the corpora quantitatively. [1] 27.58% of the diseases annotated are negated. [2] 1.90% of the diseases annotated are speculative. [3] 513 radiology reports. [4] 56% of the findings are negated.

	IxaMed-GSC	UHU-HUVR	IULA-SCRC	Cotik et al. (2017)	E3C	NUBES
Negation cue		✓	✓	✓		✓
Speculation cue				✓		✓
Scope		✓	✓			✓
Entity	✓	✓	✓	✓	✓	✓
Event					✓	
Total sentences	5,410	8,412	3,194	?[3]	1,134	29,682
w/ negation (%)	?[1]	2,298 (27.32)	1,093 (34.22)	?[4]	240 (21.16)	7,298 (24.59)
w/ speculation (%)	?[2]	-	-	?	114 (10.05)	2,229 (7.51)
Available at	-	-	[62]	-	[45]	[6]

We have explored the corpus from different perspectives: by its comparison with similar corpora, by justifying its design and by acknowledging its limitations. Annotating a corpus with extra-propositional meaning requires a thorough linguistic analysis that led to many discussions before, during and even after the process. Aspects like how to demarcate the definition of negation and uncertainty and whether some examples were actually part of them proved to be a source of disagreement. On top of that, the idiosyncrasies of medical language also posed some complications.

In the next chapters, we exploit NUBES in several experiments about automatically detecting negation and uncertainty. In Chapter 11, we model the problem as a sequence labelling task of 4 types of spans: negation cues, uncertainty cues, negation scopes, and uncertainty scopes. In Chapter 12, we address the problem of assertion classification; to be able to do so with NUBES, we transform the corpus automatically.

Chapter 11

Negation and speculation: experiments in cue and scope detection

11.1 Introduction

The study builds on Lima-López et al. (2020a), who present the first experiments with NUBes of detecting negation and speculation cues and scopes. In that work, we train biLSTM + CRF models that exploit a combination of lexical, syntactic and semantic features. Here, we evaluate a diverse set of Transformer (Vaswani et al., 2017) and Flair (Akbik et al., 2019) models, managing to improve our previous reported results, as well as the related work (Solarte Pabón et al., 2022). Furthermore, we analyse the performance of said model in a range of scenarios of varying difficulty:

- In addition to the overall performance a given model may yield, being able to achieve competitive results with as little data as possible is a most desirable trait, given that clinical data is notably hard to obtain. For this reason, we analyse the performance of the models with decreasing amounts of training data, from thousands of examples down to a few dozen.
- It has been widely reported that a few negation markers (e.g., 'no' and 'sin') are responsible for most of the negation instances in Spanish free text (Moreno Sandoval et al., 2013; Campillos Llanos et al., 2017; Cruz Díaz et al., 2017; Lima-López et al., 2020a). While previous studies on negation and uncertainty detection report overall acceptable results in multiple scenarios and datasets, it has not been studied how well predictive models perform specifically on the less frequent surface forms of negation, which are equally important in real usage scenarios.

The remainder of the chapter is structured as follows: Section 11.2 first describes the form and quantity of the data used in the experiments; then, it presents the trained and evaluated systems; finally, it explains the evaluation methodology. Section 11.3 reports the results of the evaluation and their analysis. Last, Section 11.4 summarises the chapter and presents the conclusions drawn from the presented work.

11.2 Materials and methods

11.2.1 Data

The experiments are conducted with the NUBes corpus (Chapter 10). It consists of a collection of sentences extracted from anonymous Spanish clinical records and manually annotated with negation and uncertainty cues and their scopes. For this set of experiments, we keep the train, development and testing splits of the NUBes corpus first presented in Lima-López et al. (2020a) [6], which already come converted from brat standoff format (Stenetorp et al., 2012) to token-level annotations with 4 types of entities:

- `NCue`: negation cue,
- `NSco`: negation scope,
- `UCue`: uncertainty cue, and
- `USco`: uncertainty scope.

The labelling scheme chosen for this task was BIO, in which `B-` (Beginning) marks the beginning of a entity or span, while the subsequent tokens of the span receive the tag `I-` (Inner) and tokens that do not belong to any span are marked with `O` (Outside). The sentences of Figure 9.1 would be encoded as follows:

E1 From Figure 9.1a:

CyC .... `O`
: .... `O`
Rigidez .... `O`
de .... `O`
nuca .... `O`
, ... `O`
no .... `B-NCue`
ingurgitación . `B-NSco`
yugular .... `I-NSco`
. ... `O`

E2 From Figure 9.1b:

Los .... `O`
hallazgos .... `O`
descritos .... `O`
son .... `O`
sugestivos .... `B-UCue`
de .... `I-UCue`
pielonefritis .. `B-USco`
aguda .... `I-USco`
. ... `O`

E3 From Figure 9.1c:

Tumoraciones ...... `O`
faciales .... `O`
en .... `O`
paciente .... `O`
transplantada ...... `O`
hepatica .... `O`

The total size of each data split can be consulted in Table 11.1. To compute the **train curves**, we created increasingly smaller training data subsets by randomly extracting 1/3 of the examples in 5 iterations, for a total of 6 decremental training datasets. To create the **difficult or adversarial test data set**, Adv, we remove from the original test data set, Full, the examples that contain frequent negation or uncertainty markers. We consider frequent markers any marker with relative frequency in the training set higher than 2%, which together constitute 62.11% of the markers (see Table 11.2). That is, ADV is a subset of FULL.

As Table 11.1 shows, negation instances are more than thrice more likely to occur than uncertainty in this corpus; furthermore, uncertainty markers are lexically more variable, as evidenced by the smaller drop from the regular to the difficult test set in comparison to negation.

Table 11.1: Size of the corpus for the cue and scope detection task

	Train						Dev	Test	
	1/1	1/3	$1/3^2$	$1/3^3$	$1/3^4$	$1/3^5$		FULL	ADV
Total sentences	13,802	4,600	1,533	510	169	56	1,840	2,762	1,838
w/ negation	5,265	1,761	576	210	78	24	694	1,041	240
w/ uncertainty	1,272	386	127	44	16	6	162	249	206
w/ both	364	127	53	16	4	1	64	91	11
Total spans	17,107	5,648	1,906	657	236	83	2,289	3,545	998
Negation cue (`NCue`)	6,976	2,337	775	273	97	31	919	1,423	265
Negation scope (`NSco`)	6,379	2,135	708	251	91	31	847	1,322	233
Uncertainty cue (`UCue`)	1,866	586	212	67	24	11	263	400	251
Uncertainty scope (`USco`)	1,886	590	211	66	24	10	260	400	249

Table 11.2: Cues with relative frequency > 2% on the train set

Cue	Type	#	%	C%
no	Negation	3,046	34.35	34.35
sin	Negation	1,820	20.53	54.88
probable	Speculation	264	2.98	57.86
afebril	Negation	190	2.14	60.00
asintomático	Negation	187	2.11	62.11

11.2.2 Systems

In this chapter, the task is framed as a sequence labelling problem. All the systems in this experiment approach the problem as a single task, that is, they learn to detect jointly the 4 span types, emitting for each input token one of the 9 labels defined for the task (see Section 11.2.1). We have tested 3 such neural sequence labelling frameworks:

11.2.2.1 Baseline

The baseline for this experiment was set in Lima-López et al. (2020a) with the NCRF++ (J. Yang et al., 2018b) sequence tagger. The system consists of a Convolutional Neural Network (CNN) layer for character sequence representations, followed by a biLSTM layer for word sequence representations, and an output CRF layer. The character and word embeddings are initialised randomly and trained on the given corpus. Here, we report the results of the best variant tested in Lima-López et al. (2020a), which exploits a set of lexical and morpho-syntactic features automatically extracted from the input text.

11.2.2.2 Flair

Flair is a NLP Python framework (Akbik et al., 2019) that features a specific type of character-based contextualised word embeddings of the same name (Akbik et al., 2018). Here, we train Flair's sequence tagger, which is a more sophisticated biLSTM + CRF sequence tagger and is broadly schematised in Figure 11.1.

The input embedding mechanism combines Flair's pre-trained embeddings for Spanish (`es-forward` and `es-backward`) and the fastText embeddings (Bojanowski et al., 2017) Medical Word Embeddings for Spanish or MWES (Soares et al., 2019b). Specifically, we use the v2.0 skipgram embeddings trained on uncased SciELO and Wikipedia documents [66]. Both sets of embeddings are updated during training.

In short, the core differences of this system with the baseline are that *a*) it uses pre-trained contextual character embeddings instead of static embeddings trained from scratch, and *b*) it starts off with some language and domain knowledge thanks to said pre-trained embeddings.

11.2.2.3 Transformer

The bulk of the experimentation involves Transformer (Vaswani et al., 2017) models. We have tested a diverse set of BERT- (Devlin et al., 2019) and RoBERTa-like (Y. Liu et al., 2019) pre-trained language models, both monolingual and multilingual, as well as general-purpose and domain-specific. The complete list of the tested pre-trained models can be consulted in Table 11.3.

The architecture is the same for all the system variants: first, the input to the BERT encoder is prepared according to the standard procedure; we specifically follow the same steps as those described in Chapter 4, Section 4.2.2.4, for the sensitive data BERT-based tagger. The prepared input is then passed to the encoder, which is followed by a dropout layer and one classification head consisting of a linear transformation layer that emits the logits per token for the 9 output categories. In inference, the label with the maximum probability is chosen for each token after applying the softmax function to the logits. Figure 11.2 shows a simplified diagram of the inference pipeline.

The models are trained on the cross-entropy loss of the classification head over the first subword of each input token. Subwords in suffix positions are ignored,

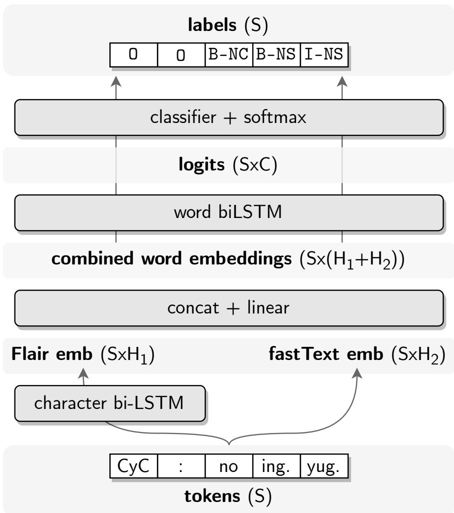


Figure 11.1: Diagram of the Flair-based cue and scope tagger. S (sequence length); H_1 = 128 or 256 (Flair embedding size); H_2 = 300 (fastText embedding size); C = 9 (number of output labels).

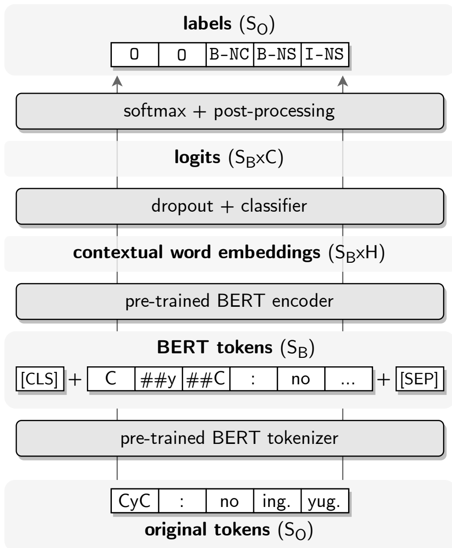


Figure 11.2: Diagram of the BERT-based cue and scope tagger. S_O (original sequence length); S_B = 220 (sequence length after BERT tokenisation); H = 768 (BERT embedding size); C = 9 (number of output labels).

that is, the output label of the input tokens is assigned from the prediction for the first corresponding subword.

Table 11.3: Pre-trained language models tested in the experimentation. References of each mentioned resource can be consulted in Table 2.4 of Chapter 2. See Appendix E for a report of the vocabulary overlap of these models with the vocabulary of the NUBes corpus.

	Lang	Corpus	Param	Vocab
BERTs				
BETO_{Base} Cased	es	Spanish Unannotated Corpora	110M	31K
mBERT_{Base} Cased	multi	Wikipedia	178M	120K
IXAmBERT_{Base} Cased	es, en, eu	Wikipedia	178M	119K
$\text{SciBERT}_{scivocab}$ Cased	en	Semantic Scholar	110M	31K
RoBERTas				
SpanBERTa_{Base} Cased	es	OSCAR	125M	50K
MarIA RoBERTa_{Base} BNE	es	BNE selective crawls	125M	50K
$\text{XLM-RoBERTa}_{Base}$	multi	Common Crawl	278M	250K

11.2.2.4 Implementation and training setup

We have optimised some hyperparameters of the Transformer variants and Flair in each data subset with 25 trials each, for a total of 1,200 models (that is, 8 optimised systems on 6 training datasets for 25 trials) in addition to the baseline. For each system and training set, the trial with the best F1-score (see Section 11.2.3) on the development data set has been chosen to compute the results on the testing data sets.

The Transformer models have been implemented with HugginFace's `transformers` Python library (Wolf et al., 2020), and optimised using Ray's `tune` Python library (Liaw et al., 2018). In the case of Flair, the Python library comes with a wrapper of Hyperopt (Bergstra et al., 2013) for hyperparamenter optimisation [67]. The hyperparameter search spaces are given in Appendix F.

As for the baseline system, the NCRF_{++} tagger is the same as that described by Lima-López et al. (2020a). Appendix F also reports its hyperparemeter setup.

11.2.3 Evaluation

The results of this chapter are again evaluated in terms of micro-average F1-score ($F1$), the harmonic mean of Precision (P) and Recall (R), repeated here for convenience:

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (8.1\ (=4.1))$$

We report **strict span-level metrics** as computed by the Python library `seqeval` [68]. To that end, the token-level predictions are converted to span-level predictions, that is, the BIO tags are interpreted to obtain predictions consisting of a span boundaries (offset and end) and the predicted category for the span. Then, the script counts true positives (TP), false positives (FP) and false negatives (FN) per category $c \in \{\texttt{NCue}, \texttt{NSco}, \texttt{UCue}, \texttt{USco}\}$ as follows:

- TP: number of predicted spans of type c that match exactly in boundaries with a gold span of type c.
- FP: number of predicted spans of type c that do not match exactly in boundaries with any gold span or that match with a gold span of a type other than c.
- FN: number of gold spans of type c that do not match exactly in boundaries with any prediction or that match with a prediction of a type other than c.

As average metrics of the different categories, we report micro-average (μ) scores. The micro-average scores are obtained by applying the same equations to the sums of the TP, FP and FN of the different categories.

This is the strictest evaluation methodology possible for this task. In order to be able to compare the results with the related work, Appendix G reports the performances of the trained sequence labelling systems following two additional evaluation methodologies, namely *SEM 2012 scores (Morante et al., 2012a) and BIO-weighted token-level scores (Solarte Pabón et al., 2022). We refer the reader to the corresponding literature for detailed explanations of these metrics.

11.3 Results

11.3.1 Cue and scope detection

Table 11.4 reports per-category and micro-average F_1-score results of models trained in the full train set and one of the train subsets (with ~1% of examples). Other metrics, including *SEM 2012 metrics, can be consulted in Appendix G.

Overall, we observe that the detection of cues (`NCue` and `UCue`) is easier than that of scopes (`NSco` and `USco`), and that speculation (`UCue` and `USco`) is more difficult to detected than negation (`NCue` and `NSco`). This is to be expected given the nature and distribution of each category, and was also noted by Lima-López et al. (2020a).

Regarding the differences among the systems trained on the full dataset, little difference among the Transformers is noted, although MarIA stands out with

Table 11.4: F_1-score results for cue and scope detection in the Full test set. The best and second-best scores are highlighted in bold and dotted underlines, respectively. N is the number of training examples.

	1/3^4 train set (N=169)					Full train set (N=13,802)				
	μ	NCue	NSco	UCue	USco	μ	NCue	NSco	UCue	USco
NCRF++	0.604	0.770	0.626	0.093	0.088	0.881	0.952	0.866	0.849	0.698
Flair+fT	0.690	0.851	0.685	0.434	0.218	0.892	0.960	0.877	0.849	0.740
BETO	**0.735**	0.861	0.728	**0.616**	0.320	0.905	0.963	**0.900**	0.870	0.759
SpanBERTa	0.691	0.865	0.650	0.537	0.207	0.898	0.960	0.894	0.850	0.743
MarIA	0.708	0.855	0.699	0.529	0.283	**0.910**	**0.968**	0.897	**0.875**	**0.781**
IXAmBERT	0.730	0.854	**0.736**	0.609	0.322	0.901	0.965	0.888	0.865	0.755
mBERT	0.714	**0.866**	0.701	0.567	0.254	0.898	0.960	0.887	0.851	0.760
XLM-R	0.730	0.864	0.726	0.577	**0.324**	0.905	0.962	0.896	0.863	0.780
SciBERT	0.678	0.859	0.642	0.502	0.113	0.890	0.959	0.868	0.861	0.750

an average F_1-score of 0.910, followed by BETO and XLM-RoBERTa (hereafter XLM-R)—both 0.905—. MarIA and XLM-R in particular achieve the greatest gains with respect to the uncertainty scope (USco) scores of the baseline set by NCRF++ , which presented the biggest opportunity for improvement in previous work. Unsurprisingly, SciBERT falls behind the other Transformers, but its performance is similar to Flair's. Still, both improve the baseline across all categories and manage to overpass prior state of the art (Solarte Pabón et al., 2022, see Table G.4 in Appendix G).

Looking at the performance of the models with the smaller train set, we see very significant gains of the Transformer models and Flair with respect to the baseline, particularly for uncertainty cues and scopes (UCue and USco respectively). It is remarkable that with only 169 examples of training, all the Transformer models yield F_1-scores above 0.5 in the detection of uncertainty cues. It is noteworthy as well that the models that fare best with this smaller training set, BETO and IXAmBERT, are not the ones that achieve the best results when presented with the full training set. The behaviour of the models with increasing amounts of training data will be analysed in greater depth in the next section.

11.3.2 Train curves and adversarial examples

Figure 11.3 shows the training curves of the 9 compared systems. These train curves have been generated by training each model with the increasing training samples and evaluating the resulting models in the two testing sets—Full and Adv, from "adversarial" or difficult. The difficult test set is a subset of the full test set that contains only examples with the least frequent negation and uncertainty cues (see Section 11.2.1). We chose to report the curves for negation and uncertainty scope detection (`NSco` and `USco`), seeing that they are the most difficult spans to detect correctly.

The baseline NCRF$_{++}$ shows the biggest gap between the scores for negation in the FULL test set and the rest of the scores along the whole curve, which evinces the poorer capability of generalisation in comparison to the Transformer models and Flair.

All the systems except the baseline surpass the 0.8 F_1-score points for negation scope (`NSco`) detection in the FULL test set with $1/9^{th}$ of the training set (1,533 examples), and reach or nearly hit 0.9 F_1-score points with all the available data (13,802).

It is striking that some models—namely, SpanBERTa, MarIA, mBERT and especially IXAmBERT—set off with great advantage over the rest of the models where negation detection is concerned, although when looking at the scores for the most difficult examples, it becomes evident that all they are doing in practice is detecting the words ‘no’ and ‘sin’ (*without*). Given more data, the other Transformers are able of catching up.

Finally, most models (NCRF$_{++}$, Flair, SciBERT and SpanBERTa most markedly) show an upwards trend still towards the end of the curve, which indicates they might be able to reach the results of the best models if given more data.

11.3.3 Error analysis

We conclude the inspection of the results with an error analysis, where we go over the confusion matrices of the compared systems (Table 11.5) and illustrate their most salient incorrect predictions. The matrices have been computed at token level ignoring the BIO tags. The values are presented in relative terms ignoring true positive `O` predictions (being the majority class, it would render the matrices uninformative). That is, each matrix adds up to 1.

As can be seen, the most frequent errors are false negative errors of scopes, both of negation and uncertainty. The baseline NCRF$_{++}$ is the system that commits this error more frequently, which accounts for ~11% of its predictions (again, not considering the true `O` tokens), while with BETO and XLM-R we manage to cut these errors by more than half. Still, the systems struggle to annotate scopes properly in the same contexts. We identified the following[1]:

1. Sentences with coordination:

   E4 **Ausencia de** *factores de riesgo vascular, cardiopatía etc, ..* (sic)
   Absence of vascular risk factors, heart disease, etc.

[1] The examples are formatted as presented in Section 10.3.1.1 of the previous chapter.

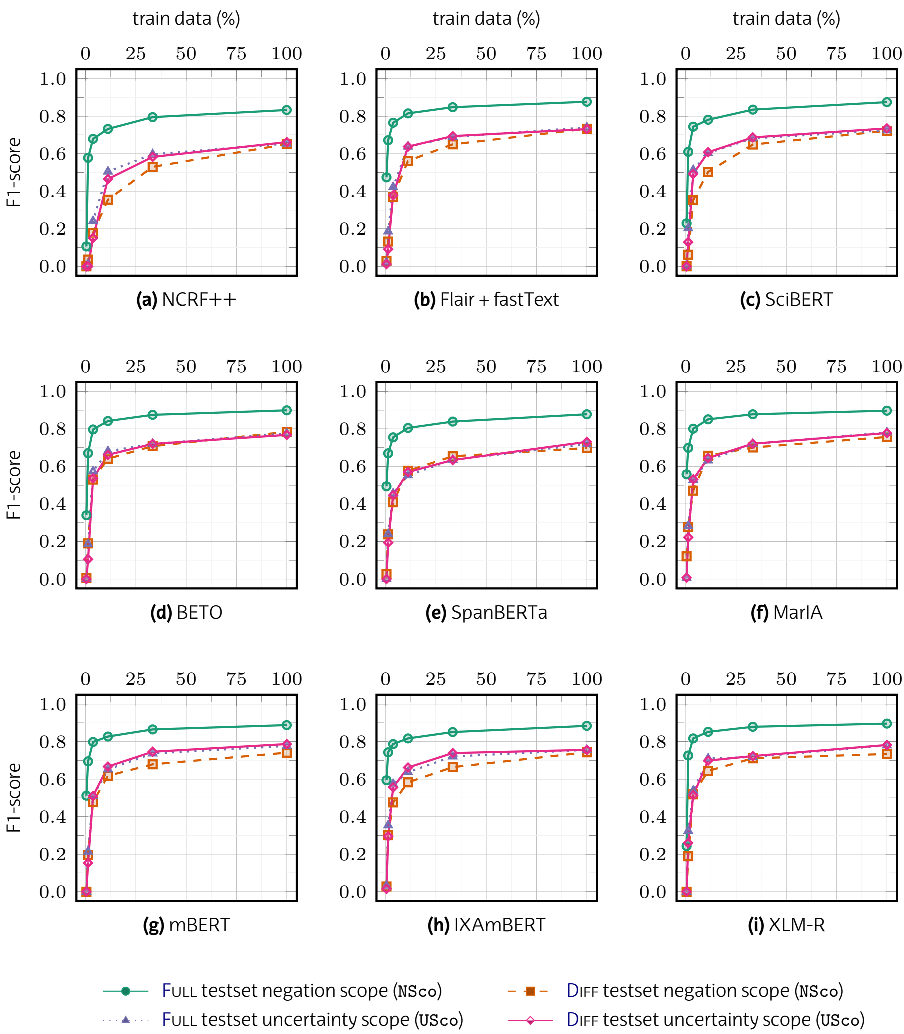


Figure 11.3: Train curves of the cue and scope detection task

Table 11.5: Confusion matrices of the cue and scope detection task; predictions made by the models trained on the entire training set for the Full test set. N is the number of true tokens for each category in absolute terms.

true	N	predicted: NCue	NSco	UCue	USco	O
NCue	1,570	0.15	0.00	0.00	0.00	0.01
NSco	5,196	0.00	0.47	0.00	0.01	0.06
UCue	597	0.00	0.00	0.05	0.00	0.01
USco	1,982	0.00	0.01	0.00	0.14	0.05
O	42K	0.00	0.03	0.00	0.01	

(a) NCRF++

true	N	predicted: NCue	NSco	UCue	USco	O
NCue	1,570	0.15	0.00	0.00	0.00	0.00
NSco	5,196	0.00	0.48	0.00	0.01	0.04
UCue	597	0.00	0.00	0.05	0.00	0.00
USco	1,982	0.00	0.01	0.00	0.16	0.03
O	42K	0.01	0.03	0.00	0.02	

(b) Flair + fastText

true	N	NCue	NSco	UCue	USco	O
NCue	1,570	0.15	0.00	0.00	0.00	0.00
NSco	5,196	0.00	0.49	0.00	0.01	0.03
UCue	597	0.00	0.00	0.05	0.00	0.00
USco	1,982	0.00	0.01	0.00	0.17	0.02
O	42K	0.01	0.02	0.00	0.03	

(c) BETO

true	N	NCue	NSco	UCue	USco	O
NCue	1,570	0.15	0.00	0.00	0.00	0.00
NSco	5,196	0.00	0.48	0.00	0.01	0.04
UCue	597	0.00	0.00	0.05	0.00	0.01
USco	1,982	0.00	0.00	0.00	0.16	0.03
O	42K	0.01	0.02	0.00	0.03	

(d) SpanBERTa

true	N	NCue	NSco	UCue	USco	O
NCue	1,570	0.15	0.00	0.00	0.00	0.00
NSco	5,196	0.00	0.49	0.00	0.01	0.04
UCue	597	0.00	0.00	0.05	0.00	0.00
USco	1,982	0.00	0.01	0.00	0.16	0.03
O	42K	0.00	0.02	0.00	0.02	

(e) MarIA

true	N	NCue	NSco	UCue	USco	O
NCue	1,570	0.15	0.00	0.00	0.00	0.00
NSco	5,196	0.00	0.48	0.00	0.01	0.04
UCue	597	0.00	0.00	0.05	0.00	0.00
USco	1,982	0.00	0.00	0.00	0.16	0.03
O	42K	0.01	0.02	0.00	0.03	

(f) mBERT

true	N	NCue	NSco	UCue	USco	O
NCue	1,570	0.15	0.00	0.00	0.00	0.00
NSco	5,196	0.00	0.49	0.00	0.01	0.04
UCue	597	0.00	0.00	0.05	0.00	0.01
USco	1,982	0.00	0.00	0.00	0.16	0.03
O	42K	0.01	0.02	0.00	0.02	

(g) IXAmBERT

true	N	NCue	NSco	UCue	USco	O
NCue	1,570	0.15	0.00	0.00	0.00	0.00
NSco	5,196	0.00	0.49	0.00	0.01	0.03
UCue	597	0.00	0.00	0.05	0.00	0.00
USco	1,982	0.00	0.00	0.00	0.17	0.02
O	42K	0.01	0.03	0.00	0.02	

(h) XLM-R

true	N	NCue	NSco	UCue	USco	O
NCue	1,570	0.15	0.00	0.00	0.00	0.00
NSco	5,196	0.00	0.48	0.00	0.01	0.03
UCue	597	0.00	0.00	0.05	0.00	0.00
USco	1,982	0.00	0.00	0.00	0.16	0.03
O	42K	0.01	0.04	0.00	0.03	

(i) SciBERT

E5 **Parece** *empeorar al apoyar la cabeza [...] y con ciertos movimientos del cuello*
It seems to be worse when resting the head [...] and with certain neck movements.

The systems may annotate only the initial coordinated phrase or clause as being part of the cue's scope. Moreover, the longer the coordinated items are, the more likely it is that the systems will miss out the non-initial items or parts of them.

2. Sentences with scopes preceding the cues:

E6 *Refiere en Agosto episodio algo semejante* **?** (sic)
[The patient] refers to a similar episode in August?

E7 [...] pensar en *componente psiquiátrico añadido* que **justificara** *las crisis*.
[...] think of an added psychiatric component that could justify the crises.

This type of examples involves mostly relative clauses, as in E7, and constitute ~3.5% of the training corpus. When presented with such input, the systems sometimes label as scopes only post-cue material.

3. Sentences with negation or uncertainty reinforcement through multiple markers:

E8 **Interpreto** *el cuadro clínico como {probable} pericarditis*.
I interpret the clinical picture as probable pericarditis.

E9 **Nos pareció** [...] *{sugestivo de} una encefalitis {o} meningoencefalitis*
We thought it was suggestive of encephalitis or meningoencephalitis

In this type of cases, the systems may annotate the nested cues and scopes, that is, they may overlook the outmost material of the negation or uncertainty expressions. However, in contrast to the two types of errors presented above, in this case the systems usually manage to include the entire focus of the negation or the uncertainty within their predicted scopes, which makes these errors less harmful. This type of error also contributes towards false negative predictions of cues.

Although to a lesser extent, the systems make false positive errors as well when it comes to the detection of scopes. The most common of these errors stems from the inability of the systems to recognise as separate syntactic constituents a phrase or clause affected by negation/uncertainty and a following adjunct, as are 'sobreinfectado' (*overinfected*) and 'en el lado derecho' (*on the right side*) in Example E10:

E10 Se observa hidrocele [...] **probablemente** *sobreinfectado* en el lado derecho.
Probably overinfected hydrocele [...] observed on the right side.

Even human annotators find these cases challenging, because the sentences may be syntactically ambiguous and must be interpreted mindfully to capture the intended meaning in the annotations.

Regarding cues, some false negative errors involve infrequent lexical expressions that the systems were not able to generalize. This is particularly the case for uncertainty cues. Here are a few examples undetected by the majority of the systems:

E11 **Hay que asumir** *que está infectada*
It must be presumed that she is infected

E12 Refiere haber ingerido lorazepam [...] con *ideación*, **al parecer**, *autolitica* (sic)
[The patient] refers having ingested lorazepam [...] with apparent suicidal ideation

Further, a minor source of false negative cue annotations are errors caused by factors unrelated to the systems themselves, and that have to do with the limitations of NUBES acknowledged in the previous chapter (see Section 10.2.3). First, a few expressions are inconsistently annotated throughout the corpus, such as the verb 'evitar' (*avoid*); the systems have learned *not* to interpret it as a negation cue, but it *is* annotated in the reference corpus in a minority of occurrences. Second, tokenization errors in sentences with ungrammatical usage of punctuation marks induce errors in the post-processing of the predicted labels, as only the prediction of the first subword is taken as final label for a word. Take the following example:

E13 Comenzar tolerancia oral.**Asintomática**. (sic)
Start oral tolerance. Asymptomatic.

While the systems may be able to detect properly that 'asintomática' (*asymptomatic*) is a negation cue, it will not be annotated as such because the word in the NUBES corpus is 'oral.Asintomática' (sic) and only the label produced for the first subword (e.g., 'oral') is taken to account to produce the final labels.

As mentioned earlier, sentences with cue reinforcement are also a source of false negative cue annotations (see Examples E8 and E9 and their explanation).

In this case, the Flair sequence labeller produces the least false negatives cues, missing out just 2% of the negation cues (`NCue`) and 6% of the uncertainty cues (`UCue`). NCRF++ is again the worst system, doubling the false prediction rates of Flair.

As for false positives predictions of cues, they actually stem for the most part from human errors, that is, these predictions capture cues overlooked by the human annotators. Interestingly, the error rates are inverted in this case, with NCRF++ committing the least false positives and XLM-R leading the rank, followed closely by SpanBERTa. Pending an example-by-example manual revision, it seems sound to assume, given the recall scores (Appendix G), that XLM-R and SpanBERTa are not committing actual errors but simply detecting more human errors of the type just explained than the rest of the systems.

Finally, there seems to be a slight confusion with some negation and speculation scopes among most systems: in ~1% of the tokens (ignoring true `O`), some systems emit the tag `USco` (uncertainty scope) when it should be `NSco` (negation scope). Upon manual analysis of these cases, we consider that the systems are actually not committing errors but again correcting what appear to be incorrect—or at best debatable—manual annotations, as exemplified in Table 11.6.

Table 11.6: Gold annotations and predictions on the sentence extract "unable to specify whether there was a loss of consciousness or not". The fact that the phrase contains what are typically negative cues ("unable to", "loss of") and that the uncertainty cue is discontinuous ("whether [...] or not") makes this example especially difficult to predict correctly. While the manual annotations interpret the phrase as a negation cue and scope, most of the systems (except Flair) retract their predictions midway in favour of speculation.

Token	Gold	NCRF++	MarIA	BETO	Flair
incapaz	B-NCue	B-NCue	B-NCue	B-NCue	B-NCue
de	I-NCue	I-NCue	I-NCue	I-NCue	I-NCue
precisar	B-NSco	I-UCue	B-USco	I-UCue	B-NSco
si	I-NSco	I-UCue	I-UCue	I-UCue	O
hubo	I-NSco	I-UCue	I-UCue	B-USco	O
o	I-NSco	I-UCue	B-UCue	B-UCue	O
no	I-NSco	I-UCue	I-UCue	I-UCue	B-NCue
perdida	I-NSco	B-USco	B-USco	B-USco	B-NSco
de	I-NSco	I-USco	I-USco	I-USco	I-NSco
conocimiento	I-NSco	I-USco	I-NSco	I-USco	I-NSco

11.4 Conclusions

In this chapter, we have evaluated multiple state-of-the-art models for sequence labelling in the tasks of negation and speculation cue and scope detection. The experiments have been conducted with NUBes, the corpus of health records written in Spanish product of the work described in Chapter 7. The evaluated systems include multiple BERT-like and RoBERTa-like Transformer-based models, Flair, and a RNN as baseline system.

The task of cue and scope detection was learned jointly by the systems. The Transformer-based labeller with the MarIA pre-trained language model (Gutiérrez-Fandiño et al., 2022) achieved the best overall results (0.91 micro-average F1-score), advancing the state of the art previously set by Lima-López et al. (2020a) and Solarte Pabón et al. (2022). The system is closely followed by most of the other Transformer-based models, while SciBERT and the Flair sequence labeller fall slightly behind (still improving the baseline). The improvement is brought predominantly by a better detection of speculation scopes as well as of the least frequent negation instances.

We also observed that neither the models with most vocabulary overlap with NUBes nor the biggest models obtained the best results, although they did follow closely MarIA. Further, the training curves showed that, while monolingual

Spanish models start off with certain advantage, being able to correctly emit predictions for the most frequent and repetitive instances, all the Transformer models manage to obtain similar results when allowed to exploit the entire training sets.

A manual error analysis revealed that the most common errors are false negative errors involving scopes, that is, the predicted scopes tend to fall short compared to the gold annotations. This is particularly true in sentences with coordination and in relative clauses, where part of a scope might precede its cue. The manual error analysis also uncovered several incorrectly annotated instances, which will help us improve the quality of the NUBes corpus.

Chapter 12

Negation and speculation: experiments in assertion classification

12.1 Introduction

Having dealt in the previous chapter with the detection of negation and uncertainty as a sequence labelling problem targeted at cues and scopes, this chapter studies perhaps the most commonplace way of modelling negation and uncertainty detection in the biomedical field: the text classification task known as *assertion classification*.

The presented experimentation follows the same methodology as that of the previous chapter, exploiting the NUBes corpus for training and testing a variety of Transformer (Vaswani et al., 2017) and Flair (Akbik et al., 2019) models. To that end, we transform automatically the NUBes corpus annotations: from cues and scopes to entities and their assertion category. To the best of our knowledge, this is the first work that studies the assertion classification of medical entities in Spanish clinical text.

The remainder of the chapter is structured as follows: Section 12.2 describes the process of transforming the NUBes corpus as wll as its results, both qualitatively and quantitatively; then, it presents the systems tested and explains the evaluation methodology. Section 12.3 reports the results of the evaluation and their analysis, including a manual error analysis. Last, Section 12.4 summarises the chapter and presents the conclusions drawn from the presented work.

12.2 Materials and methods

12.2.1 Data

In the task of assertion classification, an instance or example consists of the medical entity to be classified presented in its context. In our case, the entity of interest is marked with the HTML tag `<e></e>`. The categories of the task are the following:

- absent (`abs`): negated medical entity,
- possible (`pos`): uncertain medical entity, and
- present (`pre`): affirmed medical entity.

From the examples in Figure 9.2, we would get the following instances (the entities of interest are highlighted in boldface for convenience):

E1 CyC: <e>**Rigidez de nuca**</e>, no ingurgitación yugular. .... `pre`

E2 CyC: Rigidez de nuca, no <e>**ingurgitación yugular**</e>. .... `abs`

E3 Los hallazgos descritos son sugestivos de <e>**pielonefritis aguda**</e>. .... `pos`

E4 <e>**Tumoraciones faciales**</e> en paciente transplantada hepatica .... `pre`

E5 Tumoraciones faciales en paciente <e>**transplantada hepatica**</e> .... `pre`

At the moment of executing the experiments described in this chapter, there is no publicly available dataset in Spanish annotated with medical entities and their assertion category. Thus, in order to conduct this experiment, we automatically construct a new corpus from NUBes, with the help of the original cue, scope and entity annotations. The transformation process is as follows:

First, we automatically annotate the entire corpus with medical entities[1]. To that end we use UMLSmapper, a tool for annotating medical entities in Spanish texts and linking them to the UMLS Metathesaurus (Bodenreider, 2004), the topic of Chapter 7. Specifically, we annotate mentions of the following types of entities: clinical findings and disorders, procedures, chemicals and drugs, physiological phenomena, and some living beings (viruses, bacteria, and fungi)[2].

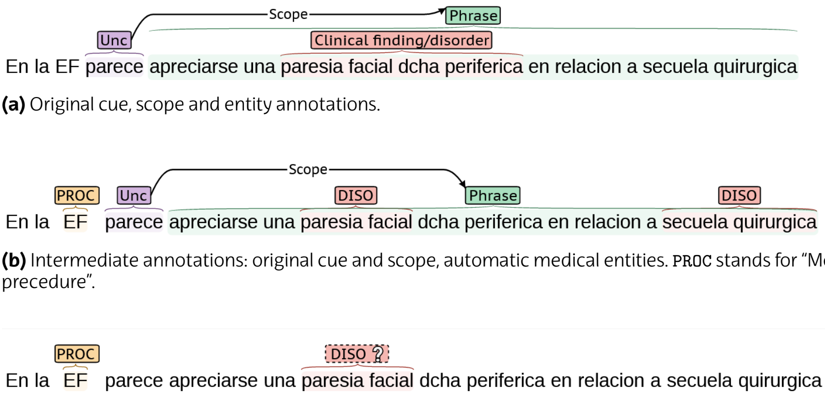
Then, we assign the categories `abs` (absent), `pos` (possible) or `pre` (present) to each annotated entity depending on whether they occur within the scope of a negation cue, an uncertainty cue or neither, respectively.

To be specific, however, not all the entities that fall within the scope of a negation or uncertainty cue are directly affected by it. Consider the sentence in

[1] NUBes has annotations of medical entities, but only of those directly affected by the cues within each negation or uncertainty scope (see Section 10.3.1.4.3 of Chapter 10). In this chapter, we are interested in being able to classify any entity, including the ones that are said to be present (`pre`). To that end, we could have kept the manual annotations of entities and complement those with the suggestions of an automatic medical entity recogniser; instead, we chose to discard the original annotations altogether and automatically annotate the entire corpus, simply to avoid inadvertently injecting artificial traces that the assertion classifiers might pick up to differentiate between entities directly affected by cues (manually annotated entities) and the rest (automatically annotated entities).

[2] The classification of types is given by the UMLS semantic groups (Bodenreider et al., 2003).

Figure 12.1. While 'secuela quirurgica' is a clinical finding under the scope of an uncertainty cue, the speculation is rather about the facial paresis than the surgical sequelae or the relation of the former to the latter (see also the discussion in Section 10.3.1.4.3 of Chapter 10). The entities annotated in NUBES are only those most prominently affected by the corresponding cue. Based on this information, we remove the entities that fall within the scope of a cue but that do not overlap with a manually annotated entity in the cases where there is one. This way, we avoid incorrectly annotating as negated or uncertain entities such as 'secuela quirurgica' in Figure 12.1.



(a) Original cue, scope and entity annotations.

(b) Intermediate annotations: original cue and scope, automatic medical entities. PROC stands for "Medical precedure".

(c) Final processed annotations of medical entities and their status.

Figure 12.1: Example of the processing of a NUBES instance to create the assertion classification corpus. Translation: "In the P[hysical] E[xamination], a peripheral right facial paresis is seemingly noticed in relation to surgical sequelae.".

Even then, we have manually revised the testing portion of the dataset, which allows us, on the one hand, to measure the validity of the proposed data conversion and, on the other hand, to ensure the reliability of the reported results and conclusions drawn therefrom. The manual revision led to correcting the assertion category of 38 instances and removing 7 instances out of the 2,474 revised examples.

Finally, each annotated entity must be converted to the text classification format presented earlier (see Examples E1 to E5). The annotations in Figure 12.1c would yield the following instances:

E6 En la <e>**EF**</e> parece apreciarse una paresia facial dcha periferica [...] .... `pre`

E7 En la EF parece apreciarse una <e>**paresia facial**</e> dcha periferica [...] .... `pos`

Notice that we do not care about the correctness of the UMLS links established by UMLSmapper nor of the entity types assigned thereof, which we simply use to filter the annotations. The task the classification models need to learn is to establish a relation between the entity and the context it occurs in, in order to emit a prediction regarding whether the entity is present, absent, or possible. The type of the entity (disorder, drug, and so on) is irrelevant to the task, even more so its link to the UMLS Metathesaurus.

In this chapter, we also work with the original training, development and test splits of the NUBEs corpus, as in Lima-López et al. (2020a) [6]. The resulting dataset is described quantitatively in Table 12.1. We followed the methodology to generate incremental training subsets (1/1 through $1/3^5$) and the more difficult testing set, ADV, as explained in the previous chapter (Section 11.2.1).

In addition, this chapter exploits a third test dataset, consisting of the original entity annotations of the NUBEs corpus, that is, the manual (MAN) annotations of entities. This test set is simply added for the sake of completeness, although, as explained above, it does not include `pre` (present) annotations (which is why the corpus was automatically re-annotated).

Table 12.1: Size of the corpus for the assertion classification task

	Train						Dev	Test		
	1/1	**1/3**	**$1/3^2$**	**$1/3^3$**	**$1/3^4$**	**$1/3^5$**		**FULL**	**ADV**	**MAN**
Total entities	12,108	4,035	1,344	447	148	49	1,659	2,467	1,507	1,300
Absent (`abs`)	2,399	782	277	92	34	11	331	460	95	973
Possible (`pos`)	1,001	332	118	39	14	5	140	197	125	327
Present (`pre`)	8,708	2,921	949	316	100	33	1,188	1,810	1,287	-
`pre` OOS	3,912	1,317	436	140	43	9	534	818	295	-

Of note, Table 12.1 specifies *out-of-scope* (OOS) present entities, that is, examples of entities mentioned in the context of a negation or uncertainty cue, but that are not affected by it (e.g., ‘EF’ in Figure 12.1c). Without OOS examples, the models would simply learn to detect the presence or absence of negation and uncertainty cues, regardless of whether they affect or not the target entity.

12.2.2 Systems

Assertion classification is a text classification task, where each medical entity whose assertion status needs to be predicted is presented to the systems delimited by special tokens in the sentence they occur in (see Examples E6 and E7). The systems tested in this chapter are the following:

12.2.2.1 Baseline

As is customary in this type of task, the NegEx (Chapman et al., 2001) system serves as a baseline in our experiments. NegEx is a rule-based system that leverages hand-crafted lexicons in order to determine the assertion categories of the given medical entities. The lexicons define 4 types of words or expressions: conjunctions, pseudo-negation cues, negation cues and uncertainty cues. The first two are used to find the boundaries of scopes and to discard false cues, respectively. Negation and uncertainty cues are further divided into two groups each, depending on whether they precede (`PRE`) or follow (`POST`) their scopes. Although NegEx has been adapted to Spanish on several occasions (see Section 9.3 in Chapter 2), only one adaptation is publicly available [69]. Unfortunately, it does not consider uncertainty. Thus, in this experiment we use the original NegEx Python implementation [70] with cues automatically extracted from our training data sets. The categories of the cues (`PRE` or `POST`) are automatically determined by choosing the most frequent position in the corpus. The conjunction and pseudo-negation lexicons have been taken from [69] as is.

12.2.2.2 Flair

The Flair NLP framework (Akbik et al., 2019) comes with a text classifier implementation as well as the sequence labeller trained in the previous chapter. The word representations are obtained following the same mechanism as described for the sequence tagger (Flair's `es-forward` and `es-backward` embeddings, and the Spanish biomedical fastText word embeddings by Soares et al. (2019b); see Section 11.2.2.1). In this case, the computed embeddings are fed into a Gated Recurrent Unit (GRU) layer to produce a document level representation, which is then used in a linear layer to make the assertion category prediction.

12.2.2.3 Transformer

As with the sequence labelling task (Chapter 11), we evaluate an assortment of text classification systems based on the Transformer architecture. The pre-trained models tested are the same as for the sequence labelling task. See Table 11.3 in the previous chapter and Appendix E for detailed information on each model; we list them here briefly for convenience:

- BETO$_{Base}$ Cased (Cañete et al., 2020), hereafter just BETO
- Multilingual BERT$_{Base}$ Cased [23], mBERT
- IXAmBERT$_{Base}$ Cased (Otegi et al., 2020), IXAmBERT
- SciBERT$_{scivocab}$ Cased (Beltagy et al., 2019), SciBERT
- SpanBERTa$_{Base}$ Cased [26], SpanBERTa
- MarIA RoBERTa$_{Base}$ BNE (Gutiérrez-Fandiño et al., 2022), MarIA

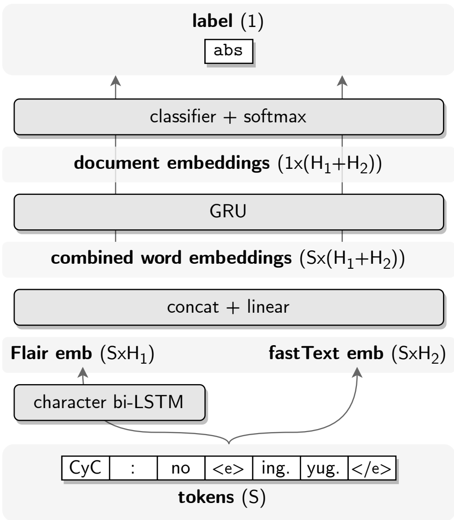


Figure 12.2: Diagram of the Flair-based assertion classifier. S (sequence length); H_1 = 128 or 256 (Flair embedding size); H_2 = 300 (fastText embedding size).

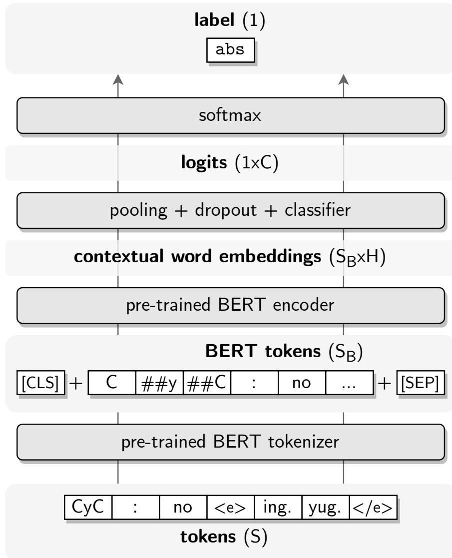


Figure 12.3: Diagram of the BERT-based assertion classifier. S_O (original sequence length); S_B = 220 (sequence length after BERT tokenisation); H = 768 (BERT embedding size); C = 3 (number of output labels).

- XLM-RoBERTa$_{Base}$ (Conneau et al., 2020), XLM-R

The classifier head is fed in this case the pooled output of the encoder. The pooled output is computed over the special token at the beginning of each sequence (i.e., BERT's [CLS] and RoBERTa's <s>) by passing its embeddings to a dense linear layer and a tanh activation function. The result is then fed to a dropout layer and the final dense linear layer which outputs the logits for the 3 categories of the task. For this task, we added the special tokens <e> and </e>, which mark the start and end of the medical entity, to the vocabularies of the pre-trained models. Again, the models are trained on the cross-entropy loss of the classification head and, for inference, the label with the maximum probability is chosen after the softmax function.

12.2.2.4 Implementation and training setup

The implementation and training setup is the same as that of the experiments on the sequence labelling task. See Section 11.2.2.4 in the previous chapter and Appendix F. As for the baseline system NegEx, we compute the train curve by extracting the negation and uncertainty cues only from the corresponding training data subset at each point.

12.2.3 Evaluation

The main evaluation metric for these experiments is again F_1 (see Equation 8.1 (=4.1)), as computed by the Python package `sklearn` (Pedregosa et al., 2011). True positive (TP), false positive (FP) and false negative (FN) are counted per category $c \in \{$`abs`, `pos`$\}$ as follows:

- TP: number of entities of type c correctly classified as c.
- FP: number of entities of a type other than c incorrectly classified as c.
- FN: number of entities of type c incorrectly classified as other than c.

The category `pre` is the negative class, in the sense that it is the unmarked, majority category—nothing to do with negative polarity—and we do not take it into account when computing our metrics to prevent misleadingly inflated results.

As average metrics of the different categories, we report micro-average (μ) scores. The micro-average scores are obtained by applying the same equations to the sums of the TP, FP and FN of the different categories.

12.3 Results

12.3.1 Assertion classification

Table 12.2 shows the main results of the chapter. It reports per-category and micro-average F_1-scores of models trained in the full train set and one of the train subsets (with ∼1% of examples). The models trained in the full train set are evaluated in two test sets: Full (of entities annotated by UMLSmapper) and Man (of entities annotated manually). Precision and recall metrics can be consulted in Appendix G.

Similarly to the cue and scope detection task, MarIA obtains the best overall results (0.937 F_1-score) in the Full test set, followed by the multilingual models mBERT (0.935) and XLM-R (0.934), and BETO (0.934). Nevertheless, the differences between the Transformer models are narrower still than in the previous chapter, and even SciBERT manages to perform on par with SpanBERTa and IXAmBERT. The system based on Flair falls in average 3 F_1-score (F_1) points behind the worst Transformer.

All these systems outperform by far the baseline set by the rule-based system NegEx when allowed to exploit the whole training set, but mostly lag behind in the ~1% training set scenario. Only SpanBERTa is capable of topping NegEx in this case, with a micro-average (μ) F_1-score of 0.660. Also noteworthy is that XLM-R achieves 0.812 F_1-score in the classification of absent entities with just 148 training examples. These questions will be discussed further in the next section.

Regarding the classification of the original entity annotations of NUBES (i.e., the MAN test set), the overall results are even higher compared to the synthetic FULL test set, with the best F_1-score, 0.978, achieved in this case by XLM-R. The generalised improvement is explained by the fact that this test set only contains `abs` and `pos` entities—no "out-of-scope" `pre` that could lead to false positive predictions; the metric that improves more markedly is indeed precision, while recall scores hardly improve or even worsen slightly (see Appendix G).

Table 12.2: F_1-score results for assertion classification. The best and second-best scores are highlighted in bold and dotted underlines, respectively. N is the number of training examples.

	FULL test						MAN test		
	$1/3^4$ train (N=148)			Full train (N=12,108)			Full train (N=12,108)		
	μ	abs	pos	μ	abs	pos	μ	abs	pos
NegEx	0.647	0.698	**0.469**	0.683	0.700	0.638	0.890	0.922	0.783
Flair+fT	0.003	0.004	0.000	0.889	0.892	0.882	0.939	0.951	0.903
BETO	0.612	0.729	0.409	0.934	**0.943**	0.914	0.972	0.979	0.952
SpanBERTa	**0.660**	0.759	0.330	0.927	0.937	0.905	0.967	0.971	0.955
MarIA	0.588	0.716	0.258	**0.937**	0.940	0.929	0.971	0.979	0.950
IXAmBERT	0.586	0.697	0.248	0.925	0.934	0.902	0.957	0.967	0.929
mBERT	0.635	0.731	0.438	0.935	0.939	0.925	0.973	0.978	**0.960**
XLM-R	0.647	**0.812**	0.292	0.934	0.934	**0.934**	**0.978**	**0.984**	0.959
SciBERT	0.458	0.586	0.149	0.927	0.931	0.916	0.967	0.975	0.943

In general, the task of assertion classification seems to be easier than cue and scope detection. The drop in performance from the negative class (`abs`) to the uncertainty class (`pos`) is also smaller. Still, the synthetic nature of the corpus is likely playing a role in this regard, particularly because it hardly contains the type of instances that could potentially induce errors the most, namely, instances with entities within negation or speculation scopes but that are not the entity most prominently affected by it (see "secuela quirurgica" in Figure 12.1b and the related discussion in Section 12.2.1).

12.3.2 Train curves and adversarial examples

The train curves in Figure 12.4 have been generated by training each model with the increasing training samples and evaluating the resulting models in the FULL and ADV test sets. The curves represent F_1-scores for absent and possible entities.

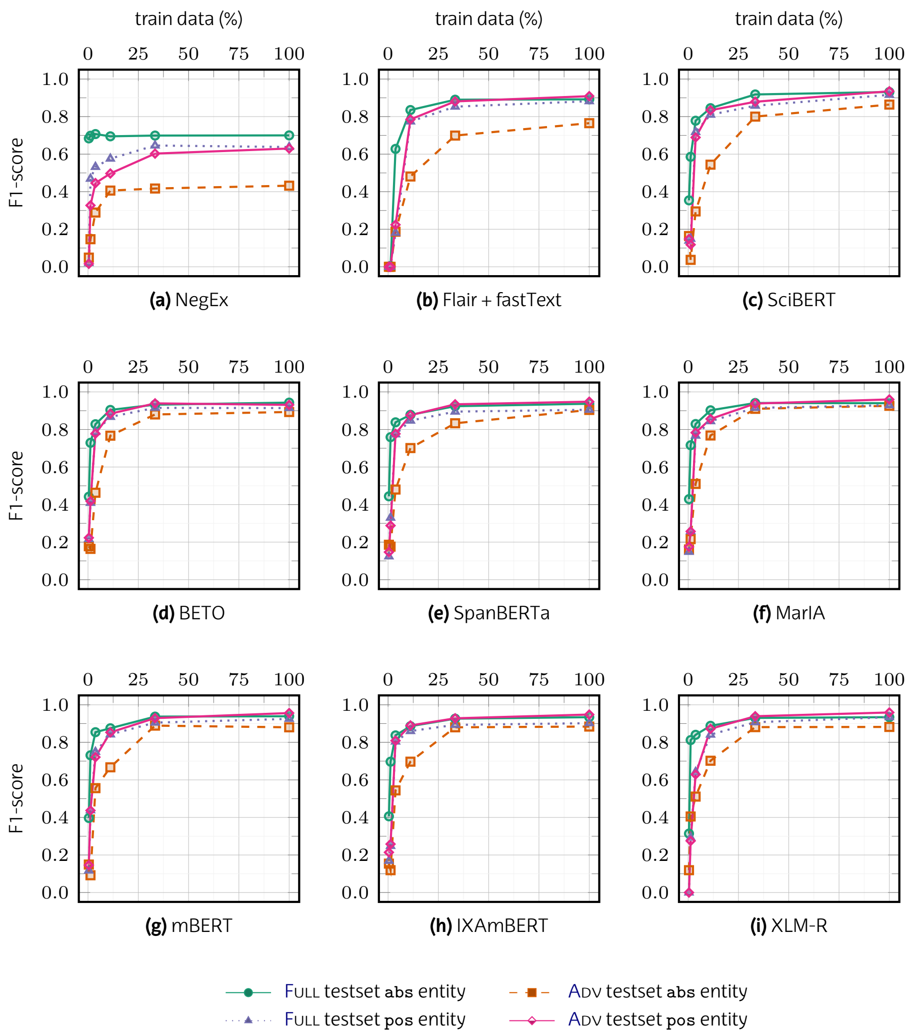


Figure 12.4: Train curves on the assertion classification task.

We observe quite a different landscape to that in the previous chapter for the task of cue and scope detection. The gap between the full and harder test sets is much narrower (except for NegEx), and the systems seem to reach a plateau earlier with around a third of the training set. Furthermore, monolingual and multilingual models do not have such markedly different behaviours in this case. Most of all, Figure 12.4 clearly demonstrates the problem of rule-based systems such as NegEx. Even if it has an excellent start at classifying the easiest negated instances, the system is just not capable of generalising to unseen cases even as the available data to enrich the tool's lexicons increases.

12.3.3 Error analysis

As shown in Figure 12.3, false positive errors are more frequent in this task than in the detection of cues and scopes and, in fact, constitute the bulk of errors made by the systems overall. A manual analysis of these errors revealed that they involve entities near cues but that are not in focus, as in the following examples (starred categories indicate that the predictions are incorrect):

E8 No mejoró con la toma de <e>**Paracetamol**</e>. .... *abs
[The patient] did not improve with Paracetamol.

E9 Cuadro confusional de probable reacción al <e>**proceso infeccioso**</e>. .... *pos
Confusional state of probable reactive character to the infectious process.

E10 Se aconseja TAC para valorar la causa de la <e>**obstrucción [...]**</e> .... *pos
CT is advised to assess the cause of the bile duct obstruction

In Example E8, the negation is about the improvement of the patient, who *did* take Paracetamol. In Example E9, it is the relation between the confusional state and the infectious process that is uncertain, not whether an infectious process took place—the use of determinate article 'the' in 'the infectious process' makes it clear that it is in fact a reference to a known past event. Finally, in Example E10, it is the origin of the obstruction that is unknown, not the existence of the obstruction itself (the same rationale applies here). These examples are particularly tricky because they require deeper understanding of the sentences than that needed to simply find cues and scopes. Even then, it is likely that fewer of this type of errors might occur if the models were trained on gold standard corpora instead of the automatically generated corpus described here.

As for false negative errors, we found two types of instances that confuse the models:

1. Sentences that express a change of state, such as disappearance of symptoms or modifications in a treatment:

Table 12.3: Confusion matrices of the assertion classification task; predictions made by the models trained on the entire training set for the FULL test set. N is the number of true examples for each category in absolute terms.

		predicted			predicted			predicted		
	N	abs	pos	pre	abs	pos	pre	abs	pos	pre
true abs	460	0.40	0.01	0.04	0.58	0.00	0.05	0.62	0.00	0.02
true pos	197	0.02	0.13	0.04	0.01	0.23	0.03	0.01	0.25	0.01
true pre	1,810	0.27	0.08		0.08	0.03		0.06	0.02	

(a) NegEx — (b) Flair + fastText — (c) SciBERT

	N	abs	pos	pre	abs	pos	pre	abs	pos	pre
true abs	460	0.63	0.00	0.01	0.62	0.00	0.02	0.63	0.00	0.02
true pos	197	0.01	0.25	0.02	0.01	0.25	0.02	0.01	0.26	0.02
true pre	1,810	0.05	0.02		0.05	0.03		0.05	0.01	

(d) BETO — (e) SpanBERTa — (f) MarIA

	N	abs	pos	pre	abs	pos	pre	abs	pos	pre
true abs	460	0.63	0.01	0.01	0.62	0.00	0.02	0.63	0.00	0.01
true pos	197	0.00	0.26	0.01	0.01	0.25	0.02	0.01	0.26	0.01
true pre	1,810	0.06	0.02		0.05	0.03		0.07	0.01	

(g) mBERT — (h) IXAmBERT — (i) XLM-R

E11 Presenta <e>**fiebre**</e> elevada que cede con tratamiento antibiótico. .... *pre
[The patient] has high fever that goes down with antibiotic treatment.

E12 Le pautaron <e>**Diclofenaco**</e> que no está tomando .... *pre
[The patient] was prescribed Diclofenaco which she does not take

In these cases, the symptom or treatment is asserted in the main clause of the sentence but negated in the relative clause. Although debatable, the guidelines of the NUBES corpus indicate that these examples should be explicitly annotated as negations, but the models seem to struggle with such instances.

2. Long sentences where the scope precedes a negation cue, which occurs towards the end of the sentence:

E13 Se obtiene <e>**cultivo de sangre**</e> y [...] siendo negativos. .... *pre
Blood culture and [...] were obtained with negative result.

The long distance between the cue and the scope, as well as their less common order in the sentence, appears to make it more difficult for the systems to establish a relation between the two.

In the case of assertion classification, there does not seem to be much confusion between instances of negated and possible entities as there was in the cue and scope detection task.

Finally, as part of the error analysis, we studied whether the errors that the systems are making in the two tasks (that is, the tasks of the previous and current chapters) coincide somehow in the same instances, given that the corpora for the two tasks originate from the same collection of sentences. Out of the 2,762 sentences for testing the cue and scope detection models, 272 have errors (made by any of the evaluated models). In the present task, assertion classification, the ratio is 196 out of 2,467. A significant amount, 92 sentences, are common to both evaluations and involve most of the situations discussed here and in the error analysis of the previous chapter (Section 11.3.3), with a prominent presence of sentences with relative clauses where the scopes of cues are split into discontinuous spans, one of which precedes the cue and the other follows it.

12.4 Conclusions

Regarding the assertion classification task, we first proposed a series of steps to convert the NUBes corpus, originally annotated for cues and scopes, to a corpus suitable for this task. A manual revision of the testing portion of the resulting corpus, as well as a manual error analyses of the results, suggest that this technique yields acceptable results and can be useful in scenarios where there is no such corpus available, as was the case in this work. In this task too, MarIA obtained the best results (0.937 micro-average F1-score), followed even more closely by the other Transformers, including SciBERT.

We observed that, in both tasks, neither the models with most vocabulary overlap with NUBes nor the biggest models obtained the best results, although they did follow closely MarIA. Further, the training curves showed that, while monolingual Spanish models start off with certain advantage, being able to correctly emit predictions for the most frequent and repetitive instances, all the Transformer models manage to obtain similar results when allowed to exploit the entire training sets. The training curves also showed that less annotated data might be necessary for the assertion classification task than for the cue and scope detection task.

A manual error analysis revealed that in the case of the assertion classification task, the most common errors involve false positive errors, where medical entities under the scope of cues but *not* in focus are incorrectly tagged as absent or possible instead of present. The manual error analysis also uncovered several incorrect annotations, which will help us improve the quality of the corpus.

PART V

CONCLUSIONS

Chapter 13

Conclusions

13.1 Summary

In this thesis, we study three key topics within the field of clinical IE, focusing specifically on content written in Spanish. We make several contributions to this field in the form of a system for term identification, a dataset annotated for negation and uncertainty, and several experiments on these topics, as well as the problem of sensitive data detection and categorisation. Throughout the thesis, we apply and compare techniques of varying levels of sophistication and novelty, which reflects the rapid advancement of the field during the years that this work has been carried out. Next, we provide a quick summary of the objectives, research and conclusions for each of the main topics of the dissertation.

13.1.1 Sensitive data detection and categorisation

Objectives

- To study the question of sensitive data in health record texts in Spanish from a technical point of view, in order to better understand how to characterise and approach it as a target of detection and classification systems based on NLP.
- To assess and compare supervised approaches in the task of sensitive data detection and categorisation in clinical text, and to identify the advantages and limits of the different methods.

In Part II of the thesis, we have tested four sequence labelling techniques, namely, CRFs (Lafferty et al., 2001), biLSTMs (J. Yang et al., 2018b), spaCy's NER tagger [37], and BERT (Devlin et al., 2019). The first belongs to traditional ML, while the rest consist of DNNs. Further, the CRF and biLSTM models have been learnt over a rich set of lexical, morphosyntactic and semantic features, while the

BERT-based model has been obtained by fine-tuning a pre-trained multilingual LM. Some of these models are available online [4].

Our first experiment has been conducted in the context of the MEDDOCAN challenge (Marimon et al., 2019), where the challenge data consisted of clinical cases manually enriched with personal data. Here, BERT has obtained the best metrics, with a greater advantage in terms of recall, followed by the biLSTM model. Still, we have not observed striking differences among the systems, all of them having obtained excellent results with F_1-scores above 0.95. We discussed that, while MEDDOCAN's synthetic data may well be a fair reflection of some types of health records, there do exist more challenging data in real scenarios.

In fact, BERT has proven to be matchless when being tested under harsher conditions. When applying the MEDDOCAN models on a corpus of real health records, BERT has demonstrated far superior generalisation capabilities, with a recall of 0.53 in the detection scenario—the second-best recall in the same scenario being 0.18. In addition, we have measured the robustness of these models to decreasing training samples. Again, the BERT-based model has proven to be more advantageous, losing only 15 points of F_1-score when trained on 230 instances instead of the entire dataset (i.e., 21,371 instances).

In line with the literature that uses BERT for other tasks, these results indicate that the knowledge transfer achieved through the pre-trained LM model not only helps obtain better results, but also diminishes the need of manually labelled data. Furthermore, this approach eliminates the dependency on feature extraction and engineering. These are decisive advantages, given the difficulties in collecting large corpora and the lack of basic linguistic analysis tools adapted to the Spanish language and the clinical domain.

13.1.2 Term identification

Objectives

- To build a system capable of performing clinical term identification natively in the Spanish language, that does not require annotated data of any kind, and that may be easily configured to meet the requirements of diverse application scenarios.
- To compare said system to other approaches proposed in the literature, most of which rely on MT at some point in the processing pipeline in order to leverage existing solutions for the English language, and to identify the advantages and limits of the tested methods.

In part III of the thesis, we have described and evaluated UMLSmapper, a prototype for biomedical term identification built on the UMLS Metathesaurus (Bodenreider, 2004). This system recognises and identifies terms in the same step based mainly on lexical similarity metrics. It is built on Apache Lucene™ for fast match retrieval, and it uses UKB (Agirre et al., 2009) to resolve ambiguities. While UMLSmapper does depend on the availability of a sufficient coverage of the UMLS Metathesaurus for the desired language, it can be easily tailored to map different categories of concepts, without depending on external NER tools adapted to each specific problem to be solved. UMLSmapper is available online for research purposes through a web API [5].

We have compared it to MetaMap (Aronson, 2001, 2006) and Transfer (Accuosto et al., 2018) on the Mantra GSC English and Spanish datasets (Kors et al., 2015). MetaMap is a well-known, robust engine for English biomedical term identification with the UMLS. Transfer is a pipeline that applies existing term identification tools like MetaMap on machine translated text, and projects the labels back to the original text through semantic similarity techniques.

Our tool has obtained an average term identification F_1-score of 0.674 and 0.626 in English and Spanish respectively. It has managed to better MetaMap by a narrow margin on the English data. As for Transfer, UMLSmapper has surpassed it in the Spanish data thanks to a greater recall. Moreover, ensembles of UMLSmapper and Transfer have improved the results of the individual pipelines, the most competitive combination being that which favours Transfer's predictions in case of overlapping predictions due to Transfer's superior precision.

13.1.3 Negation and uncertainty detection

Objectives

- To study the phenomena of negation and uncertainty in health records in Spanish, in order to propose guidelines for their annotation and to better understand how to characterise and approach them as a target of detection and classification systems based on NLP techniques.
- To build a corpus of clinical texts in Spanish manually annotated with negation and uncertainty information following the above-mentioned guidelines.
- To assess and compare supervised approaches in the task of negation and uncertainty detection in clinical text, and to identify the advantages and limits of the different methods.

In Part IV of the thesis, we have first presented a new corpus, NUBes, of clinical texts in Spanish annotated for negation and uncertainty. The corpus is publicly available for research purposes [6]. Then, we have conducted several experiments with the corpus on the automatic detection of these linguistic phenomena.

NUBES consists of 29,682 sentences extracted from health records of 18 medical specialities and 7 different EHR sections. A total of 8,855 sentences contains at least one annotation related to negation and/or uncertainty. The NUBES annotation guidelines consider syntactic, lexical and morphologic cues of negation and uncertainty, as well as their scopes. In addition, it takes into account medical entities and polarity items within said scopes.

We have exploited this corpus to tackle the problem of negation and uncertainty detection from two perspectives: first, as a sequence labelling problem, where the goal has been to detect cues and scopes; second, as a classification problem, where the task has consisted in deciding whether a given medical entity is "present", "possible", or "absent". In both cases, we have compared multiple models based on the Transformer architecture (Vaswani et al., 2017) and Flair (Akbik et al., 2019).

The model based on MarIA (Gutiérrez-Fandiño et al., 2022) has consistently achieved the best overall results. More interestingly, the training curves have shown that, while monolingual Spanish models start off with certain advantage, being able to correctly emit predictions for the most frequent and repetitive instances, all the Transformer models manage to obtain similar results when allowed to exploit the entire training sets. The training curves also showed that less annotated data might be necessary for the assertion classification task than for the cue and scope detection task.

A manual error analysis has revealed that, in the case of the sequence labelling approach, the most common errors are false negative errors of scopes, and involve more frequently sentences with coordination or relative clauses. In the case of the assertion classification task, we have observed that the most common errors involve false positive errors, where medical entities under the scope of cues but *not* in focus are incorrectly tagged as absent or possible instead of present.

13.2 Publications

In what follows, we present a list of the author's publications relevant to the research described in this document, with explanations of how they relate to specific chapters. The final section contains publications that have not been covered here but that are closely related to the research topics of the thesis.

Part II: SENSITIVE DATA DETECTION AND CLASSIFICATION

1. **Naiara Perez**, Laura García-Sardiña, Manex Serras and Arantza del Pozo (2019). "Vicomtech at MEDDOCAN: Medical Document Anonymization". In: ***Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th***

Conference of the Spanish Society for Natural Language Processing (SEPLN 2019) (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 696–703

*Indexed in **Scopus***

This paper contains Vicomtech's working notes for the MEDDOCAN challenge. It is the keystone of Chapter 4, which could be seen as an extended version of these working notes.

2. Aitor García-Pablos, **Naiara Perez** and Montse Cuadros (2020a). "Sensitive data detection and classification in Spanish clinical text: experiments with BERT". In: ***Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*** (Conference cancelled). European Language Resources Association, pp. 4486–4494

   GGS**: Class 3 Rating **B

   *Indexed in **Scopus***

   This paper describes the experiments carried out with the NUBes-PHI corpus, as well as the post-challenge evaluation of BERT with the MEDDOCAN corpus. These contributions have served to finish off Chapter 4 and build the foundation for Chapter 5. However, the experimental design of Chapter 5 is not that reported in the paper, having used different evaluation metrics and performed additional experiments, in order to maintain internal coherence with Chapter 4.

Part III: TERM IDENTIFICATION

3. **Naiara Perez** (2017). "Mapping of Electronic Health Records in Spanish to the Unified Medical Language System Metathesaurus". MA thesis. Univertiy of the Basque Country (UPV/EHU), pp. 1–87

   In the Master's thesis we described the initial version of UMLSmapper and compared it indirectly to MetaMap (Aronson, 2001, 2006). An updated description of this version of UMLSmapper is given in Chapter 7. The experimentation reported in this publication has not been included in this work, as we have since performed more informative tests on a gold standard corpus (Chapter 8).

4. **Naiara Perez**, Montse Cuadros and German Rigau (2018). "Biomedical term normalization of EHRs with UMLS". In: ***Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*** (Miyazaki, Japan, 7th–12th May 2018). European Language Resources Association, pp. 2045–2051

   GGS**: Class 3 Rating **B

   *Indexed in **Scopus***

   This paper is a summarised version of the Master's thesis.

5. **Naiara Perez**, Pablo Accuosto, Àlex Bravo, Montse Cuadros, Eva Martínez-Garcia, Horacio Saggion and German Rigau (2020). "Cross-lingual semantic annotation of biomedical literature: Experiments in Spanish and English". In: ***Bioinformatics*** 36.6, pp. 1872–1880

   *JCR™ 2020: Impact Factor 6.937, **Q1** (3/58 in Mathematical & Computational Biology)*
   ***SJR 2020:** Impact Factor 3.599, **Q1** (8/2,196 in Computer Science Applications)*
   *Indexed in **Web of Science** and **Scopus***

   This paper is the result of a collaboration with the Natural Language Processing Group (TALN) of the University Pompeu Fabra (UPF). Here, we compare several pipelines for biomedical term identification in Spanish, including UMLSmapper. Most of work and results described in Chapter 7 and Chapter 8—except the experiments over English text—are summarised in this publication.

Part IV: NEGATION AND UNCERTAINTY DETECTION

6. Salvador Lima-López, **Naiara Perez**, Montse Cuadros and German Rigau (2020a). "NUBes: A corpus of negation and uncertainty in Spanish clinical texts". In: ***Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*** (Conference cancelled). European Language Resources Association, pp. 5772–5781

   GGS**: Class 3 Rating **B
   *Indexed in **Scopus***

   This paper describes the process of creating the NUBes corpus and its outcome. These contributions have been reported in Chapter 10. The paper also includes a preliminary set of experiments with the corpus, which serve as baseline of the experiments in Chapter 11.

7. **Naiara Perez**, Montse Cuadros and German Rigau (n.d.). "Negation and speculation processing: a study on cue-scope labelling and assertion classification in Spanish clinical text". Under review as a journal article.

   This paper recounts the experimentation of Chapters 11 and 12 about approaching the detection of negation and speculation as sequence labelling and sequence classification problems, respectively.

Other related publications

8. Montse Cuadros, **Naiara Perez**, Iker Montoya and Aitor García-Pablos (2018). "Vicomtech at BARR2: Detecting biomedical abbrebiations with ML methods and

dictionary-based heuristics". In: ***Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*** (Sevilla, Spain, 18th Sept. 2018). CEUR Workshop Proceedings, pp. 322–328

*Indexed in **Scopus***

9. Salvador Lima-López, **Naiara Perez**, Laura García-Sardiña and Montse Cuadros (2020b). "HitzalMed: Anonymisation of clinical text in Spanish". In: ***Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*** (Conference cancelled). European Language Resources Association, pp. 7038–7043

   CORE 2020:** Rank **C
   *Indexed in **Scopus***

10. Aitor García-Pablos, **Naiara Perez** and Montse Cuadros (2020b). "Vicomtech at CANTEMIST 2020". In: ***Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020)*** (Online, 23rd Sept. 2020). CEUR Workshop Proceedings, pp. 489–498

    *Indexed in **Scopus***

11. Aitor García-Pablos, **Naiara Perez** and Montse Cuadros (2020c). "Vicomtech at eHealth-KD challenge 2020: deep end-to-end model for entity and relation extraction in medical text". In: ***Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020)*** (Online, 23rd Sept. 2020). CEUR Workshop Proceedings, pp. 102–111

    *Indexed in **Scopus***

12. Salvador Lima-López, **Naiara Perez** and Montse Cuadros (2021). "Grammatical error correction for Spanish health records". In: ***Procesamiento del Lenguaje Natural*** 66, pp. 121–132

    ***SJR 2020:** Impact Factor 0.149, **Q4** (1,334/2,196 in Computer Science Applications)*
    ***Dialnet Metrics 2019:** Impact Factor 0.377, **Q1** (6/70 in Linguistics)*
    ***FECYT Certificate of Excellence 2020:** **Q1** (2/52 in Linguistics)*
    *Indexed in **Web of Science** and **Scopus***

13. Aitor García-Pablos, **Naiara Perez** and Montse Cuadros (2021a). "Vicomtech at eHealth-KD challenge 2021: deep learning approaches to model health-related text in Spanish". In: ***Proceedings of the Iberian Languages Evaluation Forum***

(IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021) (Online, 21st Sept. 2021). CEUR Workshop Proceedings, pp. 712–724

Indexed in ***Scopus***

14. Aitor García-Pablos, **Naiara Perez** and Montse Cuadros (2021b). "Vicomtech at MESINESP2: BERT-based multi-label classification models for biomedical text indexing". In: ***Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*** (Online, 21st–24th Sept. 2021). CEUR Workshop Proceedings, pp. 102–111

    Indexed in ***Scopus***

15. **Naiara Perez**, Aitor Álvarez, Arantza del Pozo, Andrés Arbona, Oihane Ibarrola, Marta Suarez, Pedro de la Peña Tejada and Itziar Cuenca (2022). "ESAN: Automating medical scribing in Spanish". In: ***Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations (SEPLN-PD 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022)*** (A Coruña, Spain, 21st–23rd Sept. 2022). CEUR Workshop Proceedings, pp. 10–13

    Indexed in ***Scopus***

13.3 Future Work

As future work, we envision two distinct avenues of research.

On the one hand, there is the line of research related to the study of practicalities and viability issues that emerge when attempting to bring this technology to end users in the environment of interest (i.e., hospitals, healthcare centres). In this sense, the intrinsic evaluations presented in this thesis should be complemented with extrinsic tests that measure to what extent these tools—either alone or in combination—can help accelerate, improve, or even enable new processes in real healthcare practice and research.

Along these lines, there is the specific question of how to better bring the presented tools together into one system of clinical IE, particularly with regards to the tasks of term identification and of negation and uncertainty detection. This problem raises new pragmatic questions, such as how to handle antonymy and complex negated terms for which specific codes exist in the knowledge base of interest (e.g., "IUD not visible" corresponds to the concept identifier C1698536 [71] in the UMLS, but not all negated concepts have a code).

Furthermore, future research should investigate efficient methods for integrating the machine learning life-cycle into production healthcare settings. Due to the dynamism of the sector, techniques such as online learning are crucial for continually improving systems and keeping them updated.

On the other hand, the rapid advances of the NLP field in the last few years provide new opportunities to improve and extend the presented work.

In this respect, the results obtained in sensitive data detection and categorisation with Multilingual BERT [23] can in all probability be improved upon by fine-tuning more appropriate LMs that have been made available since then (e.g., the clinical RoBERTa LM for the Spanish language by Carrino et al. [2021]). Future work should also address techniques for document anonymisation once the sensitive data has been detected and categorised. Of particular interest in NLP is the automatic suggestion of surrogate data. Current approaches are based on language and domain specific gazetteers (Lima-López et al., 2020b; Emelyanov, 2021), a dependency that may be eliminated with pre-trained LMs.

With respect to term identification, the naive, lexically motivated approach presented here responds to the self-imposed restrictions of not requiring annotated data nor being dependent on external NER tools. The most recent related work (i.a., Wajsbürt et al., 2021; Yuan et al., 2022) employ more advanced techniques based on continuous word and/or graph embeddings and bi-encoders (Reimers et al., 2019); but these works have oracle term annotations as starting point, so their application in our use case is not straightforward. As future work, we should explore ways to incorporate them into our system in order to overcome its many limitations.

As for negation and uncertainty, multi-task learning offers a new avenue of research. In this setup, the tasks of cue and scope detection and assertion classification would be learned jointly by the same model in separate classification heads, possibly benefiting one another. Interestingly, Hartmann et al., 2021 find that learning to classify events into the affirmed or negated categories as an auxiliary task to negation scope resolution does not help and can even be detrimental. However, their setup exploits a different corpora per task and those corpora involve different languages. Furthermore, they do not look into how the task of negation scope resolution affects assertion classification.

Following the paradigm shift in the NLP community (P. Liu et al., 2021; Sun et al., 2022), future work may address all these problems with yet other emergent approaches, such as sequence-to-sequence and/or prompt-based learning, leveraging perhaps bigger language models (e.g., GPT3 [T. B. Brown et al., 2020], BART [Lewis et al., 2020], T5 [Raffel et al., 2020]). In this regard, while several works (Ettinger, 2020; Kassner et al., 2020) demonstrate that language models are not good at capturing how negation changes the meaning the sentences they appear in, others (Warstadt et al., 2019; Y. Zhao et al., 2020) found evidence for some form of encoding of negation at the syntactic level (to the best of our knowledge, similar studies have not been conducted in regard to speculation). As the processing of negation and speculation, as addressed in this work, is rather influenced by syntax than by semantics—i.e., the objective of the proposed systems is, in a nutshell, to decide *if*, not *how*, certain parts of a given sentence are affected by the presence of a cue—, these new paradigms may be found to be viable and even competitive for these tasks, as have been for others.

APPENDICES

Appendix A

MEDDOCAN category labels

In order to improve the readability of this document, we renamed the official labels of MEDDOCAN's sensitive data categories. The correspondences are listed below:

Table A.1: Official and renamed labels of MEDDOCAN category labels

Label (and abbreviation) in this document	Official MEDDOCAN label
Territory (`Ter`)	TERRITORIO
Date (`Dat`)	FECHAS
Patient's age (`Age`)	EDAD_SUJETO_ASISTENCIA
Patient's name (`Pat`)	NOMBRE_SUJETO_ASISTENCIA
Patient's sex (`Sex`)	SEXO_SUJETO_ASISTENCIA
Street (`Str`)	CALLE
Country (`Ctr`)	PAIS
Patient's ID (`Pid`)	ID_SUJETO_ASISTENCIA
E-mail address (`Ema`)	CORREO_ELECTRONICO
License ID (`Lid`)	ID_TITULACION_PERSONAL_SANITARIO
Insurance ID (`Iid`)	ID_ASEGURAMIENTO
Hospital (`Hos`)	HOSPITAL
Patient's relative (`Kin`)	FAMILIARES_SUJETO_ASISTENCIA
Institution (`Ins`)	INSTITUCION
Episode ID (`Eid`)	ID_CONTACTO_ASISTENCIAL
Phone number (`Pho`)	NUMERO_TELEFONO
Patient's profession (`Job`)	PROFESION
Fax number (`Fax`)	NUMERO_FAX
Other (`Oth`)	OTROS_SUJETO_ASISTENCIA
Outpatients clinic (`Cli`)	CENTRO_SALUD
Doctor's ID (`Did`)	ID_EMPLEO_PERSONAL_SANITARIO

Appendix B

MEDDOCAN confusion matrices

This appendix contains the confusion matrices of the 4 systems presented in Chapter 4: *The MEDDOCAN challenge*, namely, spaCy (Table B.1), CRF (Table B.2), NCRF$_{++}$ (Table B.3), and BERT (Table B.4).

The confusion matrices are computed at token-level, ignoring the BILOU tag. The values have been normalised by row and presented as percentages (i.e., each row sums 100% of the true labels). The column **N** indicates the number of tokens for each row in absolute terms. The rows and columns are ordered by the frequency of each category in the corpus (counted in number of spans).

As is usual in NER-like problems, all the systems manage to detect and categorise the most frequent categories with similar levels of success. The biggest differences lie in the least represented categories, located at the southeast quadrants of the matrices. The most remarkable difference in this area is that BERT's column Outside (`O`) is less populated in comparison to the other's, which means that BERT misses fewer sensitive data than the other compared systems.

Beyond that, eye-catching confusions have to do with semantically related or lexically similar categories, such as outpatients clinics (`Cli`) and institutions (`Ins`), phone numbers (`Pho`) and fax numbers (`Fax`), or identification numbers. All the systems commit these errors to varying degrees. Another common error is the confusion of mentions of a patient's relative (`Kin`) for the patient's age (`Age`), which is triggered by mentions of the age of a patient's relative, not of the patient themselves. Finally, although the matrices do not show it due to the normalisation of the values, the confusion of the categories territory (`Ter`), country (`Ctr`) and street (`Str`) is frequent as well, which is expected because they co-occur in the corpus in a sequential fashion very often.

Table B.1: Confusion matrix of the spaCy model in the MEDDOCAN challenge. Note that this confusion matrix has been split into two parts for convenience.

predicted

true	N	Ter	Dat	Age	Pat	Doc	Sex	Str	Ctr	Pid	Ema	Lid
Ter	1,090	94.59	00.18	00.00	00.09	00.00	00.00	01.83	00.37	00.00	00.00	00.00
Dat	779	00.13	97.18	00.26	00.00	00.00	00.13	00.00	00.00	00.00	00.00	00.00
Age	1,021	00.00	00.00	95.59	00.00	00.20	00.00	00.00	00.00	00.00	00.00	00.00
Pat	780	00.00	00.00	00.00	99.74	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Doc	1,693	00.00	00.00	00.00	00.00	99.94	00.00	00.00	00.00	00.00	00.00	00.00
Sex	461	00.00	00.00	00.22	00.00	00.00	98.70	00.00	00.00	00.00	00.00	00.00
Str	2,941	00.31	00.00	00.00	00.00	00.03	00.00	97.76	00.00	00.00	00.00	00.00
Ctr	370	00.81	00.27	00.00	00.00	00.00	00.00	00.00	95.41	00.00	00.00	00.00
Pid	290	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	96.55	00.00	00.00
Ema	271	00.00	00.00	00.00	00.00	00.00	00.00	03.32	00.00	00.00	95.94	00.00
Lid	683	00.29	00.00	00.00	00.00	00.00	00.00	00.29	00.00	00.00	00.00	99.41
Iid	588	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.17	00.00	00.00
Hos	560	00.18	00.00	00.00	00.00	00.00	00.00	01.43	00.71	00.00	00.00	00.00
Kin	131	00.00	01.53	03.05	00.00	00.00	00.76	00.00	00.00	01.53	00.00	00.00
Ins	250	02.00	00.00	00.00	00.00	00.40	00.00	01.60	00.40	00.00	00.00	00.00
Eid	39	00.00	00.00	00.00	00.00	02.56	00.00	00.00	00.00	00.00	00.00	00.00
Pho	67	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Job	21	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Fax	15	06.67	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	06.67	13.33
Oth	12	00.00	00.00	00.00	00.00	00.00	25.00	00.00	00.00	00.00	00.00	00.00
Cli	32	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
O	117K	00.01	00.00	00.35	00.00	00.00	00.00	00.04	00.02	00.00	00.00	00.00

true	N	Iid	Hos	Kin	Ins	Eid	Pho	Job	Fax	Oth	Cli	O
Ter	1,090	00.00	00.09	00.00	00.55	00.00	00.00	00.00	00.00	00.00	00.00	02.29
Dat	779	00.00	00.39	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	01.93
Age	1,021	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	04.21
Pat	780	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.26
Doc	1,693	00.00	00.00	00.00	00.06	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Sex	461	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	01.08
Str	2,941	00.00	00.78	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	01.12
Ctr	370	00.00	00.00	00.00	01.08	00.00	00.00	00.00	00.00	00.00	00.00	02.43
Pid	290	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	03.45
Ema	271	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.74
Lid	683	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Iid	588	99.83	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Hos	560	00.00	91.61	00.00	04.46	00.00	00.00	00.00	00.00	00.00	00.00	01.61
Kin	131	00.00	00.00	54.20	00.00	00.00	00.00	00.00	00.00	00.00	00.00	38.93
Ins	250	00.00	00.80	00.00	47.60	00.00	00.00	00.00	00.00	00.00	00.00	47.20
Eid	39	00.00	00.00	00.00	00.00	97.44	00.00	00.00	00.00	00.00	00.00	00.00
Pho	67	00.00	00.00	00.00	00.00	00.00	94.03	00.00	00.00	00.00	00.00	05.97
Job	21	00.00	00.00	00.00	00.00	00.00	00.00	42.86	00.00	00.00	00.00	57.14
Fax	15	00.00	00.00	00.00	00.00	00.00	33.33	00.00	40.00	00.00	00.00	00.00
Oth	12	00.00	00.00	08.33	00.00	00.00	00.00	00.00	00.00	00.00	00.00	66.67
Cli	32	00.00	00.00	00.00	90.62	00.00	00.00	00.00	00.00	00.00	00.00	09.38
O	117K	00.00	00.02	00.00	00.01	00.00	00.00	00.00	00.00	00.00	00.00	99.53

Table B.2: Confusion matrix of the CRF model in the MEDDOCAN challenge. Note that this confusion matrix has been split into two parts for convenience.

true	N	predicted Ter	Dat	Age	Pat	Doc	Sex	Str	Ctr	Pid	Ema	Lid
Ter	1,090	93.30	00.09	00.00	00.09	00.00	00.00	01.65	00.46	00.00	00.00	00.00
Dat	779	00.00	95.25	00.26	00.00	00.00	00.13	00.00	00.00	00.00	00.00	00.00
Age	1,021	00.00	00.00	96.77	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Pat	780	00.00	00.00	00.00	98.59	00.90	00.00	00.00	00.00	00.00	00.00	00.00
Doc	1,693	00.00	00.00	00.00	00.00	99.05	00.00	00.24	00.00	00.00	00.00	00.00
Sex	461	00.00	00.00	00.00	00.00	00.00	98.48	00.00	00.00	00.00	00.00	00.00
Str	2,941	00.44	00.00	00.00	00.00	00.17	00.00	97.18	00.00	00.00	00.00	00.00
Ctr	370	00.27	00.00	00.00	00.00	00.00	00.00	00.00	95.14	00.00	00.00	00.00
Pid	290	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	93.10	00.00	00.00
Ema	271	00.00	00.00	00.00	00.00	00.00	00.00	03.32	00.00	00.00	96.68	00.00
Lid	683	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	99.41
Iid	588	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.17	00.00	00.00
Hos	560	00.00	00.00	00.00	00.00	00.89	00.00	01.43	00.00	00.00	00.00	00.00
Kin	131	00.00	00.00	06.11	00.00	00.00	00.76	00.00	00.00	00.00	00.00	00.00
Ins	250	01.60	00.00	00.00	00.00	00.00	00.00	01.20	00.00	00.00	00.00	00.00
Eid	39	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Pho	67	08.96	04.48	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Job	21	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Fax	15	00.00	13.33	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Oth	12	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Cli	32	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
O	117K	00.01	00.01	00.01	00.00	00.01	00.00	00.02	00.00	00.00	00.00	00.00

true	N	Iid	Hos	Kin	Ins	Eid	Pho	Job	Fax	Oth	Cli	O
Ter	1,090	00.00	00.37	00.00	00.18	00.00	00.00	00.00	00.00	00.00	00.00	03.85
Dat	779	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	04.36
Age	1,021	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	03.23
Pat	780	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.51
Doc	1,693	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.71
Sex	461	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	01.52
Str	2,941	00.00	00.14	00.00	00.20	00.00	00.00	00.00	00.00	00.00	00.00	01.87
Ctr	370	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	04.59
Pid	290	01.03	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	05.86
Ema	271	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Lid	683	00.44	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.15
Iid	588	99.83	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Hos	560	00.00	88.04	00.00	02.14	00.00	00.00	00.00	00.00	00.00	00.00	07.50
Kin	131	00.00	00.00	38.93	00.00	00.00	00.00	00.00	00.00	00.00	00.00	54.20
Ins	250	00.00	00.00	00.00	30.00	00.00	00.00	00.00	00.00	00.00	00.00	67.20
Eid	39	00.00	00.00	00.00	00.00	100	00.00	00.00	00.00	00.00	00.00	00.00
Pho	67	00.00	00.00	00.00	00.00	00.00	71.64	00.00	00.00	00.00	00.00	14.93
Job	21	00.00	00.00	00.00	00.00	00.00	00.00	14.29	00.00	00.00	00.00	85.71
Fax	15	00.00	00.00	00.00	00.00	00.00	26.67	00.00	40.00	00.00	00.00	20.00
Oth	12	00.00	00.00	08.33	00.00	00.00	00.00	00.00	00.00	00.00	00.00	91.67
Cli	32	00.00	00.00	00.00	90.62	00.00	00.00	00.00	00.00	00.00	00.00	09.38
O	117K	00.00	00.00	00.00	00.01	00.00	00.00	00.00	00.00	00.00	00.00	99.93

Table B.3: Confusion matrix of the NCRF++ model in the MEDDOCAN challenge. Note that this confusion matrix has been split into two parts for convenience.

			predicted										
		N	Ter	Dat	Age	Pat	Doc	Sex	Str	Ctr	Pid	Ema	Lid
	Ter	1,090	94.77	00.09	00.00	00.09	00.00	00.00	01.83	00.55	00.00	00.00	00.00
	Dat	779	00.00	97.69	00.26	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
	Age	1,021	00.00	00.00	98.73	00.00	00.00	00.10	00.00	00.00	00.00	00.00	00.00
	Pat	780	00.00	00.00	00.00	99.74	00.00	00.00	00.00	00.00	00.00	00.00	00.00
	Doc	1,693	00.00	00.00	00.00	00.00	99.65	00.06	00.12	00.00	00.00	00.00	00.00
	Sex	461	00.00	00.00	00.00	00.00	00.00	99.13	00.00	00.00	00.00	00.00	00.00
	Str	2,941	00.20	00.00	00.00	00.00	00.07	00.00	98.98	00.00	00.00	00.00	00.00
	Ctr	370	01.62	00.00	00.00	00.00	00.00	00.00	00.00	95.41	00.00	00.00	00.00
	Pid	290	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	93.79	00.00	00.00
	Ema	271	00.00	00.00	00.00	00.00	00.00	00.00	03.32	00.00	00.00	93.73	00.00
	Lid	683	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
	Iid	588	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.17	00.00	00.00
	Hos	560	00.18	00.71	00.00	00.00	01.25	00.00	00.71	00.00	00.00	00.00	00.00
	Kin	131	00.00	00.00	10.69	00.00	00.00	02.29	00.00	00.00	00.00	00.00	00.00
	Ins	250	01.20	00.00	00.00	00.00	00.00	00.00	10.80	00.00	00.00	00.00	00.00
	Eid	39	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
	Pho	67	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	08.96
	Job	21	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	04.76	00.00	00.00
	Fax	15	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
	Oth	12	00.00	00.00	00.00	00.00	00.00	16.67	00.00	00.00	08.33	00.00	00.00
true	Cli	32	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
	O	117K	00.01	00.00	00.03	00.00	00.01	00.00	00.03	00.00	00.00	00.00	00.00

			Iid	Hos	Kin	Ins	Eid	Pho	Job	Fax	Oth	Cli	O
	Ter	1,090	00.00	00.00	00.00	00.83	00.00	00.09	00.00	00.00	00.00	00.00	01.74
	Dat	779	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	02.05
	Age	1,021	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	01.18
	Pat	780	00.00	00.00	00.13	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.13
	Doc	1,693	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.18
	Sex	461	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.87
	Str	2,941	00.00	00.00	00.00	00.14	00.00	00.00	00.00	00.00	00.00	00.00	00.61
	Ctr	370	00.00	00.00	00.00	00.27	00.00	00.00	00.00	00.00	00.00	00.00	02.70
	Pid	290	01.03	00.00	00.34	00.00	00.00	00.00	00.00	00.00	00.00	00.00	04.83
	Ema	271	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	02.95
	Lid	683	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
	Iid	588	99.83	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
	Hos	560	00.00	92.86	00.00	01.96	00.00	00.00	00.00	00.00	00.00	00.00	02.32
	Kin	131	00.00	00.00	54.96	00.00	00.00	00.00	00.00	00.00	00.00	00.00	32.06
	Ins	250	00.00	02.80	00.00	39.20	00.00	00.00	00.00	00.00	00.00	00.00	46.00
	Eid	39	00.00	00.00	00.00	00.00	100	00.00	00.00	00.00	00.00	00.00	00.00
	Pho	67	00.00	00.00	00.00	00.00	00.00	80.60	00.00	00.00	00.00	00.00	10.45
	Job	21	00.00	00.00	00.00	00.00	00.00	00.00	14.29	00.00	00.00	00.00	80.95
	Fax	15	00.00	00.00	00.00	00.00	00.00	60.00	00.00	40.00	00.00	00.00	00.00
	Oth	12	00.00	00.00	08.33	00.00	00.00	00.00	00.00	00.00	00.00	00.00	66.67
true	Cli	32	00.00	00.00	00.00	75.00	00.00	00.00	00.00	00.00	00.00	15.62	09.38
	O	117K	00.00	00.00	00.00	00.01	00.00	00.00	00.00	00.00	00.00	00.00	99.90

Table B.4: Confusion matrix of the BERT model in the MEDDOCAN challenge. Note that this confusion matrix has been split into two parts for convenience.

predicted

true	N	Ter	Dat	Age	Pat	Doc	Sex	Str	Ctr	Pid	Ema	Lid
Ter	1,090	97.25	00.09	00.00	00.09	00.00	00.00	01.01	00.46	00.00	00.00	00.00
Dat	779	00.00	99.49	00.26	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Age	1,021	00.00	00.00	99.80	00.00	00.00	00.10	00.00	00.00	00.00	00.00	00.00
Pat	780	00.00	00.00	00.00	100	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Doc	1,693	00.00	00.00	00.00	00.00	99.94	00.00	00.00	00.00	00.00	00.00	00.00
Sex	461	00.00	00.00	00.00	00.00	00.00	99.35	00.00	00.00	00.00	00.00	00.00
Str	2,941	00.34	00.00	00.00	00.00	00.07	00.00	99.08	00.00	00.00	00.00	00.00
Ctr	370	00.00	00.00	00.00	00.54	00.00	00.00	00.00	99.19	00.00	00.00	00.00
Pid	290	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	99.66	00.00	00.00
Ema	271	00.00	00.00	00.00	00.00	00.00	00.00	03.32	00.00	00.00	96.68	00.00
Lid	683	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
Iid	588	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.17	00.00	00.00
Hos	560	00.18	00.00	00.00	00.00	00.00	00.00	01.07	00.00	00.00	00.00	00.00
Kin	131	00.00	00.00	07.63	00.76	00.00	02.29	00.00	00.00	02.29	00.00	00.00
Ins	250	00.80	00.00	00.00	00.00	00.00	00.00	04.00	00.40	00.00	00.00	00.00
Eid	39	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	02.56
Pho	67	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Job	21	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	09.52	00.00	00.00
Fax	15	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Oth	12	00.00	00.00	08.33	00.00	00.00	08.33	00.00	00.00	16.67	00.00	00.00
Cli	32	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
O	117K	00.01	00.00	00.02	00.00	00.00	00.01	00.02	00.00	00.01	00.00	00.00

true	N	Iid	Hos	Kin	Ins	Eid	Pho	Job	Fax	Oth	Cli	O
Ter	1,090	00.00	00.00	00.00	00.28	00.00	00.09	00.00	00.00	00.00	00.00	00.73
Dat	779	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.26
Age	1,021	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.10
Pat	780	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Doc	1,693	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.06
Sex	461	00.00	00.00	00.22	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.43
Str	2,941	00.00	00.17	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.34
Ctr	370	00.00	00.00	00.00	00.27	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Pid	290	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.34
Ema	271	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Lid	683	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Iid	588	99.83	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Hos	560	00.00	96.79	00.00	01.07	00.00	00.00	00.00	00.00	00.00	00.00	00.89
Kin	131	00.00	00.00	67.94	00.00	00.00	00.00	00.00	00.00	00.00	00.00	19.08
Ins	250	00.00	00.80	00.00	69.60	00.00	00.00	00.00	00.00	00.00	00.00	24.40
Eid	39	00.00	00.00	00.00	00.00	97.44	00.00	00.00	00.00	00.00	00.00	00.00
Pho	67	00.00	00.00	00.00	00.00	00.00	98.51	00.00	00.00	00.00	00.00	01.49
Job	21	00.00	00.00	00.00	00.00	00.00	00.00	76.19	00.00	00.00	00.00	14.29
Fax	15	00.00	00.00	00.00	00.00	00.00	26.67	00.00	73.33	00.00	00.00	00.00
Oth	12	00.00	00.00	08.33	00.00	00.00	00.00	33.33	00.00	00.00	00.00	25.00
Cli	32	00.00	00.00	00.00	09.38	00.00	00.00	00.00	00.00	00.00	90.62	00.00
O	117K	00.00	00.00	00.01	00.02	00.00	00.00	00.01	00.00	00.00	00.00	99.89

Appendix C

NUBes: medical specialities and EHR sections

Table C.1 contains the average frequency per every 100 tokens of each sensitive data category in the NUBes-PHI corpus. The upper table section breaks down this information into medical specialities, while the lower section does the same for Electronic Health Record (EHR) sections.

As can be seen, reports from Obstetrics and Gynaecology (OBG) contain remarkably more sensitive information than the other specialities in relative terms—it contains particularly more doctor names (`Doc`)—, followed by Thoracic Surgery (TS) and Ophthalmology (OPH). Opposite this spectrum are specialities Plastic Surgery (PS) and Odontology (ODO). It must be noted, however, that the documents belong to the same hospital, and the number of doctors that authored them is unknown to us; in addition, some specialities are hardly represented in the dataset. It is then possible that these number simply describe the mannerisms of a few doctors. Perhaps more interestingly, Treatment Notess (TNo) and Chief Complaint (CC) are the sections that contain more sensitive information (double the average). Treatment Notes (TNo) abounds particularly with dates (`Dat`) and doctor names (`Doc`) in comparison to the other sections, while Chief Complaint (CC) has the most mentions of age of patients (`Age`) and sex of patients (`Sex`). Physical Examination (PE) is the section with least sensitive information in this comparison.

Table C.2 describes the distribution of negation and speculation annotations in the NUBes corpus, also by medical speciality (upper table section) and EHR section (lower table section). When analysed over medical specialities, Plastic Surgery (PS) and Neurology (N) reports stand out in particular for their high usage of speculative expressions; negation, on the other hand, is most frequent, in relative terms, in Cardiovascular Diseases (CD) reports (ignoring the least frequent specialities). Regarding the EHR sections, text under the Diagnostic Tests (DXT) section contributes the most negation and speculation examples, followed by History of Present Illness (HPI) and PE, while TNo hardly contain any of these phenomena.

Table C.1: Average sensitive data frequency per every 100 tokens by category and medical speciality (upper section), and EHR section (lower section). **Doc** = number of documents; **Len** = average document length in tokens. ***Ave*** = average of all medical specialities or EHR sections. The rest of abbreviations and acronyms are defined in the glossary at the end of this document.

	Doc	Len	Dat	Fac	Age	Doc	Sex	Kin	Loc	Pat	Job	Con	Oth	Tot
OBG	394	15.76	3.33	0.55	0.05	1.09	0.02	0.00	0.05	0.00	0.00	0.00	0.00	5.08
TS	18	25.17	2.00	0.00	0.44	0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.88
OPH	241	23.65	1.29	1.11	0.18	0.02	0.02	0.00	0.02	0.00	0.00	0.00	0.00	2.62
HaH	925	120.72	1.15	0.51	0.24	0.25	0.14	0.07	0.03	0.06	0.00	0.00	0.01	2.46
U	463	68.51	1.60	0.59	0.10	0.08	0.04	0.01	0.02	0.00	0.00	0.00	0.00	2.45
OTO	536	35.96	0.97	0.29	0.14	0.31	0.12	0.01	0.02	0.00	0.01	0.00	0.01	1.86
ICU	219	73.28	0.95	0.29	0.19	0.04	0.06	0.03	0.02	0.00	0.00	0.00	0.00	1.58
Ave		73.55	0.78	0.28	0.19	0.13	0.07	0.05	0.02	0.01	0.01	0.00	0.01	1.55
GCU	1,021	59.99	0.92	0.28	0.07	0.10	0.00	0.02	0.02	0.00	0.00	0.00	0.00	1.42
CD	513	92.74	0.47	0.08	0.59	0.09	0.07	0.04	0.01	0.00	0.01	0.00	0.00	1.37
GE	4	40.00	0.62	0.00	0.00	0.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.25
GS	394	108.46	0.51	0.17	0.17	0.11	0.12	0.03	0.01	0.00	0.01	0.00	0.00	1.14
OR	393	79.61	0.39	0.11	0.22	0.03	0.17	0.02	0.02	0.00	0.10	0.00	0.02	1.08
IM	507	112.46	0.47	0.12	0.21	0.07	0.05	0.06	0.02	0.00	0.00	0.00	0.01	1.00
VS	3	41.33	0.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.81
N	552	135.37	0.28	0.11	0.04	0.03	0.02	0.09	0.02	0.00	0.00	0.00	0.00	0.58
AN	16	25.88	0.00	0.00	0.00	0.24	0.00	0.00	0.00	0.00	0.24	0.00	0.00	0.48
PS	805	14.39	0.22	0.08	0.04	0.03	0.02	0.00	0.00	0.00	0.00	0.03	0.00	0.42
ODO	15	8.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TNo	686	84.90	2.14	0.60	0.01	0.60	0.00	0.00	0.02	0.00	0.00	0.02	0.01	3.40
CC	1,878	25.78	0.66	0.56	1.19	0.18	0.48	0.04	0.03	0.00	0.00	0.00	0.01	3.16
HPI	1,664	78.27	0.85	0.22	0.16	0.04	0.11	0.10	0.04	0.01	0.03	0.00	0.01	1.58
Ave		73.55	0.78	0.28	0.19	0.13	0.07	0.05	0.02	0.01	0.01	0.00	0.01	1.55
hx	118	39.36	0.93	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95
PNo	1,677	125.48	0.47	0.22	0.08	0.07	0.01	0.03	0.01	0.02	0.00	0.00	0.00	0.91
DXT	376	90.62	0.59	0.06	0.07	0.04	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.80
PE	620	51.63	0.40	0.11	0.01	0.06	0.00	0.03	0.02	0.03	0.00	0.00	0.01	0.67

Table C.2: Average negation and uncertainty marker frequency per every 100 tokens, by category and medical speciality (upper section), and by EHR section (lower section). **Doc** = number of documents; **Len** = average document length in tokens. ***Ave*** = average of all medical specialities or EHR sections. The rest of abbreviations and acronyms are defined in the glossary at the end of this document.

	Doc	Len	Neg	NSyn	NLex	NMph	Unc	ULex	USyn	Tot
CD	513	92.74	2.71	1.86	0.39	0.47	0.27	0.27	0.00	2.98
OR	393	79.61	2.19	1.80	0.33	0.06	0.57	0.56	0.01	2.76
N	552	135.37	1.74	1.08	0.38	0.28	0.91	0.90	0.01	2.65
U	463	68.51	2.06	1.44	0.49	0.12	0.58	0.58	0.00	2.64
GS	394	108.46	2.12	1.57	0.30	0.26	0.50	0.50	0.00	2.63
OTO	241	35.96	2.26	1.97	0.19	0.10	0.32	0.32	0.00	2.58
GE	4	40.00	2.50	2.50	0.00	0.00	0.00	0.00	0.00	2.50
AN	16	25.88	1.45	1.45	0.00	0.00	0.97	0.97	0.00	2.42
Ave		73.81	1.82	1.28	0.35	0.19	0.50	0.49	0.00	2.32
TS	18	25.17	2.21	2.21	0.00	0.00	0.00	0.00	0.00	2.21
IM	507	112.46	1.73	1.19	0.44	0.10	0.46	0.46	0.00	2.19
OBG	394	15.76	1.77	1.43	0.34	0.00	0.37	0.37	0.00	2.14
ICU	219	73.28	1.80	1.41	0.31	0.07	0.33	0.32	0.01	2.13
PS	805	14.39	0.90	0.75	0.13	0.02	1.09	1.09	0.00	1.99
HaH	925	120.72	1.59	1.02	0.36	0.21	0.31	0.31	0.01	1.90
GCU	1,021	59.99	1.39	1.01	0.32	0.07	0.51	0.51	0.00	1.90
ODO	15	8.07	0.00	0.00	0.00	0.00	0.83	0.83	0.00	0.83
OPH	536	23.65	0.33	0.30	0.02	0.02	0.11	0.11	0.00	0.44
VS	3	41.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DXT	376	90.62	2.10	1.90	0.15	0.04	1.21	1.21	0.01	3.31
HPI	1,664	78.27	2.23	1.72	0.33	0.17	0.53	0.52	0.00	2.75
PE	620	51.63	2.00	1.65	0.17	0.18	0.60	0.60	0.01	2.60
PNo	1,677	125.48	1.90	1.11	0.50	0.30	0.52	0.52	0.00	2.43
Ave		73.81	1.82	1.28	0.35	0.19	0.50	0.49	0.00	2.32
CC	1,878	25.78	1.85	1.49	0.25	0.11	0.34	0.33	0.00	2.19
hx	118	39.36	1.18	1.01	0.15	0.02	0.06	0.06	0.00	1.25
TNo	686	84.90	0.38	0.18	0.20	0.00	0.03	0.03	0.00	0.41

Appendix D

NUBes-PHI confusion matrices

Table D.1: Confusion matrices of spaCy for the classification task on NUBes-PHI. The matrices have been computed with token-level predictions without taking the BIO tags into account.

(a) Model trained on the MEDDOCAN corpus

true	N	Dat	Fac	Age	Tim	Doc	Sex	Kin	Loc	Pat	Job	Oth	O (predicted)
Dat	1,479	39.76	00.00	00.81	00.00	00.00	00.00	00.14	00.61	00.00	00.00	00.27	58.42
Fac	557	02.15	12.57	00.00	00.00	00.00	00.00	00.18	00.36	00.00	00.00	00.00	84.74
Age	574	00.00	00.00	58.71	00.00	00.00	00.00	00.00	00.00	00.35	00.00	00.00	40.94
Tim	407	06.14	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	93.86
Doc	401	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.25	00.25	00.00	00.00	99.50
Sex	71	00.00	00.00	00.00	00.00	00.00	81.69	01.41	00.00	00.00	00.00	00.00	16.90
Kin	44	00.00	00.00	00.00	00.00	00.00	00.00	59.09	00.00	00.00	00.00	00.00	40.91
Loc	26	00.00	07.69	00.00	00.00	00.00	00.00	00.00	03.85	00.00	00.00	00.00	88.46
Pat	14	00.00	00.00	00.00	00.00	00.00	00.00	28.57	00.00	00.00	00.00	00.00	71.43
Job	17	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
Oth	1	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
O	103K	00.07	00.02	00.00	00.00	00.00	00.01	00.08	00.05	00.02	00.00	00.00	99.75

(b) Model trained on NUBes-PHI

true	N	Dat	Fac	Age	Tim	Doc	Sex	Kin	Loc	Pat	Job	Oth	O (predicted)
Dat	1,479	92.83	00.00	01.35	00.68	00.14	00.00	00.00	00.00	00.00	00.00	00.00	05.00
Fac	557	00.00	88.33	00.00	00.00	00.54	00.00	00.00	00.90	00.18	00.00	00.00	10.05
Age	574	00.00	00.00	97.56	00.35	00.00	00.00	00.00	00.00	00.00	00.00	00.00	02.09
Tim	407	00.74	00.00	00.00	95.09	00.00	00.00	00.00	00.00	00.00	00.00	00.00	04.18
Doc	401	00.00	00.00	00.00	00.00	94.51	00.00	00.00	00.00	00.00	00.00	00.00	05.49
Sex	71	00.00	00.00	00.00	00.00	00.00	100	00.00	00.00	00.00	00.00	00.00	00.00
Kin	44	00.00	00.00	00.00	00.00	00.00	00.00	95.45	00.00	00.00	00.00	00.00	04.55
Loc	26	00.00	15.38	00.00	00.00	00.00	00.00	00.00	26.92	00.00	00.00	00.00	57.69
Pat	14	00.00	00.00	00.00	00.00	07.14	00.00	00.00	07.14	21.43	00.00	00.00	64.29
Job	17	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	11.76	00.00	88.24
Oth	1	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
O	103K	00.06	00.03	00.00	00.01	00.01	00.00	00.01	00.00	00.00	00.00	00.00	99.88

Table D.2: Confusion matrices of NCRF++ for the classification task on NUBes-PHI. The matrices have been computed with token-level predictions without taking the BIO tags into account.

(a) Model trained on the MEDDOCAN corpus

			predicted										
		N	Dat	Fac	Age	Tim	Doc	Sex	Kin	Loc	Pat	Job	Oth	O
	Dat	1,479	51.12	00.00	00.81	00.00	00.00	00.00	00.07	00.07	00.00	00.00	00.00	47.94
	Fac	557	00.00	18.49	00.00	00.00	00.00	00.00	00.00	01.80	00.54	00.00	00.00	79.17
	Age	574	00.00	00.00	60.10	00.00	00.00	00.00	02.96	00.17	00.00	00.00	00.00	36.76
	Tim	407	00.74	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	99.26
	Doc	401	00.00	00.00	00.00	00.00	01.50	00.00	03.74	00.00	01.00	00.00	00.00	93.77
	Sex	71	00.00	00.00	00.00	00.00	00.00	66.20	22.54	00.00	00.00	00.00	00.00	11.27
	Kin	44	00.00	00.00	00.00	00.00	00.00	04.55	59.09	00.00	00.00	00.00	00.00	36.36
	Loc	26	00.00	00.00	00.00	00.00	00.00	00.00	00.00	30.77	00.00	00.00	00.00	69.23
	Pat	14	00.00	00.00	00.00	00.00	00.00	00.00	28.57	14.29	00.00	00.00	00.00	57.14
	Job	17	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
true	Oth	1	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
	O	103K	00.04	00.04	00.00	00.00	00.00	00.00	00.09	00.07	00.00	00.00	00.01	99.74

(b) Model trained on NUBes-PHI

			predicted										
		N	Dat	Fac	Age	Tim	Doc	Sex	Kin	Loc	Pat	Job	Oth	O
	Dat	1,479	94.39	00.00	01.49	00.68	00.14	00.00	00.00	00.00	00.00	00.00	00.00	03.31
	Fac	557	01.08	91.56	00.00	00.00	00.72	00.00	00.18	00.90	00.36	00.00	00.00	05.21
	Age	574	00.00	00.00	98.78	00.35	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.87
	Tim	407	00.98	00.00	00.00	96.81	00.00	00.00	00.00	00.00	00.00	00.00	00.00	02.21
	Doc	401	00.00	02.00	00.00	00.00	95.76	00.25	00.00	00.00	00.00	00.00	00.00	02.00
	Sex	71	00.00	00.00	00.00	00.00	00.00	100	00.00	00.00	00.00	00.00	00.00	00.00
	Kin	44	00.00	00.00	00.00	00.00	02.27	00.00	93.18	00.00	00.00	00.00	00.00	04.55
	Loc	26	00.00	19.23	00.00	00.00	00.00	00.00	00.00	34.62	07.69	00.00	00.00	38.46
	Pat	14	07.14	00.00	00.00	00.00	07.14	00.00	00.00	00.00	71.43	00.00	00.00	14.29
	Job	17	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	05.88	00.00	94.12
true	Oth	1	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
	O	103K	00.12	00.03	00.01	00.07	00.00	00.00	00.01	00.00	00.00	00.00	00.00	99.75

Appendix E

Transformers vocabulary overlap with NUBes

Table E.1 describes the Transformers models tested in Chapters 11 and 12 in terms of their vocabulary overlap with NUBES. For comparison purposes, the same table reports the vocabulary overlap with SFU Review$_{SP}$-NEG (Jiménez-Zafra et al., 2018c), a corpus of product reviews in Spanish.

SHA is the percentage of unique words in the corpus that is covered by the vocabulary. wSHA is the percentage of all the words in the corpus (i.e., frequency weighted unique words) that is covered by the vocabulary, after removing stopwords. Similarly, UNK is the percentage of unique words in the corpus for which the tokenizer yielded the special token `[UNK]` (or analogous) and wUNK is the frequency weighted UNK (without stopwords).

The models are shown by weighted coverage in the NUBES corpus in descending order. As can be seen, the greatest vocabulary coverage, provided by SpanBERTa [26], is 28.47%. That is, 28.47% of the set of words occurring in NUBES have their own embedding. When weighted by word frequency, the coverage rises to 69.67% of the corpus. The worst model in this regard is, unsurprisingly, SciBERT (Beltagy et al., 2019)—a monolingual English model—, with just 6.02% vocabulary overlap with NUBES.

Table E.1: Vocabulary coverage by the pre-trained language models

	Vocab	**NUBes**				**SFU Review$_{SP}$-NEG**			
		SHA	wSHA	UNK	wUNK	SHA	wSHA	UNK	wUNK
SpanBERTa$_{Base}$ Cased	50,265	28.47	69.67	0.00	0.00	55.44	86.47	0.00	0.00
IXAmBERT$_{Base}$ Cased	119,101	25.63	66.84	0.73	0.31	49.10	79.41	0.55	1.12
BETO$_{Base}$ Cased	31,002	21.72	62.25	0.78	0.37	41.05	77.12	0.30	0.73
RoBERTa$_{Base}$ BNE	50,262	26.17	51.71	0.00	0.00	51.42	63.13	0.00	0.00
mBERT$_{Base}$ Cased	119,547	12.97	50.56	0.00	0.09	25.32	63.75	0.04	0.34
XLM-RoBERTa$_{Base}$	250,002	14.40	38.68	0.00	0.00	26.00	49.65	0.00	0.00
SciBERT$_{scivocab}$ Cased	31,116	6.02	29.93	0.24	0.25	7.99	33.12	0.11	0.29

Appendix F

Hyperparameters of the negation and uncertainty detection models

Table F.1: Hyperparameteres of the neural sequence taggers and text classifiers. Values between squares brackets are options or ranges for the hyperparameter optimisation. Any hyperparameter not reported here takes the default value given by the corresponding training API.

(a) NCRF++ sequence tagger (from [Lima-López et al., 2020a])

Hyperparameter	Value	Hyperparameter	Value
Character emb. dimensions	30	Batch size	16
Character CNN layers	1	Optimiser	SGD
Character hidden dimensions	50	Learning rate	0.005
Word emb. dimensions	300	L_2 regularisation	1e-8
Word bi-LSTM layers	1	Weight decay	0.001
Word hidden dimensions	200	Momentum	0
Dropout rate	0.5	Maximum epochs	40

(b) Flair sequence tagger and text classifier

Hyperparameter	Value	Hyperparameter	Value
Pre-trained word emb.	MWES	Batch size	[8, 16, 32]
Pre-trained Flair emb.	`es-forward`,	Optimiser	SGD
	`es-backward`	Learning rate	[0.05 - 0.15]
bi-LSTM/GRU layers	1	Minimum learning rate	1e-4
Hidden dimensions	[128, 256]	Weight decay	[0.0 - 0.05]
Dropout rate	[0.0 - 0.5]	Maximum epochs	60

(c) Transformer sequence taggers and text classifiers

Hyperparameter	Value	Hyperparameter	Value
Pre-trained model	*see Table 11.3*	Learning rate	[1e-5 - 1e-4]
Batch size	8	Warmup steps	[0 - 500]
Maximum input length	220	Weight decay	0.0 to 0.3
Optimiser	AdamW	Maximum epochs	30

Appendix G

Additional metrics for the experiments on negation and uncertainty detection

This appendix contains complementary result metrics of the experiments in Chapter 11: *Experiments in cue and scope detection* (Tables G.1 through G.4) and Chapter 12: *Experiments in assertion classification* (Tables G.5 and G.6).

Table G.3 reports the performance of the sequence labelling models in terms of the metrics described by Morante et al. (2012a) for the *SEM 2012 shared task on resolving the scope and focus of negation, later also employed in the NEGES workshops (Jiménez-Zafra et al., 2018a, 2019), among others. The evaluation script is publicly available from the official website of the shared task [72]. Notice that the script is prepared to count one type of cues and one type of scopes (namely, negation cues and scopes). In order to report separate scores for negation and speculation, we post-processed the outputs of the systems to contain just negation or uncertainty predictions, then applied the evaluation script.

The table includes the results of Hartmann et al. (2021), who tackle the resolution of negation scopes. Their supervised variant, consisting of a fine-tuned mBERT, outperforms all of our systems when looking at the detection of negation scopes. It must be noted, however, that our models target 3 more entity types jointly (namely, negation cues, and speculation cues and scopes).

Table G.4 reports the performance of the sequence labelling models in terms of the metric described by Solarte Pabón et al. (2022), to which we refer as 'BIO-weighted token-level' scores throughout this work. In principle, the only difference between the mBERT model reported here and that of Solarte Pabón et al. (2022) ($_{SP}$ in the table) is the optimisation of some hyperparameters (see Section 11.2.2.4), whose impact is most noticeable for uncertainty scopes, the most challenging category of all.

Table G.1: Precision results for cue and scope detection in the Full test set. The best and second-best scores are highlighted in bold and dotted underlines, respectively. N is the number of training examples.

	$1/3^4$ train set (N=169)					Full train set (N=13,802)				
	μ	NCue	NSco	UCue	USco	μ	NCue	NSco	UCue	USco
NCRF++	0.738	0.853	0.678	0.297	0.276	0.894	0.955	0.879	**0.875**	0.732
Flair+fT	0.777	0.895	0.737	0.615	0.365	0.887	0.954	0.878	0.834	0.736
BETO	0.764	0.857	0.723	**0.766**	**0.430**	0.900	0.960	0.899	0.864	0.736
SpanBERTa	0.743	0.877	0.662	0.717	0.331	0.895	0.954	0.897	0.843	0.735
MarIA	0.735	0.858	0.695	0.593	0.395	**0.911**	**0.966**	**0.902**	0.864	**0.785**
IXAmBERT	**0.795**	**0.897**	**0.790**	0.712	0.409	0.901	0.960	0.889	0.867	0.759
mBERT	0.770	0.891	0.728	0.704	0.354	0.897	0.961	0.887	0.839	0.760
XLM-R	0.777	0.874	0.758	0.692	0.422	0.897	0.956	0.891	0.843	0.766
SciBERT	0.751	0.864	0.697	0.732	0.190	0.888	0.958	0.867	0.847	0.750

Table G.2: Recall results for cue and scope detection in the Full test set. The best and second-best scores are highlighted in bold and dotted underlines, respectively. N is the number of training examples.

	$1/3^4$ train set (N=169)					Full train set (N=13,802)				
	μ	NCue	NSco	UCue	USco	μ	NCue	NSco	UCue	USco
NCRF++	0.511	0.702	0.582	0.055	0.052	0.868	0.950	0.852	0.825	0.667
Flair+fT	0.620	0.812	0.640	0.335	0.155	0.897	0.966	0.877	0.865	0.745
BETO	**0.708**	**0.866**	**0.733**	0.515	0.255	0.911	0.966	0.900	0.875	0.782
SpanBERTa	0.646	0.854	0.638	0.430	0.150	0.901	0.966	0.890	0.858	0.750
MarIA	0.683	0.852	0.703	0.477	0.220	0.910	0.969	0.893	**0.887**	0.777
IXAmBERT	0.674	0.814	0.690	**0.532**	**0.265**	0.902	**0.970**	0.887	0.863	0.750
mBERT	0.666	0.842	0.675	0.475	0.198	0.899	0.959	0.887	0.863	0.760
XLM-R	0.689	0.855	0.697	0.495	0.263	**0.914**	0.967	**0.901**	0.885	**0.795**
SciBERT	0.617	0.855	0.595	0.383	0.080	0.893	0.960	0.869	0.875	0.750

Table G.3: *SEM F1 scores for cue and scope detection in the Full test set. The best and second-best scores are highlighted in bold and dotted underlines, respectively. We refer the reader to Morante et al. (2012a) for an explanation of each metric. We include the results of Hartmann et al. (2021), who tackle the resolution of negation scopes: SU is a supervised mBERT model, while ZS ST_{cat} refers to zero-shot performance of a mBERT model trained on the BioScope corpus (Vincze et al., 2008) and the SFU Review Corpus (Konstantinova et al., 2012).

	Negation						Speculation					
	Cues	Scopes			Glob	CNS	Cues	Scopes			Glob	CNS
		CM	NCM	Token				CM	NCM	Token		
NCRF++	94.68	88.38	88.85	90.51	88.67	81.54	84.68	75.39	75.60	75.52	75.00	64.41
Flair+fT	95.38	89.49	90.01	91.58	89.38	83.22	85.71	77.89	78.69	78.67	77.83	69.71
BETO	95.78	90.86	91.76	**93.27**	90.88	85.42	86.44	80.32	81.07	81.62	80.32	**74.41**
SpanBERTa	95.50	90.63	91.37	92.81	90.57	84.81	85.19	78.18	78.97	79.86	77.92	70.59
MarIA	**96.31**	**91.42**	**92.03**	93.17	**91.48**	**85.78**	**86.72**	80.32	80.91	82.36	80.13	72.94
IXAmBERT	96.06	90.32	90.94	92.81	90.47	84.72	85.93	78.47	79.87	81.53	78.42	70.00
mBERT	95.49	90.62	91.20	92.51	90.66	84.89	86.19	78.83	79.80	79.31	78.64	71.47
XLM-R	95.77	90.98	91.66	93.24	90.97	85.42	86.58	**80.71**	**81.85**	**83.02**	**80.77**	74.12
SciBERT	95.40	89.05	89.74	91.83	89.21	82.51	86.19	77.83	78.83	79.58	77.65	70.00
SU	-	-	-	95.66	-	-	-	-	-	-	-	-
ZS ST_{cat}	-	-	-	90.24	-	-	-	-	-	-	-	-

Table G.4: BIO-tag weighted token-level scores (from Solarte Pabón et al. [2022]) for cue and scope detection in the Full test set. The best and second-best scores are highlighted in bold and dotted underlines, respectively. mBERT_{SP} is the system presented by Solarte Pabón et al. (2022).

	NCue			NSco			UCue			USco		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
NCRF++	0.95	0.94	0.95	0.93	0.88	0.90	0.87	0.82	0.85	0.84	0.69	0.76
Flair+fT	0.95	**0.97**	0.96	0.92	0.90	0.91	0.84	0.87	0.86	0.80	0.79	0.79
BETO	0.95	**0.97**	0.96	0.94	**0.92**	**0.93**	0.86	0.88	0.87	0.80	0.84	0.82
SpanBERTa	0.95	**0.97**	0.96	0.94	0.91	0.92	0.85	0.87	0.86	0.80	0.82	0.81
MarIA	**0.97**	**0.97**	**0.97**	**0.95**	0.91	**0.93**	**0.88**	0.88	**0.88**	0.84	0.82	0.83
IXAmBERT	0.96	**0.97**	0.96	0.94	0.91	**0.93**	0.86	0.87	0.87	**0.85**	0.78	0.81
mBERT	0.96	0.96	0.96	0.94	0.91	0.92	0.86	0.87	0.86	0.80	0.81	0.81
mBERT_{SP}	0.95	0.93	0.95	0.90	0.86	0.88	0.86	0.83	0.84	0.75	0.70	0.72
XLM-R	0.95	**0.97**	0.96	0.93	**0.92**	**0.93**	0.85	**0.89**	0.87	**0.82**	**0.85**	**0.84**
SciBERT	0.95	0.96	0.96	0.92	0.91	0.91	0.86	0.88	0.87	0.81	0.81	0.81

Table G.5: Precision results for assertion classification. The best and second-best scores are highlighted in bold and dotted underlines, respectively. N is the number of training examples.

	Full test						Man test		
	$1/3^4$ train (N=148)			Full train (N=12,108)			Full train (N=12,108)		
	μ	abs	pos	μ	abs	pos	μ	abs	pos
NegEx	0.643	0.631	**0.711**	0.583	0.579	0.597	0.945	0.950	0.925
Flair+fT	0.167	0.200	0.000	0.874	0.867	0.891	0.978	0.982	0.962
BETO	0.607	0.805	0.345	0.915	0.916	0.914	0.987	**0.995**	0.965
SpanBERTa	**0.790**	0.808	0.672	0.906	0.910	0.896	0.984	0.990	0.966
MarIA	0.655	0.770	0.316	**0.924**	**0.921**	0.933	0.987	**0.995**	0.965
IXAmBERT	0.748	0.800	0.478	0.906	0.911	0.895	0.981	0.990	0.955
mBERT	0.666	0.803	0.421	0.909	0.910	0.907	0.988	**0.995**	0.969
XLM-R	0.636	0.820	0.272	0.906	0.893	**0.938**	**0.991**	0.994	**0.984**
SciBERT	0.666	**0.841**	0.224	0.908	0.906	0.914	0.986	0.989	0.977

Table G.6: Recall results for assertion classification. The best and second-best scores are highlighted in bold and dotted underlines, respectively. N is the number of training examples.

	Full test						Man test		
	$1/3^4$ train (N=148)			Full train (N=12,108)			Full train (N=12,108)		
	μ	abs	pos	μ	abs	pos	μ	abs	pos
NegEx	0.651	0.780	0.350	0.825	0.885	0.685	0.841	0.895	0.679
Flair+fT	0.002	0.002	0.000	0.906	0.920	0.873	0.903	0.921	0.850
BETO	0.616	0.665	**0.503**	0.954	0.972	0.914	0.957	0.963	0.939
SpanBERTa	0.566	0.715	0.218	0.950	0.965	0.914	0.950	0.952	0.945
MarIA	0.534	0.670	0.218	0.950	0.961	0.924	0.956	0.963	0.936
IXAmBERT	0.482	0.617	0.168	0.944	0.959	0.909	0.935	0.945	0.905
mBERT	0.607	0.672	0.457	0.962	0.970	**0.944**	0.958	0.961	**0.951**
XLM-R	**0.658**	**0.804**	0.315	**0.963**	**0.978**	0.929	**0.965**	**0.975**	0.936
SciBERT	0.349	0.450	0.112	0.947	0.959	0.919	0.948	0.961	0.911

Bibliography

Accuosto, Pablo and Horacio Saggion (2018). "Improving the accessibility of biomedical texts by semantic enrichment and definition expansion". In: *Procesamiento del Lenguaje Natural* 61, pp. 57–64.

Agerri, Rodrigo, Josu Bermudez and German Rigau (2014). "IXA pipeline: Efficient and ready to use multilingual NLP tools". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)* (Reykjavik, Iceland, 26th–31st May 2014). European Language Resources Association, pp. 3823–3828.

Agirre, Eneko and Aitor Soroa (2009). "Personalizing PageRank for Word Sense Disambiguation". In: *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (Athens, Greece, 30th Mar.–3rd Apr. 2009). Association for Computational Linguistics, pp. 33–41.

Agirre, Eneko, Aitor Soroa and Mark Stevenson (2010). "Graph-based Word Sense Disambiguation of biomedical documents". In: *Bioinformatics* 26.22, pp. 2889–2896.

Ajausks, Ēriks, Victoria Arranz, Laurent Bié, Aleix Cerdà-i-Cucó, Khalid Choukri, Montse Cuadros, Hans Degroote, Amando Estela, Thierry Etchegoyhen, Mercedes García-Martínez, Aitor García-Pablos, Manuel Herranz, Alejandro Kohan, Maite Melero, Mike Rosner, Roberts Rozis, Patrick Paroubek, Artūrs Vasiļevskis and Pierre Zweigenbaum (2020). "The Multilingual Anonymisation toolkit for Public Administrations (MAPA) project". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT 2020)* (Lisboa, Portugal, 3rd–5th Nov. 2020). European Association for Machine Translation, pp. 471–472.

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter and Roland Vollgraf (2019). "FLAIR: An easy-to-use framework for state-of-the-art NLP". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies (NAACL-HLT 2019), Demonstrations (Minneapolis, MN, USA, 2nd–7th June 2019). Association for Computational Linguistics, pp. 54–59.

Akbik, Alan, Duncan Blythe and Roland Vollgraf (2018). "Contextual string embeddings for sequence labeling". In: *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)* (Santa Fe, NM, USA, 20th–26th Aug. 2018). Association for Computational Linguistics, pp. 1638–1649.

Almagro, Mario, Raquel Martínez, Víctor Fresno and Soto Montalvo (2020). "ICD-10 coding of Spanish electronic discharge summaries: An extreme classification problem". In: *IEEE Access* 8, pp. 100073–100083.

Almeida, João Rafael and Sérgio Matos (2020). "Rule-Based extraction of family history information from clinical notes". In: *Proceedings of the 35th Annual ACM Symposium on Applied Computing (SAC 2020)* (Brno, Czech Republic, 30th Mar.–3rd Apr. 2020). Association for Computing Machinery, pp. 670–675.

Alsentzer, Emily, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann and Matthew McDermott (2019). "Publicly available clinical BERT embeddings". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP 2019)* (Minneapolis, MN, USA, 7th June 2019). Association for Computational Linguistics, pp. 72–78.

Amézqueta Goñi, Carlos, Alberto Andérez González, Javier Carnicero Giménez de Azcaráte, Miguel Chavarria Díaz, Pere Crespo Molina, Fernando Escolar Castellón, José A. Falagan Mota, José Antonio Garbayo Sánchez, Marcial García Rojo, Ana Granado Hualde, Carlos Hernández Salvador, Margarita Iraburu Elizondo, Elena Manso Montes, Fernando Martín Sánchez, José Luis Monteagudo Peña, José Alberto Maldonado Segura, Javier Nogueira Fariña, Juan Reig Redondo, Montserrat Robles Viejo, Carlos Sánchez García, Jokin Sanz Ureta, Tone M. S. Birkenes and José Manuel Vázquez López (2003). "De la historia clínica a la historia de salud electrónica". Tech. rep. Sociedad Española de Informática de la Salud.

Aramaki, Eiji, Takeshi Imai, Kengo Miyo and Kazuhiko Ohe (2006). "Automatic deidentification by using sentence features and label consistency". In: *Proceedings of the i2b2 Workshop on NLP Challenges for Clinical Data* (Washington D.C., USA, 10th–11th Nov. 2006), pp. 10–11.

Aronson, Alan R. (2001). "Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program". In: *Proceedings of the AMIA Annual Symposium* (Washington D.C., USA, 3rd–7th Nov. 2001). American Medical Informatics Association, pp. 17–21.

— (2006). "MetaMap: Mapping text to the UMLS Metathesaurus". Tech. rep. NLM, NIH, DHHS.

Attardi, Giuseppe, Andrea Buzzelli and Daniele Sartiano (2013). "Machine translation for entity recognition across languages in biomedical documents". In: *Working Notes for CLEF 2013 Conference (CLEF 2013)* (Valencia, Spain, 23rd–26th Sept. 2013). CEUR Workshop Proceedings, pp. 1–6.

Báez, Pablo, Felipe Bravo-Marquez, Jocelyn Dunstan, Matías Rojas and Fabián Villena (2022). "Automatic extraction of nested entities in clinical referrals in Spanish". In: *ACM Transactions on Computing for Healthcare* 3.3, pp. 1–22.

Báez, Pablo, Fabián Villena, Matías Rojas, Manuel Durán and Jocelyn Dunstan (2020b). "The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish". In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop (CliNLP 2020)* (Online, 19th Nov. 2020). Association for Computational Linguistics, pp. 291–300.

Barnes, Jeremy, Erik Velldal and Lilja Øvrelid (2021). "Improving sentiment analysis with multi-task learning of negation". In: *Natural Language Engineering* 27.2, pp. 249–269.

Bel-Enguix, Gemma, Helena Gómez-Adorno, Alejandro Pimentel, Sergio-Luis Ojeda-Trueba and Brian Aguilar-Vizuet (2021). "Negation detection on Mexican Spanish tweets: The T-MexNeg corpus". In: *Applied Sciences* 11.9, pp. 1–22.

Beltagy, Iz, Kyle Lo and Arman Cohan (2019). "SciBERT: A pretrained language model for scientific text". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)* (Hong Kong, China, 3rd–7th Nov. 2019). Association for Computational Linguistics, pp. 3615–3620.

Beltrán, Javier and Mónica González (2019). "Detection of negation cues in Spanish: The CLiC-Neg system". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 352–360.

Bengio, Yoshua, Réjean Ducharme and Pascal Vincent (2000). "A neural probabilistic language model". In: *Advances in Neural Information Processing Systems 13 (NeurIPS 2000)* (Denver, CO, USA, 1st Jan. 2000). Vol. 13. MIT Press, pp. 1–7.

Bergstra, James, Daniel Yamins and David D. Cox (2013). "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures". In: *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)* (Atlanta, GA, USA, 17th–19th June 2013). Vol. 28. JMLR.org, pp. 115–123.

Berlanga, Rafael, Victoria Nebot and Ernesto Jimenez (2010). "Semantic annotation of biomedical texts through concept retrieval". In: *Procesamiento del Lenguaje Natural* 45, pp. 247–250.

Bethard, Steven, Leon Derczynski, Guergana Savova, James Pustejovsky and Marc Verhagen (2015). "SemEval-2015 Task 6: Clinical TempEval". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (Denver, CO, USA, 4th–5th June 2015). Association for Computational Linguistics, pp. 806–814.

Bethard, Steven, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky and Marc Verhagen (2016). "SemEval-2016 Task 12: Clinical TempEval". In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)* (San Diego, CA, USA, 16th–17th June 2016). Association for Computational Linguistics, pp. 1052–1062.

Bethard, Steven, Guergana Savova, Martha Palmer and James Pustejovsky (2017). "SemEval-2017 Task 12: Clinical TempEval". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)* (Vancouver, Canada, 3rd–4th Aug. 2017). Association for Computational Linguistics, pp. 565–572.

Blanco, Alberto, Sonja Remmer, Alicia Pérez, Hercules Dalianis and Arantza Casillas (2022). "Implementation of specialised attention mechanisms: ICD-10 classification of Gastrointestinal discharge summaries in English, Spanish and Swedish". In: *Journal of Biomedical Informatics*.

Bodenreider, Olivier (2004). "The Unified Medical Language System (UMLS): Integrating biomedical terminology". In: *Nucleic Acids Research* 32.suppl_1, pp. D267–D270.

Bodenreider, Olivier and Alexa T. McCray (2003). "Exploring semantic groups through visual approaches". In: *Journal of Biomedical Informatics* 36.6, pp. 414–432.

Bodnari, Andreea, Aurélie Névéol, Özlem Uzuner, Pierre Zweigenbaum and Peter Szolovits (2013). "Multilingual named-entity recognition from parallel corpora". In: *Working Notes for CLEF 2013 Conference (CLEF 2013)* (Valencia, Spain, 23rd–26th Sept. 2013). CEUR Workshop Proceedings, pp. 1–8.

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

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor and Marcos Zampieri (2016). "Findings of the 2016 Conference on Machine Translation (WMT2016)". In: *Proceedings of the 1st*

Conference on Machine Translation (WMT 2016): Volume 2, Shared Task Papers (Berlin, Germany, 11th–12th Aug. 2016). Association for Computational Linguistics, pp. 131–198.

Brown, Peter F., Peter V. de Souza, Robert L. Mercer, Vincent J. Della Pietra and Jenifer C. Lai (1992). "Class-based n-gram models of natural language". In: *Computational Linguistics* 18.4, pp. 467–479.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei (2020). "Language Models are few-shot learners". In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* (Online, 6th–12th Dec. 2020). Curran Associates, Inc., pp. 1877–1901.

Cabitza, Federico, Raffaele Rasoini and Gian Franco Gensini (2017). "Unintended consequences of Machine Learning in medicine". In: *Journal of the American Medical Association* 318.6, pp. 516–517.

Campillos Llanos, Leonardo, Paloma Martínez and Isabel Segura-Bedmar (2017). "A preliminary analysis of negation in a Spanish clinical records dataset". In: *Actas del Taller de NEGación en ESpañol (NEGES 2017)* (Murcia, Spain, 19th Sept. 2017), pp. 33–39.

Campillos-Llanos, Leonardo, Ana Valverde-Mateos, Adrián Capllonch-Carrión and Antonio Moreno-Sandoval (2021). "A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine". In: *BMC Medical Informatics and Decision Making* 21.

Cañete, José, Gabriel Chaperon, Rodrigo Fuentes and Jorge Pérez (2020). "Spanish pre-trained BERT model and evaluation data". In: *Proceedings of the Practical ML for Developing Countries Workshop (PML4DC 2020) at the 8th International Conference on Learning Representations (ICLR 2020)* (Online, 26th Apr.–1st May 2020), pp. 1–9.

Carreras, Xavier, Isaac Chao, Lluís Padró and Muntsa Padró (2004). "Freeling: An open-source suite of language analyzers". In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)* (Lisboa, Portugal, 26th–28th May 2004). European Language Resources Association, pp. 239–242.

Carrero, Francisco, José Carlos Cortizo and José María Gómez (2008a). "Building a Spanish MMTx by using automatic translation and biomedical ontologies. Proceedings of the 9th International Conference (IDEAL 2008)". In: *Intelligent Data Engineering and Automated Learning* (Daejeon, South Korea, 2nd–5th Nov. 2008). Vol. 5326. Lecture Notes in Computer Science. Springer, pp. 346–353.

Carrero, Francisco, José Carlos Cortizo, José María Gómez and Manuel de Buenaga (2008b). "In the development of a Spanish Metamap". In: *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM 2008)* (Napa Valley, CA, USA, 26th–30th Oct. 2008). Association for Computing Machinery, pp. 1465–1466.

Carrino, Casimiro Pio, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre and Marta Villegas (2021). "Biomedical and clinical Language Models for Spanish: On the benefits of domain-specific pretraining in a mid-resource scenario". In: *arXiv*.

Casillas, Arantza, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz and Alicia Pérez (2012). "First approaches on Spanish medical record classification using Diagnostic Term to class transduction". In: *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing (FSMNLP 2012)* (Donostia-San Sebastián, Spain, 23rd–25th July 2012). Association for Computational Linguistics, pp. 60–64.

Casillas, Arantza, Alicia Pérez, Maite Oronoz, Koldo Gojenola and Sara Santiso (2016). "Learning to extract adverse drug reaction events from electronic health records in Spanish". In: *Expert Systems With Applications* 61, pp. 235–245.

Castro, Elena, Ana Iglesias, Paloma Martínez and Leonardo Castaño (2010). "Automatic identification of biomedical concepts in Spanish language unstructured clinical texts". In: *Proceedings of the 1st ACM International Health Informatics Symposium (IHI 2010)* (Arlington, VA, USA, 11th–12th Nov. 2010). Association for Computing Machinery, pp. 751–757.

Chalapathy, Raghavendra, Ehsan Zare Borzeshi and Massimo Piccardi (2016). "Bidirectional LSTM-CRF for clinical concept extraction". In: *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP 2016)* (11th–16th Dec. 2016). Osaka, Japan: The COLING 2016 Organizing Committee, pp. 7–12.

Chapman, Wendy W., Will Bridewell, Paul Hanbury, Gregory F. Cooper and Bruce G. Buchanan (2001). "A simple algorithm for identifying negated findings and diseases in discharge summaries". In: *Journal of Biomedical Informatics* 34.5, pp. 301–310.

Chazard, Emmanuel, Capucine Mouret, Grégoire Ficheur, Aurélien Schaffar, Jean-Baptiste Beuscart and Régis Beuscart (2014). "Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records". In: *International Journal of Medical Informatics* 83.4, pp. 303–312.

Chen, Long, Wenbo Fu, Yu Gu, Zhiyong Sun, Haodan Li, Enyu Li, Li Jiang, Yuan Gao and Yang Huang (2020). "Clinical concept normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking". In: *Journal of the American Medical Informatics Association* 27.10, pp. 1576–1584.

Chen, Long, Yu Gu, Xin Ji, Chao Lou, Zhiyong Sun, Haodan Li, Yuan Gao and Yang Huang (2019). "Clinical trial cohort selection based on multi-level rule-based natural language processing system". In: *Journal of the American Medical Informatics Association* 26.11, pp. 1218–1226.

Chevrier, Raphaël, Vasiliki Foufi, Christophe Gaudet-Blavignac, Arnaud Robert and Christian Lovis (2019). "Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review". In: *Journal of Medical Internet Research* 21.5, pp. 1–15.

Chiu, Billy and Simon Baker (2020). "Word embeddings for biomedical natural language processing: A survey". In: *Language and Linguistics Compass* 14.12, e12402.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio (2014). "Learning phrase representations using RNN encoder–decoder for statistical machine translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)* (Doha, Qatar, 25th–29th Oct. 2014). Association for Computational Linguistics, pp. 1724–1734.

Church, Kenneth and Mark Liberman (2021). "The future of Computational Linguistics: On beyond alchemy". In: *Frontiers in Artificial Intelligence* 4, pp. 1–18.

Clark, Kevin, Minh-Thang Luong, Quoc V. Le and Christopher D. Manning (2020). "ELECTRA: Pre-training text encoders as discriminators rather than generators". In: *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)* (Addis Ababa, Ethiopia, 30th Apr. 2020), pp. 1–18.

Clarke, Charles L. A., Maria Maistro and Mark D. Smucker (2021). "Overview of the TREC 2021 Health Misinformation track". In: *Proceedings of the thirtieth Text REtrieval Conference (TREC 2021)* (Online, 15th–29th Nov. 2021). National Institute of Standards and Technology, pp. 1–12.

Clarke, Charles L. A., Saira Rizvi, Mark D. Smucker and Guido Zuccon (2020). "Overview of the TREC 2020 Health Misinformation track". In: *Proceedings of the twenty-ninth Text REtrieval Conference (TREC 2020)* (Online, 16th–20th Nov. 2020). National Institute of Standards and Technology, pp. 1–11.

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu and Pavel Kuksa (2011). "Natural Language Processing (almost) from scratch". In: *Journal of Machine Learning Research* 12.76, pp. 2493–2537.

Colón-Ruiz, Cristóbal and Isabel Segura-Bedmar (2019). "Protected Health Information recognition by BiLSTM-CRF". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 679–686.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer and Veselin Stoyanov (2020). "Unsupervised cross-lingual representation learning at scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)* (Online, 5th–10th July 2020). Association for Computational Linguistics, pp. 8440–8451.

Connolly, Brian, Timothy Miller, Yizhao Ni, Kevin B. Cohen, Guergana Savova, Judith W. Dexheimer and John Pestian (2016). "Natural Language Processing – Overview and History". In: *Pediatric Biomedical Informatics: Computer Applications in Pediatric Research*. Ed. by John J. Hutton. Springer Singapore, pp. 203–230.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine Learning* 20.3, pp. 273–297.

Costumero, Roberto, Federico López, Consuelo Gonzalo-Martín, Marta Millan and Ernestina Menasalvas (2014). "An approach to detect negation on medical documents in Spanish. Proceedings of the International Conference (BIH 2014)". In: *Brain Informatics and Health* (Warszawa, Poland, 11th–14th Aug. 2014). Vol. 8609. Lecture Notes in Computer Science. Springer, pp. 366–375.

Cotik, Viviana, Laura Alonso Alemany, Darío Filippo, Franco Luque, Roland Roller, Jorge Vivaldi, Ammer Ayach, Fernando Carranza, Lucas Defrancesca, Antonella Dellanzo and Macarena Fernández Urquiza (2021). "Overview of CLEF eHealth Task 1 - SpRadIE: A challenge on information extraction from Spanish radiology reports". In: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum* (Online, 21st–24th Sept. 2021). CEUR Workshop Proceedings, pp. 732–750.

Cotik, Viviana, Darío Filippo and José Castaño (2015). "An approach for automatic classification of radiology reports in Spanish". In: 216, pp. 634–638.

Cotik, Viviana, Darío Filippo, Roland Roller, Hans Uszkoreit and Feiyu Xu (2017). "Creation of an annotated corpus of Spanish radiology reports". In: *arXiv*.

Cotik, Viviana, Franco Luque and Juan Manuel Pérez (2019). "Window classifiers and Conditional Random Fields for medical report de-identification".

In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 744–754.

Crespo Miguel, Mario and Antonio Frías Delgado (2008). "Aproximación a la categorización textual en español basada en la semántica de marcos". In: *Procesamiento del Lenguaje Natural* 41, pp. 65–71.

Cruz Díaz, Noa P. and Manuel J. Maña López (2015). "An analysis of biomedical tokenization: Problems and strategies". In: *Proceedings of the 6th International Workshop on Health Text Mining and Information Analysis (Louhi 2015)* (Lisboa, Portugal, 17th Sept. 2015). Association for Computational Linguistics, pp. 40–49.

— (2019). "Negation and speculation detection". Natural Language Processing 13. John Benjamins Publishing Company.

Cruz Díaz, Noa P., Roser Morante Vallejo, Manuel J. Maña López, Jacinto Mata Vázquez and Carlos L. Parra Calderón (2017). "Annotating negation in Spanish clinical texts". In: *Proceedings of the Workshop Computational Semantics Beyond Events and Roles (SemBEaR 2017)* (Valencia, Spain, 4th Apr. 2017). Association for Computational Linguistics, pp. 53–58.

Cuadros, Montse, Naiara Perez, Iker Montoya and Aitor García-Pablos (2018). "Vicomtech at BARR2: Detecting biomedical abbrebiations with ML methods and dictionary-based heuristics". In: *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)* (Sevilla, Spain, 18th Sept. 2018). CEUR Workshop Proceedings, pp. 322–328.

Dai, Manhong, Nigam H. Shah, Wei Xuan, Mark A. Musen, Stanley J. Watson, Brian D. Athey, Fan Meng et al. (2008). "An efficient solution for mapping free text to ontology terms". In: *Proceedings of the AMIA Summit on Translational Bioinformatics* (San Francisco, CA, USA, 10th–12th Mar. 2008). Vol. 21. American Medical Informatics Association.

Dalianis, Hercules and Sumithra Velupillai (2010). "De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields". In: *Journal of Biomedical Semantics* 1.1, pp. 1–10.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2019). "BERT: pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Volume 1 (Long and Short Papers)* (Minneapolis, MN, USA, 2nd–7th June 2019). Association for Computational Linguistics, pp. 4171–4186.

Díaz de Ilarraza, Arantza, Koldo Gojenola, Lourdes Araujo and Raquel Martínez (2015). "EXTracción de RElaciones entre Conceptos Médicos en fuentes de información heterogéneas (EXTRECM)". In: *Procesamiento del Lenguaje Natural* 55, pp. 157–160.

Díaz de Ilarraza, Arantza, Koldo Gojenola, Raquel Martínez, Víctor Fresno, Jordi Turmo and Lluis Padró (2017). "PROcesamiento Semántico textual Avanzado para la detección de diagnósticos, procedimientos, otros conceptos y sus relaciones en informes MEDicos (PROSA-MED)". In: *Procesamiento del Lenguaje Natural* 59, pp. 133–136.

Domínguez-Mas, Lluís, Francesco Ronzano and Laura Furlong (2019). "Supervised learning approaches to detect negation cues in Spanish reviews". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 361–368.

Doyen, Stephane and Nicholas B. Dadario (2022). "12 plagues of AI in healthcare: A practical guide to current issues with using machine learning in a medical context". In: *Frontiers in Digital Health* 4.

Eisman, Aaron S., Nishant R. Shah, Carsten Eickhoff, George Zerveas, Elizabeth S. Chen, Wen-Chih Wu and Indra Neil Sarkar (2020). "Extracting angina symptoms from clinical notes using pre-trained Transformer architectures". In: *Proceedings of the AMIA Annual Symposium* (Online, 14th–18th Nov. 2020). American Medical Informatics Association, pp. 412–421.

Elhadad, Noémie, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy W. Chapman and Guergana Savova (2015). "SemEval-2015 Task 14: Analysis of clinical text". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (Denver, CO, USA, 4th–5th June 2015). Association for Computational Linguistics, pp. 303–310.

Elhadad, Noémie, Guergana Savova, Wendy Chapman, Glenn Zaramba, David Harris, Amy Vogel, Danielle Mowery and Sumithra Velupillai (2012). "ShARe guidelines for the annotation of modifiers for disorders in clinical notes". Tech. rep. Columbia University.

Emelyanov, Yaroslav (2021). "Towards task-agnostic privacy- and utility-Preserving models". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (Online, 1st–3rd Sept. 2021). INCOMA Ltd., pp. 394–401.

Ettinger, Allyson (2020). "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models". In: *Transactions of the Association for Computational Linguistics* 8, pp. 34–48.

Fabregat, Hermenegildo, Andres Duque, Juan Martinez-Romo and Lourdes Araujo (2019a). "De-identification through Named Entity Recognition for medical document anonymization". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 663–670.

Fabregat, Hermenegildo, Andrés Duque, Juan Martínez-Romo and Lourdes Araujo (2019b). "Extending a Deep Learning approach for negation cues detection in Spanish". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 369–377.

Fabregat, Hermenegildo, Juan Martinez-Romo and Lourdes Araujo (2018a). "Deep Learning approach for negation cues detection in Spanish". In: *Proceedings of NEGES 2018: Workshop on Negation in Spanish co-located with the 34th SEPLN Conference (SEPLN 2018)* (Sevilla, Spain, 18th Sept. 2018). CEUR Workshop Proceedings, pp. 43–48.

Fabregat, Hermenegildo, Juan Martínez-Romo and Lourdes Araujo (2018b). "Overview of the DIANN task: Disability annotation task". In: *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)* (Sevilla, Spain, 18th Sept. 2018). CEUR Workshop Proceedings, pp. 1–14.

Friedman, Carol and Noémie Elhadad (2014). "Natural Language Processing in health care and biomedicine". In: *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. Ed. by Edward H. Shortliffe and James J. Cimino. Springer London, pp. 255–284.

Fu, Sunyang, Bjoerg Thorsteinsdottir, Xin Zhang, Guilherme S. Lopes, Sandeep R. Pagali, Nathan K. LeBrasseur, Andrew Wen, Hongfang Liu, Walter A. Rocca, Janet E. Olson, Jennifer St. Sauver and Sunghwan Sohn (2022). "A hybrid model to identify fall occurrence from electronic health records". In: *International Journal of Medical Informatics* 162.

Fukuda, Ken, Akemi Tamura, Tatsuhiko Tsunoda and Toshihisa Takagi (1998). "Toward information extraction: Identifying protein names from biological papers". In: *Pacific Symposium on Biocomputing (PSB 1998)* (Maui, HI, USA, 4th–9th Jan. 1998), pp. 707–718.

Gage, Philip (1994). "A new algorithm for data compression". In: *The C Users Journal* 12.2, pp. 23–38.

García-Pablos, Aitor, Naiara Perez and Montse Cuadros (2020). "Vicomtech at eHealth-KD challenge 2020: deep end-to-end model for entity and relation extraction in medical text". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society* *for Natural Language Processing (SEPLN 2020)* (Online, 23rd Sept. 2020). CEUR Workshop Proceedings, pp. 102–111.

García-Pablos, Aitor, Naiara Perez and Montse Cuadros (2021). "Vicomtech at eHealth-KD challenge 2021: deep learning approaches to model health-related text in Spanish". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021)* (Online, 21st Sept. 2021). CEUR Workshop Proceedings, pp. 712–724.

García-Sardiña, Laura (2018). "Automating the anonymisation of textual corpora". MA thesis. University of the Basque Country (UPV/EHU), pp. 1–78.

Gascó, Luis, Anastasios Nentidis, Anastasia Krithara, Darryl Estrada-Zavala, Renato Toshiyuki Murasaki, Elena Primo-Peña, Cristina Bojo Canales, Georgios Paliouras and Martin Krallinger (2021). "Overview of BioASQ 2021-MESINESP track. Evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials." In: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum* (Online, 21st–24th Sept. 2021). CEUR Workshop Proceedings, pp. 165–187.

Geer, Lewis Y., Aron Marchler-Bauer, Renata C. Geer, Lianyi Han, Jane He, Siqian He, Chunlei Liu, Wenyao Shi and Stephen H. Bryant (2010). "The NCBI biosystems database". In: *Nucleic Acids Research* 38.suppl_1, pp. D492–D496.

Gene Ontology Consortium (2004). "The Gene Ontology (GO) database and informatics resource". In: *Nucleic Acids Research* 32.suppl_1, pp. D258–D261.

Gianola, Lucie, Ēriks Ajausks, Victoria Arranz, Chomicha Bendahman, Laurent Bié, Claudia Borg, Aleix Cerdà, Khalid Choukri, Montse Cuadros, Ona de Gibert, Hans Degroote, Elena Edelman, Thierry Etchegoyhen, Ángela Franco Torres, Mercedes García Hernandez, Aitor García Pablos, Albert Gatt, Cyril Grouin, Manuel Herranz, Alejandro Adolfo Kohan, Thomas Lavergne, Maite Melero, Patrick Paroubek, Mickaël Rigault, Mike Rosner, Roberts Rozis, Lonneke van Der Plas, Rinalds Vīksna and Pierrel Zweigenbaum (2020). "Automatic removal of identifying information in official EU languages for Public Administrations: The MAPA Project". In: Frontiers in Artificial Intelligence and Applications, pp. 223–226.

Giudice, Valentino (2019). "Aspie96 at NEGES (IberLEF 2019): negation cues detection in Spanish with character-level convolutional RNN and tokenization". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 342–351.

Goeuriot, Lorraine, Gareth J. F. Jones, Liadh Kelly, Johannes Leveling, Allan Hanbury, Henning Müller, Sanna Salanterä, Hanna Suominen and Guido Zuccon (2013). "ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports". In: *Working Notes for CLEF 2013 Conference (CLEF 2013)* (Valencia, Spain, 23rd–26th Sept. 2013). CEUR Workshop Proceedings, pp. 1–16.

Gonzalez-Agirre, Aitor, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas and Martin Krallinger (2019c). "PharmaCoNER: Pharmacological substances, Compounds and proteins Named Entity Recognition track". In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks* (Hong Kong, China, 4th Nov. 2019). Association for Computational Linguistics, pp. 1–10.

Goodfellow, Ian, Yoshua Bengio and Aaron Courville (2016). "Deep learning". MIT press.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin and Tomas Mikolov (2018). "Learning Word Vectors for 157 Languages". In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki, Japan, 7th–12th May 2018). European Language Resources Association, pp. 3483–3487.

Griffiths-Jones, Sam (2004). "The microRNA registry". In: *Nucleic Acids Research* 32.Database issue, pp. D109–D111.

Gu, Yu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao and Hoifung Poon (2022). "Domain-Specific Language Model pretraining for Biomedical Natural Language Processing". In: *ACM Transactions on Computing for Healthcare* 3.1, pp. 1–23.

Gutiérrez-Fandiño, Asier, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre and Marta Villegas (2022). "MarIA: Spanish Language Models". In: *Procesamiento del Lenguaje Natural* 68, pp. 39–60.

Harris, Zellig S. (1954). "Distributional structure". In: *Word* 10.2-3, pp. 146–162.

Hartmann, Mareike and Anders Søgaard (2021). "Multilingual negation scope resolution for clinical text". In: *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis (Louhi 2021)* (Online, 19th Apr. 2021). Association for Computational Linguistics, pp. 7–18.

Hassan, Fadi, Mohammed Jabreel, Najlaa Maarrof, David Sánchez, Josep Domingo-Ferrer and Antonio Moreno (2019). "ReCRF: Spanish medical document anonymization using automatically-crafted rules and CRF". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 727–734.

Hastings, Janna, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes and Christoph Steinbeck (2016). "ChEBI in 2016: Improved services and an expanding collection of metabolites". In: *Nucleic Acids Research* 44.D1, pp. D1214–D1219.

Haveliwala, Taher H. (2002). "Topic-sensitive PageRank". In: *Proceedings of the 11th International Conference on World Wide Web (WWW 2002)* (Honolulu, HI, USA, 7th–11th May 2002). Association for Computing Machinery, pp. 517–526.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)* (Las Vegas, NV, USA, 27th–30th June 2016). IEEE, pp. 770–778.

"Health Insurance Portability and Accountability Act" (1996). In: Public Law 104-191, pp. 1–169.

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

Huang, Zhiheng, Wei Xu and Kai Yu (2015). "Bidirectional LSTM-CRF models for sequence tagging". In: *arXiv*.

Huddleston, Rodney and Geoffrey K. Pullum (2002). "The concept 'having scope over'". In: *The Cambridge grammar of the English language*. Cambridge University Press, pp. 790–792.

Hughes, Mark, Irene Li, Spyros Kotoulas and Toyotaro Suzumura (2017). "Medical text classification using Convolutional Neural Networks". In: Studies in Health Technology and Informatics, pp. 246–250.

Iglesias, Ana, Elena Castro, Rebeca Pérez, Leonardo Castaño, Paloma Martínez, José Manuel Gómez, Sandra Kohler and Ricardo Melero (2008). "MOSTAS: Un etiquetador morfo-semántico, anonimizador y corrector de historiales clínicos". In: *Procesamiento del Lenguaje Natural* 41, pp. 229–300.

Intxaurrondo, Ander, Montserrat Marimon, Aitor Gonzalez-Agirre, Jose Antonio Lopez-Martin, Heidy Rodriguez, Jesus Santamaria, Marta Villegas and Martin Krallinger (2018). "Finding mentions of abbreviations and their definitions in Spanish clinical cases: The BARR2 shared task evaluation results". In: *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)* (Sevilla, Spain, 18th Sept. 2018). CEUR Workshop Proceedings, pp. 280–289.

Intxaurrondo, Ander, Martin Pérez-Pérez, Gael Pérez-Rodríguez, Jose Antonio López-Martín, Jesus Santamaría, Santiago de la Peña, Marta Villegas, Saber Ahmad Akhondi, Alfonso Valencia, Analia Lourenço and Martin Krallinger (2017). "The Biomedical Abbreviation Recognition and Resolution (BARR) track: benchmarking, evaluation and importance of abbreviation recognition systems applied to Spanish biomedical abstracts". In: *Proceedings of the 2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017)* (Murcia, Spain, 19th Sept. 2017). CEUR Workshop Proceedings, pp. 230–246.

Jabreel, Mohammed, Fadi Hassan, Najlaa Maarrof, David Sánchez, Josep Domingo-Ferrer and Antonio Moreno (2019). "E2EJ: Anonymization of Spanish medical records using end-to-end joint Neural Networks". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 712–719.

Jagannatha, Abhyuday N. and Hong Yu (2016a). "Bidirectional RNN for Medical Event Detection in Electronic Health Records". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)* (12th–17th June 2016). San Diego, CA, USA: Association for Computational Linguistics, pp. 473–482.

— (2016b). "Structured prediction models for RNN based sequence labeling in clinical text". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)* (1st–5th Nov. 2016). Austin, TX, USA: Association for Computational Linguistics, pp. 856–865.

Jian, Zhe, Xusheng Guo, Shijian Liu, Handong Ma, Shaodian Zhang, Rui Zhang and Jianbo Lei (2017). "A cascaded approach for Chinese clinical text de-identification with less annotation effort". In: *Journal of Biomedical Informatics* 73, pp. 76–83.

Jiang, Dehuan, Yedan Shen, Shuai Chen, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Ruifeng Xu, Jun Yan and Yi Zhou (2019). "A Deep Learning-based system for the MEDDOCAN task". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 761–767.

Jiménez-Zafra, Salud María, Noa P. Cruz Díaz, Roser Morante and María Teresa Martín-Valdivia (2018a). "NEGES 2018: Workshop on negation in Spanish". In: *Procesamiento del Lenguaje Natural* 62, pp. 21–28.

Jiménez-Zafra, Salud María, Noa P. Cruz Díaz, Roser Morante and María-Teresa Martín-Valdivia (2019). "NEGES 2019 task: negation in Spanish". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing*

(SEPLN 2019) (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 329–341.

Jiménez-Zafra, Salud María, Roser Morante, Eduardo Blanco, María Teresa Martín Valdivia and L. Alfonso Ureña López (2020a). "Detecting negation cues and scopes in Spanish". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)* (Conference cancelled). European Language Resources Association, pp. 6902–6911.

Jiménez-Zafra, Salud María, Roser Morante, María Teresa Martín-Valdivia and L. Alfonso Ureña-López (2018b). "A review of Spanish corpora annotated with negation". In: *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)* (Santa Fe, NM, USA, 20th–26th Aug. 2019). Association of Computational Linguistics, pp. 915–924.

— (2020b). "Corpora annotated with negation: An overview". In: *Computational Linguistics* 46.1, pp. 1–52.

Jiménez-Zafra, Salud María, Mariona Taulé, María Teresa Martín-Valdivia, L. Alfonso Ureña-López and M. Antònia Martí (2018c). "SFU Review$_{SP}$-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns". In: *Language Resources and Evaluation* 52, pp. 533–569.

Johnson, Alistair E. W., Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi and Roger G. Mark (2016). "MIMIC-III, a freely accessible critical care database". In: *Scientific Data* 3.

Jouffroy, Jordan, Sarah F. Feldman, Ivan Lerner, Bastien Rance, Anita Burgun and Antoine Neuraz (2021). "Hybrid Deep Learning for medication-related Information Extraction from clinical texts in French: MedExt algorithm development study". In: *JMIR Medical Informatics* 9.3, pp. 1–11.

Ju, Meizhi, Huilong Duan and Haomin Li (2015). "A CRF-based method for automatic construction of chinese symptom lexicon". In: *Proceedings of the 7th International Conference on Information Technology in Medicine and Education (ITME 2015)* (Huangshan, Anhui, China, 13th–15th Nov. 2015). IEEE, pp. 5–8.

Kalyan, Katikapalli Subramanyam, Ajit Rajasekharan and Sivanesan Sangeetha (2022). "AMMU: A survey of transformer-based biomedical pretrained language models". In: *Journal of Biomedical Informatics* 126.

Kassner, Nora and Hinrich Schütze (2020). "Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)* (Online, 5th–10th July 2020). Association for Computational Linguistics, pp. 7811–7818.

Khandelwal, Aditya and Suraj Sawant (2020). "NegBERT: A transfer learning approach for negation detection and scope resolution". In: *Proceedings of the*

12th Language Resources and Evaluation Conference (LREC 2020) (Conference cancelled). European Language Resources Association, pp. 5739–5748.

Kim, Yoon (2014). "Convolutional Neural Networks for sentence classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)* (25th–29th Oct. 2014). Doha, Qatar: Association for Computational Linguistics, pp. 1746–1751.

Konstantinova, Natalia, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada and Ruslan Mitkov (2012). "A review corpus annotated for negation, speculation and their scope". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)* (Istanbul, Turkey, 21st–27th May 2012). European Language Resources Association (ELRA), pp. 3190–3195.

Kors, Jan A., Simon Clematide, Saber A. Akhondi, Erik M. van Mulligen and Dietrich Rebholz-Schuhmann (2015). "A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC". In: *Journal of the American Medical Informatics Association* 22.5, pp. 948–956.

Koza, Walter, Darío Filippo, Viviana Cotik, Vanesa Stricker, Mirian Muñoz, Ninoska Godoy, Natalia Rivas and Ricardo Martínez-Gamboa (2019). "Automatic detection of negated findings in radiological reports for Spanish language: Methodology based on lexicon-grammatical information processing". In: *Journal of Digital Imaging* 32, pp. 19–29.

Lafferty, John D., Andrew McCallum and Fernando C. N. Pereira (2001). "Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data". In: *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)* (Williamstown, MA, USA, 28th June 2001–1st July 2011). Morgan Kaufmann, pp. 282–289.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami and Chris Dyer (2016). "Neural architectures for Named Entity Recognition". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)* (12th–17th June 2016). San Diego, CA, USA: Association for Computational Linguistics, pp. 260–270.

Landis, J. Richard and Gary G. Koch (1977). "The measurement of observer agreement for categorical data". In: *Biometrics* 33.1, pp. 159–174.

Lange, Lukas, Heike Adel and Jannik Strötgen (2019). "NLNDE: The Neither-Language-Nor-Domain-Experts' way of Spanish medical document de-identification". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 671–678.

Laparra, Egoitz, Aurelie Mascio, Sumithra Velupillai and Timothy A. Miller (2021a). "A review of recent work in Transfer Learning and Domain Adaptation for Natural Language Processing of electronic health records". In: *Yearbook of Medical Informatics* 30.01, pp. 239–244.

Laparra, Egoitz, Xin Su, Yiyun Zhao, Özlem Uzuner, Timothy Miller and Steven Bethard (2021b). "SemEval-2021 Task 10: Source-free domain adaptation for semantic processing". In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval 2021)* (Online, 5th–6th Aug. 2021). Association for Computational Linguistics, pp. 348–356.

Lara-Clares, Alicia and Ana Garcia-Serrano (2019). "Key phrases annotation in medical documents: MEDDOCAN 2019 anonymization task". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 755–760.

Leaman, Robert, Ritu Khare and Zhiyong Lu (2015). "Challenges in clinical natural language processing for automated disorder normalization". In: *Journal of Biomedical Informatics* 57, pp. 28–37.

LeCun, Yann, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard and Lawrence D. Jackel (1989). "Backpropagation applied to handwritten zip code recognition". In: *Neural Computation* 1.4, pp. 541–551.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So and Jaewoo Kang (2019). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4, pp. 1234–1240.

Leevy, Joffrey L., Taghi M. Khoshgoftaar and Flavio Villanustre (2020). "Survey on RNN and CRF models for de-identification of medical free text". In: *Journal of Big Data* 7, p. 73.

Levenshtein, Vladimir I. (1966). "Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet Physics-Doklady* 10.8, pp. 707–710.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov and Luke Zettlemoyer (2020). "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)* (Online, 5th–10th July 2020). Association for Computational Linguistics, pp. 7871–7880.

Li, Qi, Stephen Andrew Spooner, Megan Kaiser, Nataline Lingren, Jessica Robbins, Todd Lingren, Huaxiu Tang, Imre Solti and Yizhao Ni (2015). "An end-to-end hybrid algorithm for automated medication discrepancy detection". In: *BMC Medical Informatics and Decision Making* 15.37, pp. 1–12.

Liaw, Richard, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez and Ion Stoica (2018). "Tune: A research platform for distributed model selection and training". In: *International Workshop on Automatic Machine Learning (AutoML 2018) collocated with the Thirty-fifth International Conference on Machine Learning (ICML 2018)* (Stockholm, Sweden, 14th July 2018), pp. 1–8.

Lima-López, Salvador, Eulàlia Farré-Maduell, Antonio Miranda-Escalada, Vicent Brivá-Iglesias and Martin Krallinger (2021a). "NLP applied to occupational health: MEDDOPROF shared task at IberLEF 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts". In: *Procesamiento del Lenguaje Natural* 67, pp. 243–256.

Lima-López, Salvador, Naiara Perez and Montse Cuadros (2021b). "Grammatical error correction for Spanish health records". In: *Procesamiento del Lenguaje Natural* 66, pp. 121–132.

Lima-López, Salvador, Naiara Perez, Montse Cuadros and German Rigau (2020a). "NUBes: A corpus of negation and uncertainty in Spanish clinical texts". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)* (Conference cancelled). European Language Resources Association, pp. 5772–5781.

Lima-López, Salvador, Naiara Perez, Laura García-Sardiña and Montse Cuadros (2020b). "HitzalMed: Anonymisation of clinical text in Spanish". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)* (Conference cancelled). European Language Resources Association, pp. 7038–7043.

Lindberg, Donald A. B., Betsy L. Humphreys and Alexa T. McCray (1993). "The Unified Medical Language System". In: *Yearbook of Medical Informatics* 2.1, pp. 41–51.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi and Graham Neubig (2021). "Pre-train, prompt, and predict: A systematic survey of prompting methods in Natural Language Processing". In: *arXiv*.

Liu, Xiaodong, Pengcheng He, Weizhu Chen and Jianfeng Gao (2019). "Multi-task deep neural networks for natural language understanding". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)* (Firenze, Italy, 2nd–7th June 2019). Association for Computational Linguistics, pp. 4487–4496.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov (2019). "RoBERTa: A robustly optimized BERT pretraining approach". In: *arXiv*.

Loharja, Henry, Lluís Padró and Jordi Turmo (2018). "Negation cues detection using CRF on Spanish product review texts". In: *Proceedings of NEGES 2018: Workshop on Negation in Spanish co-located with the 34th SEPLN Conference (SEPLN 2018)* (Sevilla, Spain, 18th Sept. 2018). CEUR Workshop Proceedings, pp. 49–54.

Lopes, Fábio, César Teixeira and Hugo Gonçalo Oliveira (2019). "Named Entity Recognition in Portuguese neurology text using CRF. Proceedings of the 19th EPIA Conference on Artificial Intelligence (EPIA 2019)". In: *Progress in Artificial Intelligence* (Vila Real, Portugal, 3rd–6th Sept. 2019). Vol. 11804. Lecture Notes in Computer Science. Springer, pp. 336–348.

López-García, Guillermo, José M. Jerez, Nuria Ribelles, Emilio Alba and Francisco J. Veredas (2021). "Transformers for clinical coding in Spanish". In: *IEEE Access* 9, pp. 72387–72397.

López-Úbeda, Pilar, Manuel Díaz-Galiano, Luis Alfonso Ureña-López and María-Teresa Martín-Valdivia (2019). "Anonymization of clinical reports in Spanish: a hybrid method based on Machine Learning and rules". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 687–695.

López-Úbeda, Pilar, Alexandra Pomares-Quimbaya, Manuel Carlos Díaz-Galiano and Stefan Schulz (2021). "Collecting specialty-related medical terms: Development and evaluation of a resource for Spanish". In: *BMC Medical Informatics and Decision Making* 21.145, pp. 1–17.

Lu, Qiuhao, Thien Huu Nguyen and Dejing Dou (2021). "Predicting patient readmission risk from medical text via knowledge graph enhanced multiview graph convolution". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)* (11th–15th July 2021). Online: Association for Computing Machinery, pp. 1990–1994.

Luo, Yuan, Yu Cheng, Özlem Uzuner, Peter Szolovits and Justin Starren (2017). "Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes". In: *Journal of the American Medical Informatics Association* 25.1, pp. 93–98.

MacRae, Jayden, Tom Love, Michael G. Baker, Anthony Dowell, Matthew Carnachan, Maria Stubbe and Lynn McBain (2015). "Identifying influenza-like illness presentation from unstructured general practice clinical narrative using a text classifier rule-based expert system versus a clinical expert". In: *BMC Medical Informatics and Decision Making* 15.78, pp. 1–11.

Magnini, Bernardo, Begoña Altuna, Alberto Lavelli, Manuela Speranza and Roberto Zanoli (2021a). "The E3C Project: Collection and annotation of a multilingual corpus of clinical cases". In: *Proceedings of the Seventh Italian Conference on Computational Linguistics 2020 (CLiC-it 2020)* (Bologna, Italy, 1st–3rd Mar. 2021). CEUR Workshop Proceedings, pp. 1–7.

— (2021b). "The E3C Project: European Clinical Case Corpus". In: *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations (SEPLN-PD 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021)* (Málaga, Spain, 21st Sept. 2021). CEUR Workshop Proceedings, pp. 17–20.

Mamede, Nuno, Jorge Baptista and Francisco Dias (2016). "Automated anonymization of text documents". In: *Proceedings of the 2016 IEEE Congress on Evolutionary Computation (CEC 2016)* (Vancouver, Canada, 21st–29th July 2016). IEEE, pp. 1287–1294.

Mao, Jihang and Wanli Liu (2019). "Hadoken: a BERT-CRF model for medical document anonymization". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 720–726.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz and Britta Schasberger (1994). "The Penn Treebank: Annotating predicate argument structure". In: *Proceedings of the Workshop on Human Language Technology* (Plainsboro, NJ, USA, 8th–11th Mar. 1994). Association for Computational Linguistics.

Marimon, Montserrat, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodríguez, Jose Antonio Lopez Martin, Marta Villegas and Martin Krallinger (2019). "Automatic de-identification of medical texts in Spanish: The MEDDOCAN track, corpus, guidelines, methods and evaluation of results". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 618–638.

Marimon, Montserrat, Jorge Vivaldi and Núria Bel (2017a). "Annotation of negation in the IULA Spanish Clinical Record Corpus". In: *Proceedings of the Workshop Computational Semantics Beyond Events and Roles (SemBEaR 2017)* (Valencia, Spain, 4th Apr. 2017). Association for Computational Linguistics, pp. 43–52.

Martí, M. Antònia, María Teresa Martín-Valdivia, Mariona Taulé, Salud María Jiménez-Zafra, Montserrat Nofre and Laia Marsó (2016). "La negación en español: análisis y tipología de patrones de negación". In: *Procesamiento del Lenguaje Natural* 57, pp. 41–48.

Martí, M. Antònia and Mariona Taulé (2018). "Análisis comparativo de los sistemas de anotación de la negación en español". In: *Proceedings of NEGES 2018: Workshop on Negation in Spanish co-located with the 34th SEPLN Conference (SEPLN 2018)* (Sevilla, Spain, 18th Sept. 2018). CEUR Workshop Proceedings, pp. 23–28.

Martínez Cámara, Eugenio, Yudivian Almeida-Cruz, Manuel Carlos Díaz Galiano, Suilan Estévez-Velarde, Miguel Ángel García Cumbreras, Manuel García Vega, Yoan Gutiérrez, Arturo Montejo Ráez, Andres Montoyo, Rafael Muñoz, Alejandro Piad-Morffis and Julio Villena Román (2018). "Overview of TASS 2018: Opinions, health and emotions". In: *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018) co-located with 34nd SEPLN Conference (SEPLN 2018)* (Sevilla, Spain, 18th Sept. 2018). CEUR Workshop Proceedings, pp. 13–27.

McCray, Alexa T., Anita Burgun and Olivier Bodenreider (2001). "Aggregating UMLS semantic types for reducing conceptual complexity". In: Studies in Health Technology and Informatics 84.Pt 1, pp. 216–220.

McCray, Alexa T. and Stuart J. Nelson (1995). "The representation of meaning in the UMLS". In: *Methods of Information in Medicine* 34.01/02, pp. 193–201.

Medina, Salvador and Jordi Turmo (2018). "Building a Spanish/Catalan health records corpus with very sparse protected information labelled". In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki, Japan, 7th–12th May 2018). European Language Resources Association, pp. 7–12.

Menger, Vincent, Floor Scheepers, Lisette Maria van Wijk and Marco Spruit (2018). "DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text". In: *Telematics and Informatics* 35.4, pp. 727–736.

Miftahutdinov, Zulfat, Ilseyar Alimova and Elena Tutubalina (2019). "KFU NLP team at SMM4H 2019 tasks: Want to extract adverse drugs reactions from tweets? BERT to the rescue". In: *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H 2019) Workshop & Shared Task* (2nd Aug. 2019). Florence, Italy: Association for Computational Linguistics, pp. 52–57.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean (2013a). "Distributed representations of words and phrases and their compositionality". In: *Advances in Neural Information Processing Systems 26 (NIPS 2013)* (Harrah's Lake Tahoe, NV, USA, 5th–10th Dec. 2013). Curran Associates Inc., pp. 3111–3119.

Mikolov, Tomas, Wen-tau Yih and Geoffrey Zweig (2013b). "Linguistic regularities in continuous space word representations". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computa-*

tional Linguistics: Human Language Technologies (NAACL HLT 2013) (Atlanta, GA, USA, 9th–14th June 2013). Association for Computational Linguistics, pp. 746–751.

Ministerio de Sanidad (2021). "Sistema HCDSNS: Historia Clínica Digital del Sistema Nacional de Salud. Informe de Situación 31 de enero de 2021". Tech. rep.

Miranda-Escalada, Antonio, Eulàlia Farré-Maduell, Salvador Lima-López, Luis Gascó, Vicent Briva-Iglesias, Marvin Agüero-Torales and Martin Krallinger (2021). "The ProfNER shared task on automatic recognition of occupation mentions in social media: Systems, evaluation, guidelines, embeddings and corpora". In: *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task* (Mexico City, Mexico, 10th June 2021). Association for Computational Linguistics, pp. 13–20.

Miranda-Escalada, Antonio, Eulàlia Farréa and Martin Krallingera (2020a). "Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist track for cancer text mining in Spanish, corpus, guidelines, methods and results". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020)* (Online, 23rd Sept. 2020). CEUR Workshop Proceedings, pp. 303–323.

Miranda-Escalada, Antonio, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé and Martin Krallinger (2020b). "Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of CLEF eHealth 2020." In: *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum* (Thessaloníki, Greece, 22nd–25th Sept. 2020). CEUR Workshop Proceedings, pp. 1–29.

Montoya, Iker (2017). "Análisis, normalización, enriquecimiento y codificación de historia clínica electrónica (HCE)". MA thesis. University of the Basque Country (UPV/EHU), pp. 1–113.

Morante, Roser and Eduardo Blanco (2012a). "*SEM 2012 shared task: Resolving the scope and focus of negation". In: **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (Montréal, Canada, 7th–8th June 2012). Association for Computational Linguistics, pp. 265–274.

— (2021). "Recent advances in processing negation". In: *Natural Language Engineering* 27.2, pp. 121–130.

Morante, Roser and Walter Daelemans (2012b). "ConanDoyle-neg: Annotation of negation cues and their scope in Conan Doyle stories". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC*

2012) (İstanbul, Turkey, 21st–27th May 2012). European Language Resources Association, pp. 1563–1568.

Morante, Roser and Caroline Sporleder (2012c). "Modality and negation: An introduction to the special issue". In: *Computational Linguistics* 38.2, pp. 223–260.

Moreno, Antonio, Susana López, Fernando Sánchez and Ralph Grishman (2003). "Developing a syntactic annotation scheme and tools for a Spanish treebank". In: *Treebanks*. Ed. by Anne Abeillé. Vol. 20. Text, Speech and Language Technology. Springer, pp. 149–163.

Moreno Sandoval, Antonio and Marta Salazar Garrote (2013). "La anotación de la negación en un corpus escrito etiquetado sintácticamente". In: *Revista Iberoamericana de Lingüística* 8, pp. 45–60.

Mowery, Danielle L., Brett R. Southe, Lee Christensen, Laura-Maria Murtola, Sanna Salanterä, Hanna Suominen, David Martinez, Noemie Elhadad, Sameer Pradhan, Guergana Savova and Wendy W. Chapman (2013). "Task 2: ShARe/CLEF eHealth Evaluation Lab 2013". In: *Working Notes for CLEF 2013 Conference (CLEF 2013)* (Valencia, Spain, 23rd–26th Sept. 2013). CEUR Workshop Proceedings, pp. 1–11.

Nakamura, Yuta, Shouhei Hanaoka, Yukihiro Nomura, Takahiro Nakao, Soichiro Miki, Takeyuki Watadani, Takeharu Yoshikawa, Naoto Hayashi and Osamu Abe (2021). "Automatic detection of actionable radiology reports using Bidirectional Encoder Representations from Transformers". In: *BMC Medical Informatics and Decision Making* 21.1, pp. 1–19.

Naseem, Usman, Adam G. Dunn, Matloob Khushi and Jinman Kim (2022). "Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT". In: *BMC Bioinformatics* 23, pp. 1–15.

Névéol, Aurélie, Hercules Dalianis, Sumithra Velupillai, Guergana Savova and Pierre Zweigenbaum (2018a). "Clinical Natural Language Processing in languages other than English: Opportunities and challenges". In: *Journal of Biomedical Semantics* 9, pp. 1–13.

Névéol, Aurélie, Aude Robert, Francesco Grippo, Claire Morgand, Chiara Orsi, László Pleikán, Lionel Ramadier, Grégoire Rey and Pierre Zweigenbaum (2018b). "CLEF eHealth 2018 Multilingual Information Extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italian". In: *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum* (Avignon, France, 10th–14th Sept. 2018). CEUR Workshop Proceedings, pp. 1–18.

Neves, Mariana, Daniel Butzke, Antje Dörendahl, Nora Leich, Benedkit Hummel, Gilbert Schönfelder and Barbara Grune (2019). "Overview of the CLEF eHealth 2019 Multilingual Information Extraction". In: *Working Notes of*

CLEF 2019 - Conference and Labs of the Evaluation Forum (Lugano, Switzerland, 9th–12th Sept. 2019). CEUR Workshop Proceedings, pp. 1–9.

Nunes, Tiago, David Campos, Sérgio Matos and José Luís Oliveira (2013). "BeCAS: Biomedical concept recognition services and visualization". In: *Bioinformatics* 29.15, pp. 1915–1916.

Oronoz, Maite, Arantza Casillas, Koldo Gojenola and Alicia Pérez (2013). "Automatic annotation of medical records in Spanish with disease, drug and substance names. Proceedings of the 18th Iberoamerican Congress on Pattern Recognition (CIARP 2013)". In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (Havana, Cuba, 20th–23rd Nov. 2013). Vol. 8259. Lecture Notes in Computer Science. Springer, pp. 536–543.

Oronoz, Maite, Koldo Gojenola, Alicia Pérez, Arantza Díaz De Ilarraza and Arantza Casillas (2015). "On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions". In: *Journal of Biomedical Informatics* 56, pp. 318–332.

Otegi, Arantxa, Aitor Agirre, Jon Ander Campos, Aitor Soroa and Eneko Agirre (2020). "Conversational Question Answering in low resource scenarios: A dataset and case study for Basque". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)* (Conference cancelled). European Language Resources Association, pp. 436–442.

Pan, Sinno Jialin and Qiang Yang (2010). "A survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10, pp. 1345–1359.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot and Édouard Duchesnay (2011). "Scikit-learn: machine learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830.

Pennington, Jeffrey, Richard Socher and Christopher Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)* (Doha, Qatar, 25th–29th Oct. 2014). Association for Computational Linguistics, pp. 1532–1543.

Percha, Bethany (2021). "Modern clinical text mining: A guide and review". In: *Annual Review of Biomedical Data Science* 4.1, pp. 165–187.

Perez, Naiara, Pablo Accuosto, Àlex Bravo, Montse Cuadros, Eva Martínez-Garcia, Horacio Saggion and German Rigau (2020). "Cross-lingual semantic annotation of biomedical literature: Experiments in Spanish and English". In: *Bioinformatics* 36.6, pp. 1872–1880.

Perez, Naiara, Montse Cuadros and German Rigau (2018). "Biomedical term normalization of EHRs with UMLS". In: *Proceedings of the 11th Inter-*

national Conference on Language Resources and Evaluation (LREC 2018) (Miyazaki, Japan, 7th–12th May 2018). European Language Resources Association, pp. 2045–2051.

Pérez, Alicia, Arantza Casillas, Koldo Gojenola, Maite Oronoz, Nerea Aguirre and Estibaliz Amillano (2014). "The aid of machine learning to overcome the classification of real health discharge reports written in Spanish". In: *Procesamiento del Lenguaje Natural* 53, pp. 77–84.

Perez de Viñaspre Garralda, Olatz (2017). "Osasun-alorreko termino-sorkuntza automatikoa: SNOMED CTren eduki terminologikoaren euskaratzea". PhD thesis. University of the Basque Country (UPV/EHU), pp. 1–156.

Pérez-Díez, Irene, Raúl Pérez-Moraga, Adolfo López-Cerdán, Jose-Maria Salinas-Serrano and María de la Iglesia-Vayá (2021). "De-identifying Spanish medical texts - named entity recognition applied to radiology reports". In: *Journal of Biomedical Semantics* 12, pp. 1–13.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer (2018). "Deep contextualized word representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018), Volume 1 (Long Papers)* (New Orleans, LA, USA, 1st–6th June 2018). Association for Computational Linguistics, pp. 2227–2237.

Piad-Morffis, Alejandro, Suilan Estevez-Velarde, Yoan Gutierrez, Yudivian Almeida-Cruz, Andrés Montoyo and Rafael Muñoz (2021). "Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2021". In: *Procesamiento del Lenguaje Natural* 67, pp. 233–242.

Piad-Morffis, Alejandro, Yoan Gutiérrez, Hian Cañizares-Diaz, Suilan Estevez-Velarde, Rafael Muñoz, Andrés Montoyo and Yudivián Almeida-Cruz (2020). "Overview eHealth-KD 2020". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020)* (Online, 23rd Sept. 2020). CEUR Workshop Proceedings, pp. 71–84.

Piad-Morffis, Alejandro, Yoan Gutiérrez, Juan Pablo Consuegra-Ayala, Suilan Estevez-Velarde, Yudivián Almeida-Cruz, Rafael Muñoz and Andrés Montoyo (2019). "Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2019". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 1–16.

Piccialli, Francesco, Vittorio di Somma, Fabio Giampaolo, Salvatore Cuomo and Giancarlo Fortino (2021). "A survey on deep learning in medicine: Why, how and when?" In: *Information Fusion* 66, pp. 111–137.

Porta-Zamorano, Jordi (2019). "Spanish medical document anonymization with three-channel Convolutional Neural Networks". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 639–646.

Pradhan, Sameer, Noemie Elhadad, Brett R. South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman and Guergana Savova (2013). "Task 1: ShARe/CLEF eHealth Evaluation Lab 2013". In: *Working Notes for CLEF 2013 Conference (CLEF 2013)* (Valencia, Spain, 23rd–26th Sept. 2013). CEUR Workshop Proceedings, pp. 1–6.

Pradhan, Sameer, Noémie Elhadad, Wendy W. Chapman, Suresh Manandhar and Guergana Savova (2014). "SemEval-2014 Task 7: Analysis of clinical text". In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (Dublin, Ireland, 23rd–24th Aug. 2014). Association for Computational Linguistics, pp. 54–62.

Prakash G., Bino Patric, Shomona Gracia Jacob and Radhameena S. (2014). "Mining semantic representation from medical text: A Bayesian approach". In: *Proceedings of the 2014 International Conference on Recent Trends in Information Technology (ICRTIT 2014)* (Chennai, India, 10th–12th Apr. 2014). IEEE, pp. 1–4.

Pratt, A. W. (1973). "Medicine, Computers, and Linguistics". In: *Advances in Biomedical Engineering*. Ed. by J. H. U. Brown and James F. Dickson. Academic Press, pp. 97–140.

Price W. Nicholson, II, Sara Gerke and I. Glenn Cohen (2019). "Potential liability for physicians using Artificial Intelligence". In: *Journal of the American Medical Association* 322.18, pp. 1765–1766.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li and Peter J. Liu (2020). "Exploring the limits of transfer learning with a unified text-to-text Transformer". In: *Journal of Machine Learning Research* 21.140, pp. 1–67.

Ras, Gabriëlle, Ning Xie, Marcel van Gerven and Derek Doran (2022). "Explainable Deep Learning: A field guide for the uninitiated". In: *Journal of Artificial Intelligence Research* 73, pp. 329–296.

Rasmy, Laila, Yonghui Wu, Ningtao Wang, Xin Geng, W. Jim Zheng, Fei Wang, Hulin Wu, Hua Xu and Degui Zhi (2018). "A study of generalizability of Recurrent Neural Network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set". In: *Journal of Biomedical Informatics* 84, pp. 11–16.

Ravì, Daniele, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo and Guang-Zhong Yang (2017). "Deep Learning for

Health Informatics". In: *IEEE Journal of Biomedical and Health Informatics* 21.1, pp. 4–21.

Rebholz-Schuhmann, Dietrich, Simon Clematide, Fabio Rinaldi, Senay Kafkas, Erik M. van Mulligen, Chinh Bui, Johannes Hellrich, Ian Lewin, David Milward, Michael Poprat, Antonio Jimeno-Yepes, Udo Hahn and Jan A. Kors (2013a). "Entity recognition in parallel multi-lingual biomedical corpora: The CLEF-ER laboratory overview. Proceedings of the 4th International Conference of the CLEF Initiative (CLEF 2013)". In: *Information Access Evaluation. Multilinguality, Multimodality, and Visualization* (Valencia, Spain, 23rd–26th Sept. 2013). Vol. 8138. Lecture Notes in Computer Science. Springer, pp. 353–367.

Rebholz-Schuhmann, Dietrich, Jee-Hyub Kim, Ying Yan, Abhishek Dixit, Caroline Friteyre, Robert Hoehndorf, Rolf Backofen and Ian Lewin (2013b). "Evaluation and cross-comparison of lexical entities of biological interest (LexEBI)". In: *PloS One* 8.10, e75185.

"Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)" (2016). In: *Official Journal of the European Union* L 119, pp. 1–88.

Reimers, Nils and Iryna Gurevych (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)* (Hong Kong, China, 3rd–7th Nov. 2019). Association for Computational Linguistics, pp. 3982–3992.

Rivera Zavala, Renzo and Paloma Martínez (2020). "The impact of pretrained language models on negation and speculation detection in cross-lingual medical text: Comparative study". In: *JMIR Medical Informatics* 8.12, e18953.

Roberts, Kirk, Dina Demner-Fushman, Ellen M. Voorhees and William Hersh (2016). "Overview of the TREC 2016 Clinical Decision Support track". In: *Proceedings of the twenty-fifth Text REtrieval Conference (TREC 2016)* (Gaithersburg, MD, USA, 15th–18th Nov. 2016). National Institute of Standards and Technology, pp. 1–14.

Roberts, Kirk, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick and Alexander J. Lazar (2018). "Overview of the TREC 2018 Precision Medicine track". In: *Proceedings of the twenty-seventh Text REtrieval Conference (TREC 2018)* (Gaithersburg, MD, USA, 14th–16th Nov. 2018). National Institute of Standards and Technology, pp. 1–12.

— (2020). "Overview of the TREC 2020 Precision Medicine track". In: *Proceedings of the twenty-ninth Text REtrieval Conference (TREC 2020)* (Online, 16th–20th Nov. 2020). National Institute of Standards and Technology, pp. 1–10.

Roberts, Kirk, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, Alexander J. Lazar and Shubham Pant (2017). "Overview of the TREC 2017 Precision Medicine track". In: *Proceedings of the twenty-sixth Text REtrieval Conference (TREC 2017)* (Gaithersburg, MD, USA, 15th–17th Nov. 2017). National Institute of Standards and Technology, pp. 1–13.

Roberts, Kirk, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, Alexander J. Lazar, Shubham Pant and Funda Meric-Bernstam (2019). "Overview of the TREC 2019 Precision Medicine track". In: *Proceedings of the twenty-eighth Text REtrieval Conference (TREC 2019)* (Gaithersburg, MD, USA, 13th–15th Nov. 2019). National Institute of Standards and Technology, pp. 1–12.

Roberts, Kirk, Matthew S. Simpson, Ellen M. Voorhees and William Hersh (2015). "Overview of the TREC 2015 Clinical Decision Support track". In: *Proceedings of the twenty-fourth Text REtrieval Conference (TREC 2015)* (Gaithersburg, MD, USA, 17th–20th Nov. 2015). National Institute of Standards and Technology, pp. 1–12.

Rodriguez-Penagos, Carlos, Anastasios Nentidis, Aitor Gonzalez-Agirre, Alejandro Asensio, Jordi Armengol-Estapé, Anastasia Krithara, Marta Villegas, Georgios Paliouras and Martin Krallinger (2020). "Overview of MESINESP, a Spanish medical semantic indexing task within BioASQ 2020". In: *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum* (Thessaloníki, Greece, 22nd–25th Sept. 2020). CEUR Workshop Proceedings, pp. 1–12.

Rojas, Matías, Jocelyn Dunstan and Fabián Villena (2022). "Clinical Flair: A pre-trained Language Model for Spanish clinical Natural Language Processing". In: *Proceedings of the 4th Clinical Natural Language Processing Workshop (ClinicalNLP 2022)* (Seattle, WA, USA, 14th July 2022). Association for Computational Linguistics, pp. 87–92.

Roller, Roland, Madeleine Kittner, Dirk Weissenborn and Ulf Leser (2018). "Cross-lingual candidate search for biomedical concept normalization". In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki, Japan, 7th–12th May 2018). European Language Resources Association, pp. 16–20.

de la Rosa, Javier, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas and María Grandury (2022). "BERTIN: Efficient pre-training of a Spanish Language Model using perplexity sampling". In: *Procesamiento del Lenguaje Natural* 68, pp. 13–23.

Rumelhart, David E., Geoffrey E. Hinton and Ronald J. Williams (1986). "Learning representations by back-propagating errors". In: *Nature* 323, pp. 533–536.

Safran, Charles, Meryl Bloomrosen, W. Edward Hammond, Steven Labkoff, Suzanne Markel-Fox, Paul C. Tang and Don E. Detmer (2007). "Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper". In: *Journal of the American Medical Informatics Association* 14.1, pp. 1–9.

Sager, Naomi (1972). "Syntactic formatting of science information". In: *Proceedings of the Fall Joint Computer Conference (AFIPS '72 Fall, Part II)* (Anaheim, CA, USA, 5th–7th Dec. 1972). Association for Computing Machinery, pp. 791–800.

— (1978). "Natural language information formatting: The automatic conversion of texts to a structured data base". In: ed. by Marshall C. Yovits. Vol. 17. Advances in Computers. Elsevier, pp. 89–162.

— (1980). "Computational Linguistics In Medicine". In: *Americal Journal of Computational Linguistics* 6.1, pp. 44–47.

Sánchez-León, Fernando (2019). "Resource-based anonymization for Spanish clinical cases". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 704–711.

Santiso, Sara, Arantza Casillas, Alicia Pérez and Maite Oronoz (2017). "Medical entity recognition and negation extraction: Assessment of NegEx on health records in Spanish. Proceedings of the 5th International Work-Conference (IWBBIO 2017)". In: *Bioinformatics and Biomedical Engineering* (Granada, Spain, 26th–28th Apr. 2017). Vol. 10208. Lecture Notes in Computer Science. Springer, pp. 177–188.

— (2019). "Word embeddings for negation detection in health records written in Spanish". In: *Soft Computing* 23, pp. 10969–10975.

Santiso, Sara, Alicia Pérez and Arantza Casillas (2021). "Adverse Drug Reaction extraction: Tolerance to entity recognition errors and sub-domain variants". In: *Computer Methods and Programs in Biomedicine* 199.

Santiso, Sara, Alicia Pérez, Arantza Casillas and Maite Oronoz (2020). "Neural negated entity recognition in Spanish electronic health records". In: *Journal of Biomedical Informatics* 105, p. 103419.

dos Santos, Cícero and Maíra Gatti (2014). "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts". In: *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014): Technical Papers* (23rd–29th Aug. 2014). Dublin, Ireland: Dublin City University and Association for Computational Linguistics, pp. 69–78.

Savova, Guergana K., James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler and Christopher G. Chute (2010). "Mayo clinical Text Analysis and Knowledge Extraction system (cTAKES): architecture, component evaluation and applications”. In: *Journal of the American Medical Informatics Association* 17.5, pp. 507–513.

Schneider, Werner and Anna Lena Sågvall Hein, eds. (1977). “Computational Linguistics in Medicine”. North-Holland Publishing Co., pp. 1–181.

Schuster, Mike and Kaisuke Nakajima (2012). “Japanese and Korean voice search”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)* (Kyoto, Japan, 25th–30th Mar. 2012). IEEE, pp. 5149–5152.

Schuster, Mike and Kuldip K. Paliwal (1997). “Bidirectional recurrent neural networks”. In: *IEEE Transactions on Signal Processing* 45.11, pp. 2673–2681.

Sennrich, Rico, Barry Haddow and Alexandra Birch (2016). “Neural Machine Translation of rare words with subword units”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016) – Volume 1: Long Papers* (Berlin, Germany, 7th–12th Aug. 2016). Association for Computational Linguistics, pp. 1715–1725.

Serrà, Joan and Alexandros Karatzoglou (2017). “Getting Deep Recommenders Fit: Bloom Embeddings for Sparse Binary Input/Output Networks”. In: *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys 2017)* (Como, Italy, 27th–31st Aug. 2017). Association for Computing Machinery, pp. 279–287.

Seuss, Hannes, Peter Dankerl, Matthias Ihle, Andrea Grandjean, Rebecca Hammon, Nicola Kaestle, Peter A. Fasching, Christian Maier, Jan Christoph, Martin Sedlmayr, Michael Uder, Alexander Cavallaro and Matthias Hammon (2017). “Semi-automated de-identification of German content sensitive reports for big data analytics”. In: *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren* 189.7, pp. 661–671.

Shaitarova, Anastassia, Lenz Furrer and Fabio Rinaldi (2020). “Cross-lingual transfer-learning approach to negation scope resolution”. In: *Proceedings of the 5th Swiss Text Analytics Conference (SwissText 2020) & 16th Conference on Natural Language Processing (KONVENS 2020)* (Online, 23rd–25th June 2020). CEUR Workshop Proceedings, pp. 1–7.

Shaitarova, Anastassia and Fabio Rinaldi (2021). “Negation typology and general representation models for cross-lingual zero-shot negation scope resolution in Russian, French, and Spanish”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021), Student Research Workshop* (Online, 6th–11th June 2021). Association for Computational Linguistics, pp. 15–23.

Simpson, Matthew S., Ellen M. Voorhees and William Hersh (2014). “Overview of the TREC 2014 Clinical Decision Support track”. In: *Proceedings of the twenty-third Text REtrieval Conference (TREC 2014)* (Gaithersburg, MD, USA, 19th–21st Nov. 2014). National Institute of Standards and Technology, pp. 1–8.

Sineva, Elizaveta, Stefan Grünewald, Annemarie Friedrich and Jonas Kuhn (2021). "Negation-Instance Based Evaluation of End-to-End Negation Resolution". In: *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL 2021)* (Online, 10th–11th Nov. 2021). Association for Computational Linguistics, pp. 528–543.

Soares, Felipe, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger and Jordi Armengol-Estapé (2019b). "Medical word embeddings for Spanish: Development and evaluation". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP 2019)* (Minneapolis, MN, USA, 7th June 2019). Association for Computational Linguistics, pp. 124–133.

Sohrab, Mohammad Golam, Pham Minh Thang and Makoto Miwa (2019). "A generic neural exhaustive approach for entity recognition and sensitive span detection". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 735–743.

Solarte Pabón, Oswaldo, Orlando Montenegro, Maria Torrente, Alejandro Rodríguez González, Mariano Provencio and Ernestina Menasalvas (2022). "Negation and uncertainty detection in clinical texts written in Spanish: a deep learning-based approach". In: *PeerJ Computer Science* 8, e913.

Solarte-Pabón, Oswaldo, Ernestina Menasalvas and Alejandro Rodriguez-González (2020). "Spa-neg: An approach for negation detection in clinical text written in Spanish. Proceedings of the 8th International Work-Conference (IWBBIO 2020)". In: *Bioinformatics and Biomedical Engineering* (Granada, Spain, 30th Sept.–2nd Oct. 2020). Vol. 12108. Lecture Notes in Computer Science. Springer, pp. 323–337.

Soysal, Ering, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu and Hua Xu (2017). "CLAMP – A toolkit for efficiently building customized clinical natural language processing pipelines". In: *Journal of the American Medical Informatics Association* 25.3, pp. 331–336.

Spasic, Irene and Goran Nenadic (2020). "Clinical text data in Machine Learning: Systematic review". In: *JMIR Medican Informatics* 8.3, pp. 1–19.

Spasić, Irena, Pete Burnap, Mark Greenwood and Michael Arribas-Ayllon (2012). "A Naïve Bayes approach to classifying topics in suicide notes". In: *Biomedical Informatics Insights* 5.Suppl. 1, pp. 87–97.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii (2012). "brat: a web-based tool for NLP-assisted text annotation". In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguis-*

tics (EACL 2012) (Avignon, France, 23rd–27th Apr. 2012). Association for Computational Linguistics, pp. 102–107.

Stricker, Vanesa, Ignacio Iacobacci and Viviana Cotik (2015). "Negated findings detection in radiology reports in Spanish: an adaptation of NegEx to Spanish". In: *Workshop on Replicability and Reproducibility in Natural Language Processing: adaptative methods, resources and software (RRANLP 2015) co-located with the 24th International Joint Conference on Articial Intelligence (IJCAI 2015)* (Buenos Aires, Argentina, 26th July 2015), pp. 1–7.

Stubbs, Amber, Michele Filannino, Ergin Soysal, Samuel Henry and Özlem Uzuner (2019). "Cohort selection for clinical trials: n2c2 2018 shared task track 1". In: *Journal of the American Medical Informatics Association* 26.11, pp. 1163–1171.

Stubbs, Amber, Christopher Kotfila and Özlem Uzuner (2015). "Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1". In: *Journal of Biomedical Informatics* 58.Supplement, S11–S19.

Styler IV, William F., Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova and James Pustejovsky (2014). "Temporal annotation in the clinical domain". In: *Transactions of the Association for Computational Linguistics* 2, pp. 143–154.

Suárez-Paniagua, Víctor (2019). "VSP at MEDDOCAN 2019 de-identification of medical documents in Spanish with Recurrent Neural Networks". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (Bilbao, Spain, 24th Sept. 2019). CEUR Workshop Proceedings, pp. 654–662.

Suárez-Paniagua, Víctor, Hong Dong and Arlene Casey (2021). "A multi-BERT hybrid system for Named Entity Recognition in Spanish radiology reports". In: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum* (Online, 21st–24th Sept. 2021). CEUR Workshop Proceedings, pp. 846–856.

Sun, Tian-Xiang, Xiang-Yang Liu, Xi-Peng Qiu and Xuan-Jing Huang (2022). "Paradigm shift in Natural Language Processing". In: *Machine Intelligence Research* 19, pp. 169–183.

Szarvas, György, Veronika Vincze, Richárd Farkas and János Csirik (2008). "The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts". In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (Columbus, OH, USA, 19th June 2008). Association for Computational Linguistics, pp. 38–45.

Taboada, Maite, Caroline Anthony and Kimberly Voll (2006). "Methods for creating semantic orientation dictionaries". In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (Genova, Italy, 22nd–28th May 2006). European Language Resources Association (ELRA).

Tang, Buzhou, Hongxin Cao, Yonghui Wu, Min Jiang and Hua Xu (2012). "Clinical entity recognition using Structural Support Vector Machines with rich features". In: *Proceedings of the ACM Sixth International Workshop on Data and Text Mining in Biomedical Informatics (DTMBIO 2012)* (Maui, HI, USA, 29th Oct. 2012). Association for Computing Machinery, pp. 13–20.

Taulé, Mariona, M. Antònia Martí and Marta Recasens (2008). "AnCora: Multilevel annotated corpora for Catalan and Spanish". In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)* (Marrakech, Morocco, 18th–30th May 2008). European Language Resources Association.

Taulé, Mariona, Montserrat Nofre, Mónica González and Maria Antònia Martí (2021). "Focus of negation: Its identification in Spanish". In: *Natural Language Engineering* 27.2, pp. 131–152.

Torregrossa, François, Robin Allesiardo, Vincent Claveau, Nihel Kooli and Guillaume Gravier (2021). "A survey on training and evaluation of word embeddings". In: *International Journal of Data Science and Analytics* 11.2, pp. 85–103.

Tseytlin, Eugene, Kevin Mitchell, Elizabeth Legowski, Julia Corrigan, Girish Chavan and Rebecca S. Jacobson (2016). "NOBLE – Flexible concept recognition for large-scale biomedical natural language processing". In: *BMC Bioinformatics* 17, pp. 1–15.

Tsochantaridis, Ioannis, Thorsten Joachims, Thomas Hofmann and Yasemin Altun (2005). "Large margin methods for structured and interdependent output variables". In: *Journal of Machine Learning Research* 6.50, pp. 1453–1484.

Tveit, Amund, Ole Edsberg, Thomas Røst, Arild Faxvaag, Øystein Nytrø, Torbjørn Nordgård, Martin Ranang and Anders Grimsmo (2004). "Anonymization of general practioner medical records". In: *Proceedings of the 2nd HelsIT Conference (HelsIT 2004)* (Trondheim, Norway, 10th–24th Sept. 2004), pp. 1–6.

Uzuner, Özlem (2009). "Recognizing obesity and comorbidities in sparse data". In: *Journal of the American Medical Informatics Association* 16.4, pp. 561–570.

Uzuner, Özlem, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian and Brett R South (2012). "Evaluating the state of the art in coreference resolution for electronic medical records". In: *Journal of the American Medical Informatics Association* 19.5, pp. 786–791.

Uzuner, Özlem, Ira Goldstein, Yuan Luo and Isaac Kohane (2008). "Identifying patient smoking status from medical discharge records". In: *Journal of the American Medical Informatics Association* 15.1, pp. 14–24.

Uzuner, Özlem, Yuan Luo and Peter Szolovits (2007). "Evaluating the state-of-the-art in automatic de-identification". In: *Journal of the American Medical Informatics Association* 14.5, pp. 550–563.

Uzuner, Özlem and Amber Stubbs (2015). "Practical applications for Natural Language Processing in clinical research: The 2014 i2b2/UTHealth shared tasks". In: *Journal of Biomedical Informatics* 58, S1–S5.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin (2017). "Attention is all you need". In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (Long Beach, CA, USA). Curran Associates Inc., pp. 6000–6010.

Vellido, Alfredo (2020). "The importance of interpretability and visualization in machine learning for applications in medicine and health care". In: *Neural Computing and Applications* 32, pp. 18069–18083.

Velupillai, Sumithra, Hercules Dalianis, Martin Hassel and Gunnar H. Nilsson (2009). "Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial". In: *International Journal of Medical Informatics* 78.12, e19–e26.

Viani, Natalia, Timothy A. Miller, Carlo Napolitano, Silvia G. Priori, Guergana K. Savova, Riccardo Bellazzi and Lucia Sacchi (2019). "Supervised methods to extract clinical events from cardiology reports in Italian". In: *Journal of Biomedical Informatics* 95.103219, pp. 1–10.

Vincze, Veronika, György Szarvas, Richárd Farkas, György Móra and János Csirik (2008). "The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes". In: *BMC Bioinformatics* 9.Suppl 11, S9.

Vivaldi, Jorge and Horacio Rodríguez (2010). "Using Wikipedia for term extraction in the biomedical domain: first experiences". In: *Procesamiento del Lenguaje Natural* 45, pp. 251–254.

Voorhees, Ellen M. and William Hersh (2012). "Overview of the TREC 2012 Medical Records track". In: *Proceedings of the twenty-first Text REtrieval Conference (TREC 2012)* (Gaithersburg, MD, USA, 6th–9th Nov. 2012). National Institute of Standards and Technology, pp. 1–6.

Wajsbürt, Perceval, Arnaud Sarfati and Xavier Tannier (2021). "Medical concept normalization in French using multilingual terminologies and contextual embeddings". In: *Journal of Biomedical Informatics* 114, pp. 1–9.

Wang, Jing, Huan Deng, Bangtao Liu, Anbin Hu, Jun Liang, Lingye Fan, Xu Zheng, Tong Wang and Jianbo Lei (2020). "Systematic evaluation of research progress on Natural Language Processing in medicine over the past 20 years:

Bibliometric study on PubMed". In: *Journal of Medical Internet Research* 22.1.

Wang, Yanshan, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn and Hongfang Liu (2018). "Clinical information extraction applications: A literature review". In: *Journal of Biomedical Informatics* 77, pp. 34–49.

Warstadt, Alex, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic and Samuel R. Bowman (2019). "Investigating BERT's knowledge of language: Five analysis methods with NPIs". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)* (Hong Kong, China, 3rd–7th Nov. 2019). Association for Computational Linguistics, pp. 2877–2887.

Weegar, Rebecka, Arantza Casillas, Arantza Diaz de Ilarraza, Maite Oronoz, Alicia Pérez and Koldo Gojenola (2016). "The impact of simple feature engineering in multilingual medical NER". In: *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP 2016)* (Osaka, Japan, 11th–16th Dec. 2016). The COLING 2016 Organizing Committee, pp. 1–6.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest and Alexander Rush (2020). "Transformers: State-of-the-art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): System Demonstrations* (Online, 16th–20th Nov. 2020). Association for Computational Linguistics, pp. 38–45.

World Health Organization (2021). "Global strategy on digital health 2020-2025". Tech. rep.

Wu, Stephen, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, Bo Zhao and Hua Xu (2019). "Deep Learning in clinical Natural Language Processing: a methodical review". In: *Journal of the American Medical Informatics Association* 27.3, pp. 457–470.

Yamada, Ikuya, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji and Yuji Matsumoto (2020). "Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and Entities from Wikipedia". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): System Demonstrations*

(Online, 16th–20th Nov. 2020). Association for Computational Linguistics, pp. 23–30.

Yang, Jie, Shuailong Liang and Yue Zhang (2018a). “Design Challenges and Misconceptions in Neural Sequence Labeling”. In: *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)* (Santa Fe, NM, USA, 20th–26th Aug. 2019). Association for Computational Linguistics, pp. 3879–3889.

Yang, Jie and Yue Zhang (2018b). “NCRF++: An open-source neural sequence labeling toolkit”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2018), System Demonstrations* (Melbourne, Australia, 15th–20th July 2018). Association for Computational Linguistics, pp. 74–79.

Yang, Xi, Jiang Bian, Yan Gong, William R. Hoga and Yonghui Wu (2019). “MADEx: A system for detecting medications, adverse drug events, and their relations from clinical notes”. In: *Drug Safety* 42, pp. 123–133.

Yang, Zhongliang, Yongfeng Huang, Yiran Jiang, Yuxi Sun, Yu-Jin Zhang and Pengcheng Luo (2018). “Clinical assistant diagnosis for Electronic Medical Record based on Convolutional Neural Network”. In: *Scientific Reports* 8.6329, pp. 1–9.

Yetano Laguna, Javier and Vicent Alberola (2003). “Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias”. Ministerio de Sanidad y Consumo, Centro de Publicaciones.

Yuan, Zheng, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang and Sheng Yu (2022). “CODER: Knowledge-infused cross-lingual medical term embedding for term normalization”. In: *Journal of Biomedical Informatics* 126, pp. 1–11.

Zhao, Jing, Aron Henriksson, Lars Asker and Henrik Boström (2015). “Predictive modeling of structured electronic health records for adverse drug event detection”. In: *BMC Medical Informatics and Decision Making* 15.Suppl. 4, pp. 1–15.

Zhao, Yiyun and Steven Bethard (2020). “How does BERT’s attention change when you fine-tune? An analysis methodology and a case study in negation scope”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)* (Online, 5th–10th July 2020). Association for Computational Linguistics, pp. 4729–4747.

Zubillaga, Aitor, Paula Laccourreye, Jon Kerexeta, Nekane Larburu, Eduardo Alonso, D Jesús Gómez, Francisco Martínez and Maykel Alonso-Arce (2022). “Hospital readmission prediction via keyword extraction and sentiment analysis on clinical notes”. In: Studies in Health Technology and Informatics 295, pp. 339–342.

Online Resources and References

[1] Instituto de Salud Carlos III (2022). *Memorias BNCS*. URL: https://www.isciii.es/QuienesSomos/CentrosPropios/BNCS/Paginas/Publicaciones_actividades.aspx (visited on 15/12/2022).

[2] SciELO (2022). *Analytics Visualizations*. Scope=Document, Tabulation=by language. URL: http://visual.scielo.org/v1 (visited on 14/12/2022).

[3] SIL International (2022). *What are the top 200 most spoken languages?* URL: https://www.ethnologue.com/guides/ethnologue200 (visited on 20/04/2022).

[4] Vicomtech (2019). *HitzalMed*. URL: https://hitzalmed.nlp.vicomtech.org (visited on 06/12/2022).

[5] — (2018). *UMLSmapper web API*. URL: https://um-public.nlp.vicomtech.org (visited on 11/09/2022).

[6] — (2020). *Vicomtech/NUBes-negation-uncertainty-biomedical-corpus*. URL: https://github.com/Vicomtech/nubes-negation-uncertainty-biomedical-corpus (visited on 30/12/2021).

[7] National Institute of Standards and Technology (2012). *TREC 2011 Medical track*. URL: https://trec.nist.gov/data/medical2011.html (visited on 23/07/2022).

[8] — (2021). *TREC 22021 Clinical Trials track*. URL: http://www.trec-cds.org/2021.html (visited on 23/07/2022).

[9] Radford, Alec, Karthik Narasimhan, Tim Salimans and Ilya Sutskever (2018). *Improving language understanding by generative pre-training*. URL: https://openai.com/blog/language-unsupervised (visited on 23/04/2022).

[10] Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever (2019). *Language models are unsupervised multitask learners*. URL: https://openai.com/blog/better-language-models (visited on 23/04/2022).

[11] Rahimi, Ali (2017). *NIPS 2017 Test-of-time award presentation.* URL: https://www.youtube.com/watch?v=ORHFOnaEzPc (visited on 17/04/2022).

[12] Rahimi, Ali and Ben Recht (2017). *Reflections on Random Kitchen Sinks.* URL: http://www.argmin.net/2017/12/05/kitchen-sinks (visited on 17/04/2022).

[13] World Health Organization (2022). *International Statistical Classification of Diseases and Related Health Problems (ICD).* URL: https://web.archive.org/web/20220204085603/https://www.who.int/classifications/classification-of-diseases (visited on 19/04/2022).

[14] Ministerio de Asuntos Económicos y Transformación Digital (2015). *Plan de Impulso de las Tecnologías del Lenguaje.* URL: https://plantl.mineco.gob.es (visited on 20/04/2022).

[15] — (2018). *II Hacktahon de Tecnologías del Lenguaje en 4YFN.* URL: https://plantl.mineco.gob.es/tecnologias-lenguaje/comunicacion-formacion/eventos/Hackathon-MWC-2018/Paginas/II-Hackathon-MWC-2018.aspx (visited on 20/04/2022).

[16] Cardellino, Cristian (2019). *Spanish Billion Words Corpus and Embeddings.* URL: https://crscardellino.github.io/SBWCE (visited on 22/04/2022).

[17] Cañete, José (2019a). *Spanish Word Embeddings.* URL: https://github.com/BotCenter/spanishWordEmbeddings (visited on 22/04/2022).

[18] — (2019b). *Compilation of Large Spanish Unannotated Corpora.* URL: https://zenodo.org/record/3247731 (visited on 30/12/2021).

[19] Pérez, Jorge (2019). *FastText embeddings from SBWC.* URL: https://github.com/dccuchile/spanish-word-embeddings#fasttext-embeddings-from-sbwc (visited on 22/04/2022).

[20] Campillos-Llanos, Leonardo (2021). *NLPMedTerm deliverables.* URL: http://www.lllf.uam.es/ESP/nlpmedterm_en.html#deliverables (visited on 22/04/2022).

[21] flair (2019). *Flair Embeddings.* URL: https://github.com/flairNLP/flair/blob/master/resources/docs/embeddings/FLAIR_EMBEDDINGS.md (visited on 31/12/2021).

[22] Báez, Pablo, Fabián Villena, Manuel Durán, Matías Rojas and Jocelyn Dunstan (2020a). *The Chilean Waiting List Corpus.* URL: https://doi.org/10.5281/zenodo.3926705 (visited on 20/04/2022).

[23] Google Research (2019). *bert/multilingual.md at master · google-research/bert.* URL: https://github.com/google-research/bert/blob/master/multilingual.md (visited on 11/11/2021).

[24] Johnson, Alistair, Tom Pollard and Roger Mark (2016). *MIMIC-III clinical database.* URL: https://physionet.org/content/mimiciii/1.4 (visited on 23/04/2022).

[25] Allen Institute for AI (2021). *The C4 Multilingual Dataset*. URL: https://github.com/allenai/allennlp/discussions/5265 (visited on 20/04/2022).
[26] Tran, Chris (2020). *Pretrain RoBERTa for Spanish from scratch and perform NER on Spanish documents*. URL: https://github.com/chriskhanhtran/spanish-bert (visited on 30/12/2021).
[27] Ortiz Suárez, Pedro Javier, Benoît Sagot and Laurent Romary (2021). *OSCAR*. URL: https://oscar-corpus.com (visited on 30/12/2021).
[28] Blibioteca Nacional de España (2017). *Recolecciones selectivas. National Library of Spain*. URL: http://www.bne.es/en/Colecciones/ArchivoWeb/Subcolecciones/selectivas.html (visited on 30/12/2021).
[29] Department of Biomedical Informatics at Harvard Medical School (2022). *n2c2 NLP Research Data Sets*. URL: https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp (visited on 20/11/2022).
[30] Gonzalez-Agirre, Aitor, Ander Intxaurrondo, Jose Antonio Lopez-Martin, Montserrat Marimon, Jesus Santamaria, Felipe Soares, Marta Villegas and Martin Krallinger (2019b). *MEDDOCAN*. URL: https://temu.bsc.es/meddocan (visited on 18/11/2021).
[31] Medical Imaging Databank of the Valencia Region (2021). *De-identifying Spanish medical texts - Named Entity Recognition applied to radiology reports*. URL: https://bimcv.cipf.es/bimcv-projects/dismed (visited on 20/11/2022).
[32] MAPA EU Project (2021a). *MAPA trained models for entity detection*. URL: https://gitlab.com/MAPA-EU-Project/mapa_project/-/blob/master/avaliable_mapa_trained_models.md (visited on 20/11/2022).
[33] — (2021b). *MAPA - Anonymization packages*. URL: https://elrc-share.eu/repository/search/?q=mfsp:b550e1a88a8311ec9c1a00155d026706687917f92f64482587c6382175dffd76 (visited on 20/11/2022).
[34] Soares, Felipe, Aitor Gonzalez-Agirre and Martin Krallinger (2019b). *Spanish Clinical Case Corpus Part-of-Speech Tagger*. URL: https://github.com/PlanTL-GOB-ES/SPACCC_POS-TAGGER (visited on 10/11/2021).
[35] Gonzalez-Agirre, Aitor, Ander Intxaurrondo, Jose Antonio Lopez-Martin, Montserrat Marimon, Jesus Santamaria, Felipe Soares, Marta Villegas and Martin Krallinger (2019c). *Resources – MEDDOCAN*. URL: https://temu.bsc.es/meddocan/index.php/resources (visited on 10/11/2021).
[36] Heilman, Michael (2013). *Hierarchical word clustering, following "Brown clustering" (Brown et al., 1992)*. URL: https://github.com/mheilman/tan-clustering (visited on 11/11/2021).
[37] Explosion (2016). *spaCy · Industrial-strength Natural Language Processing in Python*. URL: https://spacy.io (visited on 11/11/2021).

[38] Explosion (2017a). *spaCy's NER model · spaCy Universe*. URL: https://spacy.io/universe/project/video-spacys-ner-model (visited on 11/11/2021).

[39] — (2017b). *Training spaCy's Statistical Models · spaCy Usage Documentation (legacy)*. URL: https://v2.spacy.io/usage/training#ner (visited on 11/11/2021).

[40] Yang, Jie (2018). *NCRF++, a Neural Sequence Labeling Toolkit. Easy use to any sequence labeling tasks (e.g. NER, POS, Segmentation). It includes character LSTM/CNN, word LSTM/CNN and softmax/CRF components.* URL: https://github.com/jiesutd/NCRFpp (visited on 11/11/2021).

[41] Hugging Face (2018). *Transformers: State-of-the-art Natural Language Processing for Pytorch, TensorFlow, and JAX*. URL: https://github.com/huggingface/transformers (visited on 11/11/2021).

[42] Instituto Nacional de Estadística (2019). *INEbase/Demografía y población/Padrón. Población por municipios/Apellidos y nombres más frecuentes/Resultados*. URL: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177009&menu=resultados&idp=1254734710990 (visited on 11/11/2021).

[43] National Library of Medicine (2021a). *UMLS Metathesaurus Browser - Concept C0004521*. URL: https://uts.nlm.nih.gov/uts/umls/concept/C0004521 (visited on 22/11/2021).

[44] — (2022). *UMLS Release File Archives: 2004-2022AA*. URL: https://www.nlm.nih.gov/research/umls/licensedcontent/umlsarchives04.html (visited on 26/11/2022).

[45] Bruno Kessler Foundation (2020). *ELG - European Clinical Case Corpus*. URL: https://live.european-language-grid.eu/catalogue/project/1312 (visited on 31/03/2022).

[46] The Apache Software Foundation (2011). *Apache Lucene - Welcome to Apache Lucene*. URL: https://lucene.apache.org (visited on 11/11/2021).

[47] IXA NLP Group (2014). *ixa2.si.ehu.eus/ixa-pipes*. URL: https://ixa2.si.ehu.eus/ixa-pipes (visited on 11/11/2021).

[48] — (2009). *UKB: Graph based Word Sense Disambiguation and Similarity*. URL: https://ixa2.si.ehu.eus/ukb (visited on 11/11/2021).

[49] National Library of Medicine (2021b). *Metathesaurus - Rich Release Format (RRF) - UMLS® Reference Manual - NCBI Bookshelf*. URL: https://www.ncbi.nlm.nih.gov/books/NBK9685 (visited on 11/11/2021).

[50] — (2016). *UMLS Glossary - Supressibility*. URL: https://www.nlm.nih.gov/research/umls/new_users/glossary.html#s (visited on 10/11/2021).

[51] Soriano, Bàrbara, Oriol Borrega, Mariona Taulé and Mª Antònia Martí (2008). *Guidelines*. URL: http://.ub.edu/corpus/webfm_send/17 (visited on 29/11/2021).

[52] The Apache Software Foundation (2021a). *Package org.apache.lucene.queryparser.classic (Lucene 8.11.0 API)*. URL: https://lucene.apache.org/core/8_11_0/queryparser/org/apache/lucene/queryparser/classic/package-summary.html (visited on 16/11/2021).

[53] — (2021b). *Class TFIDFSimilarity (Lucene 8.11.0 API)*. URL: https://lucene.apache.org/core/8_11_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html (visited on 16/11/2021).

[54] National Library of Medicine (2020a). *MetaMap*. URL: https://lhncbc.nlm.nih.gov/ii/tools/MetaMap.html (visited on 11/11/2021).

[55] — (2020b). *Data File Builder*. URL: https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/additional-tools/DataFileBuilder.html (visited on 11/11/2021).

[56] UFAL (2017). *UFAL Medical Corpus | ÚFAL*. URL: https://ufal.mff.cuni.cz/ufal_medical_corpus (visited on 11/11/2021).

[57] National Library of Medicine (2021c). *UMLS Metathesaurus Browser - Concept C0029235*. URL: https://uts.nlm.nih.gov/uts/umls/concept/C0029235 (visited on 13/10/2021).

[58] — (2021d). *UMLS Metathesaurus Browser - Concept C0199182*. URL: https://uts.nlm.nih.gov/uts/umls/concept/C0199182 (visited on 13/10/2021).

[59] — (2021e). *UMLS Metathesaurus Browser - Concept C1708335*. URL: https://uts.nlm.nih.gov/uts/umls/concept/C1708335 (visited on 13/10/2021).

[60] — (2021f). *UMLS Metathesaurus Browser - Concept C0015264*. URL: https://uts.nlm.nih.gov/uts/umls/concept/C0015264 (visited on 13/10/2021).

[61] Laboratorio de Lingüística Informática UAM (1999). *UAM Spanish Treebank*. URL: http://www.lllf.uam.es/ESP/Treebank.html (visited on 31/03/2022).

[62] Marimon, Montserrat, Jorge Vivaldi and Núria Bel (2017b). *Negation On CR IULA*. URL: http://eines.iula.upf.edu/brat//#/NegationOnCR_IULA (visited on 31/03/2022).

[63] CLiC - Centre de Llenguatge i Computació (2016). *SFU Review-NEG*. URL: http://clic.ub.edu/corpus/en/node/172 (visited on 31/03/2022).

[64] — (2022). *NewsCom-NEG*. URL: http://clic.ub.edu/corpus/newscom_neg-ca (visited on 14/04/2022).

[65] Grupo de Ingeniería Lingüística - UNAM (2019). *Negation and Sentiment Detection on Mexican Spanish Tweets: The T-MexNeg Corpus*. URL: https://gitlab.com/gil.iingen/negation_twitter_mexican_spanish (visited on 31/03/2022).

[66] Soares, Felipe, Marta Villegas, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé and Martin Krallinger (2020). *FastText and Word2Vec Spanish Medical Embeddings*. URL: https://zenodo.org/record/3626806 (visited on 31/12/2021).

[67] flair (2020). *Tutorial 8: Model optimization*. Removed as of 15/11/2021. URL: https://github.com/flairNLP/flair/blob/master/resources/docs/TUTORIAL_8_MODEL_OPTIMIZATION.md.

[68] chakki (2018). *A Python framework for sequence labeling evaluation (named-entity recognition, pos tagging, etc...)* URL: https://github.com/chakki-works/seqeval (visited on 31/12/2021).

[69] Santamaría, Jesús (2018). *NegEx-MES*. URL: https://github.com/PlanTL-SANIDAD/NegEx-MES (visited on 06/12/2021).

[70] Kang, Peter (2009). *A python module to implement Wendy Chapman's NegEx algorithm*. URL: https://github.com/chapmanbe/negex/tree/master/negex.python (visited on 21/01/2022).

[71] National Library of Medicine (2023). *UMLS Metathesaurus Browser - Concept C1698536*. URL: https://uts.nlm.nih.gov/uts/umls/concept/C1698536 (visited on 20/01/2023).

[72] Antwerpen, Universiteit (2012). *Resolving the Scope and Focus of Negation - *SEM 2012 Shared Task*. URL: https://www.clips.uantwerpen.be/sem2012-st-neg (visited on 12/04/2022).

List of Abbreviations

ADR adverse drug reaction
AI Artificial Intelligence
API Application Programming Interface
brat brat rapid annotation tool
CC clinical case
CDS clinical decision support
CLEF Cross-Lingual Evaluation Forum
CSF Castro et al. (2010) Scoring Function
cTAKES clinical Text Analysis and Knowledge Extraction System
CUI Concept Unique Identifier
DeCS Descriptores en Ciencias de la Salud
EMEA European Medicines Agency
GDPR General Data Protection Regulation
GUI Graphic User Interface
HIPAA Health Insurance Portability and Accountability Act
HTTP Hypertext Transfer Protocol
i2b2 Informatics for Integrating Biology and the Bedside
IberEval Workshop on Evaluation of Human Language Technologies for Iberian Languages
IberLEF Iberian Languages Evaluation Forum
ICD International Classification of Diseases
ID Identification
INE Insituto Nacional de Estadística
JSON JavaScript Object Notation
LOINC Logical Observation Identifiers Names and Codes
LSF Lucene Scoring Function
Mantra GSC Mantra Gold Standard Corpus

MEDDOCAN Medical Document Anonymization
MedDRA Medical Dictionary of Regulatory Activities
MeSH Medical Subject Headings®
n2c2 National NLP Clinical Challanges
NLM National Library of Medicine
NLNDE Neither-Language-nor-Domain-Experts
PHI Personal Health Information
PM precision medicine
RRF Rich Release Format
SCTSPA Spanish translation of SNOMED CT
SemEval International Workshop on Semantic Evaluation
SNOMED CT Systematized Nomenclature of Medicine – Clinical Terms
SPACCC Spanish Clinical Case Corpus
TASS Taller de Análisis Semántico
TCP Transmission Control Protocol
TREC Text REtrieval Conference
TUI Type Unique Identifier
UIMA Unstructured Information Management applications
UMLS Unified Medical Language System
URL Uniform Resource Locator
WHO World Health Organization
XAI Explainable AI

Electronic Health Records

CC Chief Complaint
DXT Diagnostic Tests
EHR Electronic Health Record
HPI History of Present Illness
hx Patient History
PE Physical Examination
PNo Progress Notes
TNo Treatment Notes

Languages

de German
en English
es Spanish
eu Basque
fr French
hu Hungarian
it Italian
multi Multilingual

SPA Spanish

Latinisms

cf. confer (compare)
e.g. exempli gratia (for example)
etc. et cetera (and so on)
i.a. inter alia (among others)
ibid. ibidem (in the same place)
i.e. id est (that is)
lit. literal meaning
N.B. nota bene (in the same place)

Machine Learning

ANN Artificial Neural Network
biGRU bidirectional GRU
biLSTM bidirectional LSTM
BPE Byte-Pair Encoding
CNN Convolutional Neural Network
CRF Conditional Random Field
Dev Development data split
DL Deep Learning
DNN Deep Neural Network
emb embedding
FFNN Feedforward Neural Network
GRU Gated Recurrent Unit
LSTM Long Short-Term Memory
ML Machine Learning
MLP Multilayer Perceptron
NB Naïve Bayes
RNN Recurrent Neural Network
SGD Stochastic Gradient Descent
SVM Support Vector Machine

Measures and Symbols

Acc accuracy
Ave average
BLEU Bilingual Evaluation Understudy
BP Brevity Penalty
$\mathbf{F_1}$ F_1-score
FN false negative
FP false positive
HE human error

J Jaccard coefficient
κ Cohen's kappa coefficient
Lk leak
lwκ linearly weighted κ
Max maximum
μ micro-average
Min minimum
OP overlap percentage
P precision
R recall
Tot total
TP true positive

Medical Specialities and Care Units

AN Anaesthesiology
CD Cardiovascular Diseases
GCU Geriatric Convalescence Unit
GE Gastroenterology
GS General Surgery
HaH Hospital at Home
ICU Intensive Care Unit
IM Internal Medicine
N Neurology
OBG Obstetrics and Gynaecology
ODO Odontology
OPH Ophthalmology
OR Orthopaedics
OTO Otolaryngology
PS Plastic Surgery
TS Thoracic Surgery
U Urology
VS Vascular Surgery

Natural Language Processing

ASR Automatic Speech Recognition
BERT Bidirectional Encoder Representations from Transformers
bioNLP biomedical NLP
BLP Biomedical Language Processing
CBOW continuous bag-of-words
CL Computational Linguistics
DD Domain Dependency
ELMo Embeddings from Language Models

GloVe Global Vectors
GPT Generative Pre-trained Transformer
IE Information Extraction
IR Information Retrieval
LD Language Dependency
LM Language Model
LT Language Technology
mBERT Multilingual BERT
MER Medical Entity Recognition
MERC Medical Entity Recognition and Classification
MLM Masked Language Model
MT Machine Translation
NAF NLP Annotation Format
NER Named Entity Recognition
NERC Named Entity Recognition and Classification
NLG Natural Language Generation
NLP Natural Language Processing
NLU Natural Language Understanding
NMT Neural MT
NSP Next Sentence Prediction
OOV out-of-vocabulary words
PoS Part of Speech
QA Question Answering
WSD Word Sense Disambiguation

Negation and Uncertainty

`abs` absent
`NCue` negation cue
`Neg` negation
`NLex` lexical negation cue
`NMph` morphological negation cue
NPI negative polarity-sensitive item
`NSco` negation scope
`NSyn` syntactic negation cue
`Pol` polarity item
`pos` possible
`pre` present
`UCue` uncertainty cue
`ULex` lexical uncertainty cue
`Unc` uncertainty
`USco` uncertainty scope

USyn syntactic uncertainty cue

Semantic Groups of the UMLS Metathesaurus

anat anatomy
chem chemical or drug
devi device
diso disorder
geog geographic area
livb living being
objc object
phen phenomenon
phys physiology
proc procedure

Semantic Types of the UMLS Metathesaurus (partial list)

acty activity
bhvr behaviour
cnce conceptual entity
enty entity
evnt event
fndg finding
sosy sign or symptom

Sensitive Data Categories

Age patient's age
Cli outpatients clinic
Con contact information
Ctr country
Dat date
Did doctor's ID
Doc doctor's name
Eid episode ID
Ema e-mail address
Fac healthcare facility
Fax fax number
Hos hospital
Ide identification number
Iid insurance ID
Ins institution
Job patient's profession
Kin patient's relative
Lid license ID

`Loc` location
`Oth` other
`Pat` patient's name
`Pho` phone number
`Pid` patient's ID
`Sex` patient's sex
`Str` street
`Ter` territory
`Tim` time

Sequence Tagging Categories

`B-` Beginning
`I-` Inner
`L-` Last
`O` Outside
`U-` Unique

Syntax

A adjective
AP adjective phrase
N noun
NP noun phrase
P preposition
PP prepositional phrase
R relative clause
S clause
V verb
VP verb phrase