Extracción de información en informes radiológicos escritos en español

Cotik, Viviana Erica

2018

UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
DEPARTAMENTO DE COMPUTACIÓN

Extracción de información en informes radiológicos escritos en español

Tesis presentada para optar al título de Doctora de la Universidad de Buenos Aires
en el área de Ciencias de la Computación

Viviana Erica Cotik

Director de tesis:	Dr. José Castaño
Consejera de estudios:	Dra. Irene Loiseau
Lugar de trabajo:	Departamento de Computación Facultad de Ciencias Exactas y Naturales Universidad de Buenos Aires

Buenos Aires, Abril 2018

Extracción de información en informes radiológicos escritos en español

Resumen:

En los últimos años, la cantidad de información clínica disponible en formato digital ha crecido constantemente debido a la adopción del uso de sistemas de informática médica. En la mayoría de los casos, dicha información se encuentra representada en forma textual. La extracción de información contenida en dichos textos puede utilizarse para colaborar en tareas relacionadas con la clínica médica y para la toma de decisiones, y resulta esencial para la mejora de la atención médica.

El dominio biomédico tiene vocabulario altamente especializado, local a distintos países, regiones e instituciones. Se utilizan abreviaturas ambiguas y no estándares. Por otro lado, algunos tipos de informes médicos suelen presentar faltas ortográficas y errores gramaticales. Además, la cantidad de datos anotados disponibles es escasa, debido a la dificultad de obtenerlos y a temas relacionados con la confidencialidad de la información. Esta situación dificulta el avance en el área de extracción de información.

Pese a ser el segundo idioma con mayor cantidad de hablantes nativos en el mundo, poco trabajo se ha realizado hasta ahora en extracción de información de informes médicos escritos en español. A los desafíos anteriormente descriptos se agregan la ausencia de terminologías específicas para ciertos dominios médicos y la menor disponibilidad de recursos lingüísticos que los existentes para otros idiomas.

En este trabajo contribuimos al dominio de la biomedicina en español, proveyendo métodos con resultados competitivos para el desarrollo de componentes fundamentales de un proceso de extracción de información médico, específicamente para informes radiológicos.

Con este fin, creamos un corpus anotado de informes radiológicos en español para el reconocimiento de entidades, negación y especulación y extracción de relaciones. Publicamos el proceso seguido para la anotación y el esquema desarrollado.

Implementamos dos algoritmos de detección de entidades nombradas con el fin de encontrar entidades anatómicas y hallazgos clínicos. El primero está basado en un diccionario especializado del dominio no disponible en español y en el uso de reglas basadas en conocimiento morfosintáctico y está pensado para trabajar con lenguajes sin muchos recursos lingüísticos. El segundo está basado en campos aleatorios condicionales y arroja mejores resultados.

Adicionalmente, estudiamos e implementamos distintas soluciones para la detección de hallazgos clínicos negados. Para esto, adaptamos al español un conocido algoritmo de detección de negaciones en textos médicos escritos en inglés y desarrollamos un método basado en reglas creadas a partir de patrones inferidos del análisis de caminos en árboles de dependencias. También adaptamos el primer método, que arrojó los mejores resultados, para la detección de negación y especulación en resúmenes de alta hospitalaria y notas de evolución clínica escritos en alemán.

Consideramos que los resultados obtenidos y la publicación de criterios de anotación y evaluación contribuirán a seguir avanzando en la extracción de información de informes clínicos escritos en español.

palabras clave: detección de entidades nombradas, detección de negación y especulación, BioNLP, biomedicina, anotación de corpus, informes radiológicos, extracción de información, minería de textos.

Information extraction from Spanish radiology reports

Abstract:

In the last years, the number of digitized clinical data has been growing steadily, due to the adoption of clinical information systems. A great amount of this data is in textual format. The extraction of information contained in texts can be used to support clinical tasks and decisions and is essential for improving health care.

The biomedical domain uses a highly specialized and local vocabulary, with abundance of non-standard and ambiguous abbreviations. Moreover, some type of medical reports present ill-formed sentences and lack of diacritics. Publicly accessible annotated data is scarce, due to two main reasons: the difficulty of creating it and the confidential nature of the data, that demands de-identification. This situation hinders the advance of information extraction in the biomedical domain area.

Although Spanish is the second language in terms of numbers of native speakers in the world, not much work has been done in information extraction from Spanish medical reports. Challenges include the absence of specific terminologies for certain medical domains in Spanish and the availability of linguistic resources, that are less developed than those of high resources languages, such as English.

In this thesis, we contribute to the BioNLP domain by providing methods with competitive results to apply a fragment of a medical information extraction pipeline to Spanish radiology reports.

Therefore, an annotated dataset for entity recognition, negation and speculation detection, and relation extraction was created. The annotation process followed and the annotation schema developed were shared with the community.

Two named entity recognition algorithms were implemented for the detection of anatomical entities and clinical findings. The first algorithm developed is based on a specialized dictionary of the radiology domain not available in Spanish and in the use of rules based on morphosyntactic knowledge and is designed for named entity recognition in medium or low resource languages. The second one, based on conditional random fields, was implemented when we were able to obtain a larger set of annotated data and achieves better results.

We also studied and implemented different solutions for negation detection of clinical findings: an adaptation to Spanish of a popular negation detection algorithm for English medical reports and a rule-based method that detects negations based on patterns inferred from the analysis of paths of dependency parse trees. The first method obtained the best results and was also adapted for negation and speculation detection in German clinical notes and discharge summaries.

We consider that the results obtained, and the annotation guidelines provided will bring new benefits to further advance in the field of information extraction from Spanish medical reports.

keywords: named entity recognition, negation and speculation detection, BioNLP, annotation guidelines, annotation schema, Spanish radiology reports, information extraction, text mining.

Agradecimientos

Quiero agradecer a todos aquellos que con su comprensión, colaboración, estímulo y reconocimiento académico permitieron la concreción de este trabajo.

A mis padres, por su constante apoyo y su inmenso amor. A Tomás, So-Min y Yuni.

A Dany, Naty e Ignacio, que me ayudaron a pensar y resolver distintas situaciones que se fueron presentando. A Laura, Lore, Vale, Maiu, Ale, Leti, Celina, Nat, Jime, Ana, Adri y San por su apoyo, ayuda y contención. A Vir, Flavia, Rodo, Dany M., Andre C., Mariana, Matilde, Graciela y Andre B., que siempre están. Por su contención y por comprender que me haya recluído para terminar este trabajo. También a Marco, Barbi, Laura, X, Joke, Ana, Diego y Gaby. A Ale, Vivi E. y Cyn. A Jor, Pablo y Fer, que me hacen reir y me dan ánimos. A Andrés por acompañarme en parte de este recorrido y ayudarme a pensar. También a Ignacio A.H., Viviana B., Nurit, Susana L., Martha y Hector.

A Aleksandra, Ilka, Kathrin, Iliana, Bárbara, Andrej y Larissa, por ser parte de mi familia durante mi estadía en Alemania. A Renlong, Seong, Philippe, Leonhard y Roland.

A José por dedicar su tiempo e introducirme en la realización de este trabajo. A Horacio Rodríguez y a Jorge Vivaldi por guiarme, por las innumerables discusiones, lecturas y comentarios respecto a mi trabajo. De todos aprendí mucho durante este proceso.

A Hans Uszkoreit y Feiju Xu por confiar en mi y ofrecerme hacer una estadía de investigación en la Universidad de Saarland y en el Centro Alemán de Investigación en inteligencia artificial (DFKI). A Rob Gaizauskas y Angus Roberts por recibirme tan cálidamente en la Universidad de Sheffield y a Horacio Rodríguez por abrirme las puertas en la Universidad Politécnica de Cataluña (UPC).

A Darío por sus palabras de apoyo, por su entusiasmo y ayuda constante.

A mis amigos y compañeros del Departamento de Computación. En especial a Alexandra, Rosana, Irene, Santiago, Ricardo, Verónica, Isabel y Agustín. A Pablo L. y a Mariano B., quienes siempre me ayudaron. A Paula, Azul y Vicky. A Mariano C., David, Juan Manuel, Pablo B., Matías, Mariano R., Mariano M. y a mis compañeros de oficina históricos. A Vanesa, Aída, Ale y Nico R. A mis compañeros de Tleng.

Al Departamento de Computación, que colaboró en múltiples ocasiones para que pudiera presentar mis trabajos en congresos de la especialidad, al DFKI y a la Universidad de Saarland, que, en reconocimiento a mi trabajo, me ayudaron a presentarlo en el exterior y a la ACL por los mismos motivos.

A todos aquellos que leyeron secciones de mi tesis y me ayudaron a mejorarla (en especial a Jorge, Horacio R., Rodo, Leo, Horacio C., Mariano C., Andrés, Laura, Ale, Vale y Franco). A Olivier, por ayudarme a preparar la presentación.

A los revisores de mis trabajos, en particular a los de este, por sus valiosos comentarios.

A todos aquellos, a los que por descuido y agotamiento no incluí en esta lista.

A Luna y a Lenka.

Contents

List of Figures

List of Tables

Part I

Prelude

CHAPTER 1

Introduction

We will begin this chapter with Section 1.1, where we present the problem addressed in our research, why it is important and problematic. Then, we will provide background information about the area by briefly reviewing previous research and current importance of the topic. Next, we will point at some aspects that are missing in previous research. Section 1.2 presents our main research questions. Later, our contributions are presented in Section 1.3, and finally the structure of the thesis and the dissemination of the obtained results are shown in Sections 1.4 and 1.5.

1.1 Motivation

In the last thirty years there has been an exponential growth in the amount of digital literature available [126, 13]. This large amount of accumulated information - mainly of textual nature- has promoted research in textual information systems that facilitate the analysis, management, access and automatic processing of this great amount of data. The necessary technology to face this problem needs disciplines such as natural language processing (NLP) and information extraction (IE).

Particularly, in areas such as molecular biology, where there is a large number of scientific articles and an accelerated discovery of information (e.g. the sequencing of the human genome, some years ago), biologists, bioinformaticians and other researchers are not able to keep up with the hundreds of scientific articles retrieved by a query. It would be more convenient to reduce the set of documents to access or to access a database, with already processed information in order to retrieve later the scientific articles that are of interest for the researcher.

On the other hand, in the medical domain, the information produced by physicians in the format of electronic health records (EHR), clinical notes, discharge summaries and radiology reports, among others, is being digitized steadily with the adoption of information systems in the medical domain [26, 32, 112].[1,2] A great

[1] It is also important to notice that in many sanitary centers due to different reasons, medical records are still being completed on paper [105, 4, 205], leading to difficulties in finding them and to retrieving information in a timely manner. Statistics about penetration of electronic health records in Latin America can be seen in http://globalhealthintelligence.com/ghi-analysis/electronic-medical-records-growing-in-latin-america/ (accessed Mar. 2018).

[2] The first EHR system was created more than 40 years ago, but its adoption has been increasing steadily only in the last years [112]. https://oli.cmu.edu/jcourse/workbook/activity/page?

amount of this information is in textual format because writing provides a much richer detail than completing forms with structured fields [272, 83]. Physicians often have to read through these texts to gain insight on the details of the patients diseases or history. This is a very time-consuming activity, and time is a scarce resource. However, the availability of information in structured format can contribute to trigger automatic alerts about situations that require action or attention of a physician (obtaining timely information is critical in case of urgent or important findings [24, 37, 23]). It is also useful as input to systems that help physicians in decision making with assessments based on matching patient information with knowledge bases (clinical decision support systems) [72, 239]. Finally, precision medicine, an approach that proposes the customization of healthcare by taking into account genes, environment and lifestyle of a person in order to provide tailored treatment and prevention strategies, is benefited by the availability of structured data.[3]

Therefore, it is of utmost importance, to have mechanisms to obtain structured information from unstructured texts. The codification of this information to entries in known terminologies is also useful.

The action of turning information into structured data is not only useful for reducing costs and accelerating knowledge discovery of explicit and implicit pieces of information that help to diagnose, to prevent health problems and to treat patients [229, 50]. It is also significant for knowing the patients, the number of cases of different illnesses, preparing budgets, working on global prevention plans and accessing medical records based on their content [272]. Basically, it is useful to support clinical tasks and decisions. Moreover, automatic identification of relevant terms in medical reports is a preliminary step for indexing and for search tools and it is also useful for clinical and research purposes.

Besides, clinical reports often contain a large number of expressions of negation and speculation. It is important to recognize whether the mentions of medical conditions occurring in reports are presented as factual, as counterfactual (absent) or as speculated (suspected), since extracted information that is under the scope of negation or speculation cannot be presented as factual [173].

Finally, being able to classify reports among those containing an asserted medical condition from those that do not, is also important. In the case of radiology reports, it can be useful to automatically retrieve texts with their belonging images corresponding to specific medical findings and use them for educational purposes. It can also serve for selecting which reports to read with attention.

There are many works that address the extraction of information in different domains, but not all domains have the same difficulty. The biomedical domain has highly specialized vocabulary. Besides, the medical jargon is local and differs in distinct countries (where the same language is spoken), regions, medical institutions or even among professionals of the same institution. Finally, the naming of biomedical terms is imprecise, a variety of names can be used for the same concept (due to synonymia, different Graeco-Latin transliterations, spelling and orthographic variations and different form of abbreviating terms, among others) [13, 148, 90] and a name can have different meanings.

There exist different kind of texts in the biomedical domain: from research papers, abstracts and carefully written documents (such as drug leaflets) to medical narratives, such as reports originated from imaging studies (imaging reports).

context=e6f7b56780020ca60106943387dcc70b (accessed Mar. 2018).

[3] National Institute of Health (NIH) definition of precision medicine can be seen in: https://ghr.nlm.nih.gov/primer/precisionmedicine/definition (accessed Mar. 2018).

Among them, there are different formats: unstructured or semi-structured, longer or shorter, well written or with many orthographic and syntactic errors. Some institutions provide tools in order to register diseases in a structured way, presenting different options provided by a terminology service [97]. One of these institutions is the *Hospital Italiano de Buenos Aires*, Argentina.[4]

Clinical narrative can proceed from data entry, or dictation and transcription. Particularly, reports that are written in front of the patient, or in a short amount of time -for example, imaging studies-, have usually orthographic and grammatical errors. In fact, unlike many other text types, such as scientific articles and discharge summaries, radiology reports are often written in a rather telegraphic style (sometimes without verbs, without punctuation signs and with missing diacritics -such as the ∼ symbol and accents-) [272]. There are few guidelines for writing medical reports and there is a lack of consensus about what constitutes a good report. According to Halls' description about what constitutes a good radiology report [111], sentences should be short, and acronyms should be used. He also describes the abundance of *non-sentences*, such as "no evidence of malignancy" and the bad use of *hedges*.[5] Furthermore, all kind of medical reports contain many technical terms as well as non-standard abbreviations, whose ambiguity is much higher than in other domains [142, 204]. Even the same abbreviation can be used with different meaning by distinct medicine specialties [204] or by different physicians of the same specialty in a hospital. Furthermore, a large number of abbreviations do not follow a naming convention. Many of these issues also appear in social media texts [211]. According to Simpson and Demner-Fushman [229], "the meaning and grammar of biomedical texts are so intertwined that all surveys dedicate a section to natural language preprocessing and grammatical analysis".

In order to compare results of different proposals it is of utmost importance to have common datasets and evaluation standards. Annotated data, where experts say which is the solution to the problem, is needed to be able to evaluate the performance of information extraction techniques. In the medical domain publicly available annotated datasets are very scarce. Obtaining annotated corpora is more difficult than in other domains, given that:

- medical data is of sensitive nature, due to the presence of personal information, and is therefore usually non-publicly available nor easy to obtain privately. In order to be shared, data has to be anonymized, in such a way that not only the identity of the patients and of the intervening physicians is unknown, but also that it is not inferable. It also has to remain useful for subsequent analysis (non-perturbative anonymization) [100, 84, 85, 101]. Permission to publish the data is rarely given (it depends on the owners of the data and on the

[4] Hospital Italiano de Buenos Aires, https://www.hospitalitaliano.org.ar/#!/home/principal (accessed Mar. 2018).

[5] Hedges are epistemic modality markers.

institutional, country or region data sharing policies and legislation),[6,7,8,9]

- annotation expertise is required. It is not easy to find annotators. A special training or certain educational degree is needed to be able to understand biomedical texts and it is not easy to find human resources who meet these requirements and that are available to work in annotation projects [244]. Due to the above-mentioned difficulties it is also difficult to obtain a high agreement among annotators, even in the cases where precisely defined annotation guidelines, training and a controlled annotation methodology is followed. Furthermore, the difficulty of the domain makes it not an easy task for crowdsourcing.

There is also no agreement on the criteria to be used for evaluating the results. Even among experts there might be criteria differences about which are the boundaries of an information unit of interest and about which is the type of the information unit, and therefore not only exact matches, but also partial matches are considered [229, 270, 189, 131, 77]. Furthermore, there is no standard definition for partial matches [255, 87].

Given the scarce availability of public accessible data and the lack of common evaluation standards results can be very biased to the dataset they were tried with and it is difficult to compare different techniques.

The area of study that deals with information extraction from biomedical texts is called BioNLP or biomedical text mining. It is also referred to as information extraction in the biomedical domain.

Much work has been done in extracting information. Some examples of these information extraction systems are: Tacitus, Fastus and TextPro (all of them from the Stanford Research Institute -SRI-), Proteus (New York University), QXtract (Columbia 2003), LaSIE -Large Scale Information Extraction- (University of Sheffield, England) and Avatar (IBM). Information extraction systems were also developed for specific topics, for example, MedLEE and CaTIES (both of the medical domain) and SUISEKI (biological domain). Negation and speculation detection have also been studied in the general [207, 280] and in the biomedical domain [43, 284, 187, 125].

BioNLP is a very active area at this moment. It is growing steadily. A large number of events, such as workshops, tutorials and invited talks in the most prominent conferences on natural language processing are being produced yearly. Also, special issues in specialized journals, books and surveys have been published in the last years [138, 11, 50, 72, 293, 291, 41, 229, 124]. Furthermore, there is an important tradition of challenges in this area (such as BioCreAtiVe, i2b2, ConLL, CLEF and BioASQ), which help advance the field and provide useful corpora for research.

[6] For example, the *Health Insurance Portability Accountability Act* (HIPAA) is a United States legislation, that requires the anonymizaton with regards to patient information of all medical records that are going to be accessed outside of the clinical setting in which they were produced.

[7] The European Data Protection Directive 95/46/EC states that "'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity".

[8] In Argentina laws 26,529 (*Patient rights in relation with health professionals and institutions*) -art. 2-, 17,132 (*of the exercise of medicine*) -art. 11- and 25,326 (*protection of personal data -habeas data-*) -art. 8- mention the obligation to preserve the confidentiality of medical reports concerning patients, which can only be disclosed if a judicial authority or another law so determines it or if the patient authorizes it. Law 17,132 -art. 11- establishes that medical reports may be published for research goals.

[9] Other data sharing policies are *UK Data Protection Act* (1998), and *Canada Personal Information Protection and Electronic Documents Act* (2000).

There also exist commercial applications that deal with this subject, such as SAP's Healthcare,[10] an application designed for clinical research, which, among others, combines structured and unstructured information in order to obtain meaningful medical information from big data, and IBM Watson's Health,[11] that is applying NLP techniques to improve clinical decision support systems. There are also worldwide projects addressing the issue, such as Khresmoi project,[12] that uses information extraction from unstructured biomedical texts in a cross-lingual environment. There also exist different kind of resources for text analysis, such as terminologies, and tools for extracting information from texts, some of them adapted to the biomedical domain. The importance of BioNLP can be highlighted if we consider, for example, the areas of work of the *American Medical Informatics Association* (AMIA),[13] that include information management for clinical data; development of clinical decision support systems; the use of knowledge to enhance public health policies and consumer health informatics, focused in empowering patients so they are able to understand and participate in the management of their own health, by means of having access to the information of comprehensible nature. Moreover, the *Health Natural Language Processing Center* (hNLP)[14] targets the lack of shared biomedical data, a fact that hinders reproducibility and improvement of results, by focusing on providing de-identified clinical notes from several institutions and by sponsoring research programs, BioNLP challenges and scientific events. hNLP follows the tradition of the *Linguistic Data Consortium* (LDC)[15] and of the *European Language Resources Association* (ELRA)[16] for general language resources. Finally, there exist important fundings for NLP projects applied to medical diagnosis.[17]

More attention has been given to the biological domain than to the medical or clinical domain and to well-written texts rather than informal texts.

The extraction of information possess dissimilar challenges for different languages and regions. In some countries, such as India and China more than 15 and 9 official languages exist [228].[18,19,20] German has compound nouns, so compound-splitter modules are needed. It also has long dependencies. East Asian languages don't have spaces between words, so segmentation at a character level has to be used to detect words.

The greater amount of work has been carried out for English, although less than 10% of the world population primarily speak this language [63]. There is also work

[10] SAP Healthcare https://www.sap.com/industries/healthcare.html (accessed Mar. 2018).

[11] IBM Watson Health https://www.ibm.com/watson/health/.

[12] Kreshmoi project http://www.khresmoi.eu/ (accessed Mar. 2018).

[13] AMIA: https://www.amia.org/about-amia (accessed Mar. 2018).

[14] hNLP https://healthnlp.hms.harvard.edu/center/pages/home.html, hNLP publications: https://healthnlp.hms.harvard.edu/center/pages/publications.html (both accessed Mar. 2018).

[15] The Linguistic Data Consortium (LDC) is a consortium of entities, found to address the data shortage for natural language processing research. Currently it is a repository of language resources and sponsors research programs, among others https://www.ldc.upenn.edu/about (accessed Mar. 2018).

[16] European Language Resources Association (ELRA) http://www.elra.info/en/ (accessed Mar. 2018).

[17] Funding for automated language understanding projects with applications to medical diagnosis, among others. https://www.colorado.edu/linguistics/2015/12/07/martha-palmer-and-colleagues-receive-funding-build-language-understanding-systems.

[18] Ethnologue: https://www.ethnologue.com/ (accessed Mar. 2018).

[19] Languages spoken in China: http://no2.mofcom.gov.cn/article/aboutchina/population/200903/20090306117656.shtml (accessed Mar. 2018).

[20] The world factbook https://www.cia.gov/library/publications/the-world-factbook/ (accessed Mar. 2018).

done for other languages, such as Swedish [232, 231], French [71, 196, 190, 192, 193], German [30, 108, 279, 213, 214] and Spanish [53, 36, 197, 198, 165, 61]. They involve corpus annotation, negation and hedge detection and identification of terms and relations of interests, among others. But, although Spanish is the second language in number of native speakers in the world (after Chinese), the third more used in internet and the fourth in available PubMed articles[21] [274],[22] not much work has been done for information extraction in Spanish texts of the biomedical domain. There is still less work done for clinical reports and for the radiology domain. Spanish does not have the previously described difficulties but has more complex morphology and inflection than English. Furthermore, it has much less lexical resources and to the best of our knowledge there are no publicly available corpora for the detection of terms of interest in clinical reports and relations among them. Only recently, in 2017 resources for the assertion detection task have been made available for Spanish [165]. Besides, there are only a few research groups -usually with limited resources-working for this language.[23] Finally, although English is the scientific language by excellence, the language in which the clinical reports are written is not necessarily English, if not the language that is spoken in the region where the medical institution that makes the study is based. Therefore, it is necessary to increase the work done for languages other than English.

1.2 Research questions

The main research questions of our work are:

- how can named entity recognition[24] be handled in informal texts in the medical domain written in medium resource languages,[25] for which there are limited specific vocabularies, less linguistic resources than for high resources languages and where there are scarce annotated datasets?
- is the exact match an appropriate criterion for measuring results in the named entity recognition task in the biomedical domain or are results based on partial match more desirable?
- is NegEx[26] suitable for the detection of negation in Spanish radiology reports? or are there other more sophisticated syntactically based methods, that are more appropriate for negation detection?
- is the knowledge acquired by NegEx implementation for Spanish easily transferable to other Indo-European languages, like German?

[21] PubMed is a search engine of scientific papers in the biomedical domain.

[22] Vítores [274] also states, among others, that the percentage of Spanish speakers is growing, whilst the one of Chinese and English speakers is descending, that 7.7% of Internet users communicate themselves in Spanish -preceded by English and Chinese-, that Spanish Wikipedia is the sixth in number of hourly visits and that although Spanish is relegated in scientific and technical areas, where English predominates, it is the fourth language used in PubMed articles -with 0.8% articles and 92.1% in English-. In addition to English, French -1.2%- and German -1%- precede Spanish.

[23] The research groups in Spain working currently in the area are informed in [7].

[24] Named entity recognition is defined in Chapter 2.

[25] Languages can be classified into three types: high, medium and low resource. High resource languages, like English, have more linguistic resources, such as PoS taggers, dependency parsers, vocabularies of different domains. Low resources languages, like Basque, have less available data to train models and less terminologies available. Medium resource languages are in the middle [230].

[26] NegEx is a negation detection method proposed by Chapman et al. [43], that will be presented in Chapter 5.

1.3 Contributions

As we previously mentioned, not much attention has been given to texts written in Spanish in spite of its relevance and, as for any medium resource language, there are less supporting tools, available data and knowledge sources than for high resources languages. Recently, some more research groups are working in BioNLP in Spanish.

In this work we contribute to the BioNLP research area by providing methods with competitive results to apply an important part of an information extraction pipeline to Spanish radiology reports. Radiology reports are a particular case of medical reports. They are unstructured and have abundance of orthographic and grammatical errors. Reports were provided by a public hospital from Argentina. Contributions are made in text classification, named entity recognition of anatomical entities and clinical findings, and negation detection of clinical findings. In order to have some insights of the applicability of our proposal to other Indo-European languages we applied our implementation of the negation detection algorithm to German text.

In the named entity recognition task, the lack of availability of an annotated corpus and of a specific knowledge source in Spanish guided us to explore initial solutions, that achieved results with lower performance than the existing in other languages. Given the need to have data in order to be able to better test our algorithms and to improve our results and the lack of existence of a corpus -to the best of our knowledge-, a lot of effort has been put by us in order to create an annotated corpus of Spanish radiology reports. Its creation provided us with the possibility to apply other methods, based on annotated data, to solve our needs. The first applied methods are still valid, since they can be used as a guide for the resolution of the named entity recognition problem in cases where there is a reduced number of annotated resources available and there are no specific lexical resources in the same language.

The contributions of this work can be summarized as follows:

- the development of an annotated dataset for Spanish clinical reports, that is useful for named entity recognition of anatomical entities, clinical findings, measures, negation and speculation detection and relation extraction. The dataset corresponds to radiology reports and has an inter annotator agreement (IAA)[27] of 0.89. To the best of our knowledge, there are no publicly available annotated corpora of Spanish clinical reports,
- the development and dissemination of annotation guidelines for clinical reports, which can be used for other annotation projects,
- the implementation of different named entity recognition algorithms to automatically detect *anatomical entities* and *clinical findings* in radiology reports written in Spanish in short texts, with lack of diacritics, scarcity of standardized nomenclatures, abundance of abbreviations, syntactic problems and with the inexistence of specific lexical resources for Spanish,
- the implementation of exact and partial match metrics, that avoid the penalization for retrieving entities with boundary errors, a very common situation in biomedical texts. The precise definition of the partial match metrics allows the comparison with different algorithms,
- the implementation and comparison of different methods of negation detection in radiology reports written in Spanish,
- the implementation of a negation and speculation detection algorithm for two

[27] Inter annotator agreement (IAA) is defined in Chapter 3.

types of German clinical texts (clinical notes and discharge summaries),[28] for which, to the best of our knowledge only one negation detection algorithm implementation existed (tested with a limited set of 12 reports),
- the proposal of a possible simplification of NegEx [43], a well known algorithm, originally developed for the detection of negation in English clinical texts,
- the publication of NegEx triggers for German.[29,30]
- an extensive survey of previous works,
- the implementation of a classification method among medical reports containing affirmed clinical findings and medical records that do not, and finally,
- the report of each step of our IE pipeline, that makes it useful to be reproduced by other researchers for application to medium resource languages.

We plan to make available the Spanish annotated dataset after the thesis dissertation.

The publication of NegEx triggers for German, the implementation of reproducible evaluation metrics and the future publication of the Spanish annotated dataset and Spanish NegEx triggers contributes to the advance in the area of BioNLP research in information extraction (particularly in recognition of named entities, of relations and assertion identification) in informal texts of the clinical domain written in Spanish, given that results will be comparable with other methods tested on the same data and with the same evaluation metrics.

The work required the collection of data of Spanish radiology reports, their selection, normalization and anonymization; the retrieval of lexical data (RadLex and SNOMED CT anatomical entities and clinical findings);[31] the translation of RadLex terms; the annotation of data, that involved the creation of an annotation schema and annotation guideline, to get annotators that meet the requirements needed to work in this complex domain, and the setting and monitoring of a reproducible annotation process; the annotation of a dataset for classification, and for negation detection in German and in Spanish; the analysis and definition of evaluation criteria and the review of differences between English and Spanish, and between both languages and German for the implementation of NegEx in Spanish and in German. It also involved interdisciplinary work and the incorporation of linguistic knowledge. With the work carried out we are able to answer to the research questions that were stated in a previous section of this chapter.

1.4 Thesis Structure

The rest of this document is structured as follows. Part I is closed in Chapter 2, where we present the area of study called BioNLP, some basic definitions, the existing resources, tasks and challenges and previous work in the area.

Part II is devoted to the construction of our solution. Chapter 3 introduces the importance and difficulties of the annotation process, the data used, the annotation schema and guidelines created and an analysis of the resulting annotated dataset. In Chapter 4 we present two methods applied to the detection of entities in Spanish radiology reports and a method for classification of reports among those containing findings and those, that do not contain findings. In Chapter 5 we introduce the negation and speculation detection problem and we present different methods

[28] Different types of medical reports will be described in Chapter 2.

[29] NegEx triggers are a list of negation and speculation terms used by the algorithm. They will be introduced in Chapter 5.

[30] Published NegEx German triggers: http://macss.dfki.de/german_trigger_set.html (accessed Mar. 2018).

[31] RadLex and SNOMED CT are presented in Chapter 2.

applied for negation detection in Spanish radiology reports and negation and speculation detection in German clinical notes and discharge summaries. We also present the annotation for negation carried out for each dataset, an analysis of our datasets, and the results or the implemented algorithms. Each of the chapters of this section includes a review of previous work of the corresponding task focused in the biomedical domain and when possible, in Spanish -or German for the negation and speculation detection task-.

We close this work in Part III. Chapter 6 contains some final words and an outlook of the road ahead. After that, we include the bibliography referenced. Finally, Appendix A lists the main acronyms and abbreviations used throughout this work and their expansions. Appendix B has information that might be of interests, for further understanding of this work, and Appendix C contains definitions of some of the terms that are used throughout the work.

1.5 Dissemination of Results

Abbreviated versions of the results presented in this thesis have been published previously in peer-reviewed publications in [55, 54, 60, 243, 57, 56, 58, 59]. A journalistic dissemination of part of the presented results can be seen in Nexciencia web site.[32]

This chapter and Chapter 2 include publications [54, 55]. Chapter 3 includes following publications [58, 59]. Chapter 4 includes [55]. Chapter 5 contains following publications [243, 57, 56]. A portion of the chapter was also part of Vanesa Stricker master thesis dissertation [242].

1.6 Resumen

En los últimos treinta años la cantidad de textos disponibles digitalmente ha crecido exponencialmente [126, 13]. Esta gran cantidad de información de carácter textual ha promovido la investigación en sistemas de información que faciliten el análisis, la gestión y el acceso a los mismos. La tecnología necesaria se basa en disciplinas tales como el procesamiento del lenguaje natural (NLP) y la extracción de información (IE).

En particular en el área de la biología molecular la gran cantidad de artículos científicos y el aceleramiento del descubrimiento de la información (por ej. con la secuenciación del genoma humano hace unos años) dificulta la posibilidad de mantenerse al día con los cientos de artículos recuperados por una consulta en internet. Sería más conveniente reducir la cantidad de documentos a leer o acceder a una base de datos con información procesada, que permita luego al investigador acceder a los artículos de su interés. Por otro lado, en el área de la medicina, la información producida por los médicos en el formato de historias clínicas electrónicas (EHR), resúmenes de alta hospitalaria e informes radiológicos, entre otros, se está digitalizando constantemente con la adopción de sistemas de información en el dominio médico [26, 32, 112]. En la mayoría de los casos, dicha información se encuentra representada en forma narrativa (o textual), ya que esto proporciona detalles mucho más ricos que el resultado de completar formularios con campos estructurados.

Contar con información en formato estructurado contribuye al aviso automático ante situaciones que requieren acción rápida o inmediata [24, 37, 23]. También es de

[32] Journalistic scientific dissemination of part of our results: http://nexciencia.exactas.uba.ar/hospital-garraham-diagnostico-imagenes-radiologia-computacion-jose-castano-viviana-cotik-dario-fillipo (accessed Mar. 2018).

utilidad para sistemas de soporte a la decisión clínica [72, 239] y para la medicina de precisión, que intenta proporcionar tratamientos a medida y estrategias de prevención de acuerdo con la sintomática, a la genética, el entorno y el estilo de vida del paciente. Finalmente, contar con datos estructurados ayuda en la reducción de costos, a la toma de decisiones y en el aceleramiento de descubrimiento de información implícita y explícita [229, 50]. Por otro lado, es importante determinar qué condiciones clínicas están presentes y cuáles están ausentes o son hipotéticas [173]. Finalmente, la clasificación de informes, entre aquellos que contienen un hallazgo clínico afirmado y aquellos que no lo contienen es útil para recuperar informes con sus correspondientes imágenes y utilizar las mismas con fines educativos.

El dominio biomédico tiene varias dificultades que no aparecen en otros dominios. Cuenta con vocabulario altamente especializado y la jerga que se utiliza es local y puede diferir en distintos países, regiones, instituciones médicas y hasta entre los profesionales de una misma institución. Un mismo concepto puede ser escrito de distintas formas (debido a sinonimia, distintas transliteraciones grecolatinas, variaciones ortográficas y por el uso de abreviaturas no estándar) [13, 148, 90]. Por otro lado, a diferencia de los artículos científicos, algunos tipos de informes médicos se escriben de manera rápida, lo que ocasiona gran cantidad de errores gramaticales y ortográficos. En particular en los informes radiológicos se espera que las oraciones sean cortas y carentes de la estructura sintáctica habitual y que se usen acrónimos [111]. Por último, para poder comparar resultados de distintas propuestas es importante contar con conjuntos de datos disponibles públicamente y estándares de evaluación. La existencia de datos públicos es muy escasa en el dominio médico. Esto se debe a que 1) los datos son sensibles, por la presencia de información personal. La publicación de los mismos requiere su anonimización y depende de la existencia y aplicación de normativas a nivel país, institucional o regional; 2) el proceso de anotación requiere contar con cierto grado de entrenamiento. No es fácil conseguir recursos que lo tengan y que estén disponibles para este tipo de trabajos [244]. Por otro lado, la complejidad de la terminología y las otras características de los textos convierten a la definición de criterios de anotación y a la anotación en una tarea compleja.

BioNLP, el área de estudio que trata con extracción de información de textos biomédicos ha crecido mucho y está muy activa en este momento. Existe gran cantidad de conferencias y talleres, tutoriales y charlas invitadas en las conferencias más importantes de NLP y una importante tradición de competencias en el área. También hay proyectos de colaboración internacional y se está tratando el tema comercialmente. El área se ha desarrollado más para textos científicos y otros textos formales y para el idioma inglés. Poco se ha hecho para el tratamiento de textos médicos en español, pese a que es el segundo idioma más hablado en el mundo y a que los informes médicos en lugares de habla hispana se escriben en español. El español tiene una morfología más rica que el inglés y cuenta con menos recursos para el procesamiento lingüístico de textos. Por otro lado, a nuestro mejor saber y entender no hay corpus de datos disponibles para la detección de entidades de interés y relaciones en textos clínicos.[33]

En el presente trabajo queremos responder las siguientes preguntas: 1) ¿cómo se puede trabajar con reconocimiento de entidades nombradas en textos informales del dominio médico escritos en lenguajes sin muchos recursos disponibles?[34], 2) ¿la coincidencia total es un criterio adecuado para evaluar resultados en la evaluación de

[33] Recientemente, en 2017, se ha puesto a disposición del público un corpus de negaciones [165].

[34] A diferencia del inglés, el español tiene menos recursos lingüísticos, menos terminologías y menos corpus disponibles.

detección de entidades nombradas en el dominio biomédico o es más deseable utilizar un criterio de coincidencia parcial?[35] 3) ¿es posible adecuar NegEx, el algoritmo propuesto por Chapman et˜al. [43] para la detección de negaciones en resúmenes de alta hospitalaria a la detección de negaciones en informes de radiología escritos en español o hay métodos más sofisticados basados en sintaxis que son más adecuados para esta problemática? y 4) ¿el conocimiento adquirido por la implementación de NegEx al español se puede transferir fácilmente a otras lenguas indoeuropeas, como el alemán?

En este trabajo contribuimos al dominio de BioNLP proveyendo métodos con resultados competitivos, para aplicar a un importante subconjunto de un proceso de extracción de información médica de informes radiológicos escritos en español. Los informes fueron provistos por un hospital público de Argentina. Entre las contribuciones del trabajo se encuentran: la creación de un corpus anotado de informes radiológicos (RR) en español, útil para el reconocimiento de entidades anatómicas, hallazgos clínicos, medidas, negación, especulación y extracción de relaciones; el desarrollo y divulgación de lineamientos para realizar anotaciones para informes clínicos, que pueden ser utilizados en futuros proyectos; la implementación de distintos algoritmos para identificar entidades anatómicas y hallazgos clínicos en RR en español; la implementación de métricas con distinto criterio de coincidencia, con el efecto de evitar la penalización al recuperar entidades con coincidencia parcial, una situación habitual en textos del dominio biomédico; la implementación de distintos métodos de detección de negaciones para el español y para el alemán estudiando cuán adecuadas son las distintas técnicas según el tipo de informe médico; la publicación de los *triggers* de NegEx[36] para el alemán y una extensa revisión de los trabajos previos.

Está prevista la publicación de los datos anotados luego de la presentación de la tesis.

[35] Nos referimos a los términos *exact* y *partial match* del inglés.

[36] Los *triggers* de NegEx son una lista de términos utilizados por el algoritmo que indican negación y especulación.

CHAPTER 2

Biomedical Text mining

In this chapter we introduce the research area called biomedical text mining. First, we provide an introduction. Then, we supply some basic definitions. Next, the basic steps involved in a natural language processing workflow and the main information extraction tasks are presented. Finally, we introduce existing resources available for biomedical text mining and evaluation metrics usually used for these tasks.

2.1 Introduction

As already mentioned, in the last thirty years there has been an exponential growth in the amount of digital texts available. Among them, there are different textual genres (e.g. journalistic, scientific, medical reports, and social media messages) and distinct domains (e.g. law, biomedicine, tourism and entertainment). Texts can be of *formal* nature, that is, written in a correct way, usually with lack of orthographic errors and with well-formed sentences (e.g. scientific articles and drug leaflets) or of *informal* nature, which usually contain orthographic errors, lack of punctuation signs and non-standardized abbreviations and vocabulary (e.g. some type of medical reports, SMS and tweets).

Particularly in the medical domain, the amount of digital texts available has also increased steadily in the last years with the penetration of electronic health record systems [26, 32, 112]. The extraction of information from medical reports can serve to support clinical tasks (such as the trigger of automatic alerts and as input for decision support systems) to take decisions (e.g. prevention plans, preparing budgets) and to target which medical records to access based on their content.

Research on BioNLP focuses on texts referring to the biological domain (usually scientific articles and abstracts) and to texts of the clinical or medical domain, that include scientific articles and medical reports, among others. More attention has been provided to formal texts and to the biological domain. We are going to work with informal texts of the medical domain.

There exist different kind of medical reports, such as electronic health reports, discharge summaries and radiology reports. Some of them are semi-structured, others are unstructured, some have many sections, while others only have one, but all of them have abundance of non-standard abbreviations, many of them ambiguous

[204], and specialized language. Also, many negation and speculation terms appear [42, 267, 61]. We are going to work with a particular kind of medical report: radiology reports.

There are no standards referring to how to write a medical report. Particularly, radiology reports have very specialized vocabulary -not only because of belonging to the clinical domain, but also given that they serve as a form of communication between the radiologist and the referring physician (not with the patient)-. They also contain a high number of ill-formed sentences (for example, "no fever"), grammatical and orthographic errors and lack of punctuation signs. There is also a high variation in the way that physicians refer to the same clinical finding [235]. Reports are also short. This is mainly due to the scarce time that physicians have to write and to the guidelines about how to write them.

Existing publications that address how to write radiology reports mention the importance of writing a *good* report and the lack of training in the field. Also, grammar issues, style of writing and use of abbreviations and standardization are addressed [111, 282]. As previously mentioned, guidelines usually recommend being brief, despite writing in a telegraphic way. For example, "Linear atelectasis right lower lobe" is recommended instead of "There is an area of linear atelectasis in the right lower lobe".[1] Also avoidance of over-hedging is recommended. Templates for writing radiology reports are provided by the RSNA (Radiological Society of North America).[2]

Research about the amount of details wanted by physicians who referred patients to the radiology department and about the quality of radiology reports have been done [49, 166, 235]. It is of utmost importance that the communication among the radiologist and the referring physician is clear and that information is received in a timely manner (i.e. in case of an important finding a call or a manual or automatically sent message should be received by the referring physician) [24, 37, 23].

In order to evaluate information extraction results, reports have to be annotated. One of the main difficulties to work with medical reports is that the number of publicly available annotated medical reports is scarce. This situation is mainly due to confidentiality issues, regulated by institutions and by country or regional laws, that enforce to anonymize reports. There are different degrees and kinds of anonymization. The difficulty in obtaining or sharing annotated reports hinders the growth of the area, since different approaches to solve a problem can usually not be compared. A large number of challenges have been created in the biomedical domain, which help obtain annotated datasets and improve the state of the art.

As mentioned in the previous chapter, research in BioNLP is growing steadily. Only in 2016 and 2017 more than 13 events including conferences, workshops, tutorials and invited talks in top NLP conferences have been organized. They are described in Section B.1 of Appendix B. Also, special issues in specialized journals, books and surveys have been published in the last years [11, 50, 72, 293, 291, 41, 229, 124]. In 2009 the International Journal of Medical Informatics published a special issue of BioNLP [138] and in 2016 a special issue on information retrieval in biomedicine [102] appeared in the Journal of Information Retrieval.

[1] Example taken from "Guidelines about how to compose a radiology report" http://www.chestx-ray.com/index.php/practice/how-to-compose-a-radiology-report-guidelines (accessed Jan. 2018).

[2] Templates of reports provided by the Radiological Society of North America https://www.rsna.org/Reporting_Initiative.aspx, http://www.radreport.org/ (both accessed Feb. 2018).

The rest of the chapter is organized as follows. In the next section we will provide some basic definitions. Then, in Section 2.3 we will present the architecture of a text mining system and briefly present each of the steps involved in an NLP workflow and the main IE tasks. We will also present available resources for natural language processing and for other text mining tasks. In Section 2.4 we will mention existing resources available for biomedical text mining. Next, in Section 2.5 we will present some of the evaluation metrics usually used to judge the correctness of system outputs. Finally, we close the chapter in Section 2.6 providing an overview of previous work in biomedical text mining related, but not central to our solutions. Previous work related to corpus annotation, named entity recognition and negation detection will be presented in next chapters. An overview of the challenges in the BioNLP area is also presented. More detail about them is provided in the related chapters.

2.2 Definitions

This section provides an introduction to important concepts relevant throughout this work. First, we will describe artificial intelligence tools for the processing of textual information. Then, terms related to medical informatics. Next, some terms related with linguistics, and finally, other terms of interest. Some other linguistic terms referred in this chapter and throughout the work are defined in Appendix C.

2.2.1 Artificial intelligence tools for the processing of textual information

Natural language processing (NLP)

Natural language is the language written and spoken by people. Natural language processing (NLP) are the computational techniques to process natural language in order to analyze it or to generate it. According to Jurafsky and Martin [137] what differences an NLP application from other data processing systems is that the first has to have knowledge about language. NLP can be used as a component of systems that understand and generate speech (speech processing systems), but this thesis covers only the analysis of written text.[3] There is a wide range of applications of NLP from word counting to question answering systems. Natural language generation, speech recognition, dialogue systems, machine translation, text summarization, author profiling, automatic text correction, automatic prediction of words, opinion mining and sentiment analysis are some of the applications of NLP. Some areas of study of NLP are: text segmentation, word-sense disambiguation, textual entailment, anaphora and coreference resolution.

There are many aspects of language that can be taken into account, among others: morphology, syntax, semantics, pragmatics and discourse [137]. One of the aspects that stands in the way of analyzing natural language is its ambiguity. There exist many kinds of ambiguity, for example *bank* can refer to a financial entity or to the bank of a river (semantic ambiguity) and *play* can be a noun (a theater play) or a verb (syntactic ambiguity).

[3] Texts are also refereed by us as narratives and unstructured texts. Structured data is data stored, managed and queried by a database system. Textual data can be semi-structured or unstructured. Unstructured text refers to free text. Semi-structured texts include unstructured and structured sections.

Information extraction (IE)

Information extraction is the area of research that deals with the extraction of structured information from machine-readable unstructured or semi-structured texts [106, 3]. It can be thought as the task of filing templates, containing slots, that represent semantic information [137]. From these templates databases can be populated and decisions can be taken. The Message Understanding Conferences (MUC) and the Automatic Content Extraction (ACE) program helped advance the IE field. IE is domain dependent. Information extraction tasks include named entity recognition and relation extraction.

Information Retrieval (IR)

Given a document collection and a set of information needs, information retrieval is the action of producing a list of documents matching the information needs. Documents are pre-processed in order to produce term indexes, which contain information about where the terms occur in the collection. Once relevant documents are retrieved, the user has to read them to find the desired information.[4] Web search engines are an example of information retrieval applications. Text retrieval is a branch of information retrieval. The Text Retrieval Conference (TREC) contributed to IR systems development and evaluation. See Section 2.6, in particular Table 2.1 for more information about TREC. IR techniques are domain independent.

Text Mining (TM)

Text mining deals with the analysis of information to discover patterns or knowledge that are not explicitly mentioned in texts. The goal is to be able to analyze information and facilitate decision making [3, 115]. A broader definition of text mining includes any system that extracts information from texts [293].

We are taking the broader definition of text mining, so although we are applying information extraction techniques for named entity recognition, the chapter is called biomedical text mining and the area of study BioNLP.

2.2.2 Medical informatics definitions

We call **clinical finding** to an observation that indicates or might indicate the presence of a disease or health problem. Clinical signs are also considered *clinical findings* by us. For example, adenomegalies, fever, cough, headache, swelling, appendicitis and cyst, are considered findings by us. An observation that does not indicate a problem or a possible problem is not considered a finding by us (e.g. pregnancy). In the case of negation or speculation terms (e.g. no fever), we still consider fever as a finding, but affected by a negation.

We call **anatomical entity** to the anatomical structures, organs of the body, tissues and their components. The *Foundational Model of Anatomy*[5] defines an anatomical entity as an “Organismal continuant entity which is enclosed by the bona fide boundary of an organism or is an attribute of its structural organization”. Some examples are *vein*, *liver*, *finger* and *bone*. At a certain point it is not easy

[4] NLP for information access. Horacio Saggion. Winter School of Computer Science (Escuela de Ciencias Informáticas, ECI) 2008.

[5] Foundational Model of Anatomy https://bioportal.bioontology.org/ontologies/FMA-SUBSET?p=classes (accessed Feb. 2018).

to define what constitutes an anatomical entity and what not [31] (consider, for example, *wall of the liver*), but we stick to the previous definition.

The different kind of reports written by physicians are called **clinical or medical reports**. There are differences among them, some of them have semi-structured format (a part consists in structured data and other fragment in unstructured text), and others have unstructured format. Depending on the institution where they are produced, they might have different sections and be longer or shorter. We provide a brief description of some of them in order to provide tools to understand this work. We add the definition of drug insert, since we are going to refer to it in the rest of the work.

An **electronic health report (EHR)** is the collection of all the health-related information of a patient. Usually each patient has different EHRs in different institutions, but there is a tendency towards the exchange of information from a patient between institutions and countries when needed [34].[6] The implementation of unified EHRs is also being discussed [45, 276, 122].

A **discharge summary (DS)** is a report written at the end of a hospital stay. It includes reason of admission, diagnosis made, laboratory results, therapy undertaken, medication taken and recommendations for follow up. It is incorporated in the EHR.

A **clinical note** is a register of the observations recorded by a physician during a medical consultation. It can include physical findings, results of laboratory tests, conclusions drawn by the physician or group of physicians involved in the treatment of the patient, comments about changes in medication and further action.

A **radiology report (RR)** is a text written by a radiologist after having done an imaging study, such as an MRI (magnetic resonance imaging), CT scan (computer tomography), ultrasound or an X-ray.[7] In the report the physician summarizes what he saw, the interpretation of the findings within the clinical context, possible diagnosis and suggests further steps for a definite diagnosis if needed. Among others, it might include measures of organs, organs seen and impressions.[8] Texts are usually short and with complex vocabulary. For a general description about RRs see [111].

A **progress note** is usually a short and unstructured text explaining the changes of a patient's health during a treatment or hospitalization period.

A **drug insert or drug leaflet** is a document provided along with a medicine in order to give information about the drug, its interactions and secondary effects, among others.

An **anamnesis report** is a text containing information obtained by a physician based on questions asked to the patient.

2.2.3 Linguistic terms

We provide the definition of some terms, that are useful for understanding lexical semantics and language resources that will be introduced later. Other useful terms

[6] Mapping out the obstacles of free movement of electronic health records in the EU in the light of single digital market https://www.eu2017.ee/sites/default/files/inline-files/final_inglk_etervise_uuring.pdf (accessed Mar. 2018).

[7] In the United States of America, ultrasounds are carried out by technicians, but the report is written by a physician. In Argentina and in many other countries of the region, the ultrasound is carried out and written by a physician, except in some cases, where an administrative writes what a physician dictates. In some other countries, the ultrasound is carried out by the physician and the image is interpreted by a technician, that writes the radiology report.

[8] More information about radiology reports can be seen in https://www.radiologyinfo.org/en/info.cfm?pg=article-read-radiology-report (accessed Feb. 2018).

are defined in Appendix C.
A **corpus** (plural *corpora*) is a collection of texts or speech used for a specific purpose, which may be enriched with some type of annotation.
A **gold standard** is a dataset annotated by specialists, that can be used as a reference to evaluate software tools.
A **lexeme** is a minimal unit of meaning, independent of the inflectional endings that words related to it may have.
A **terminology** is the set of terms used in a particular domain.

2.2.4 Other definitions

Abbreviations and acronyms

Abbreviations are sequence of letters used to shorten the representation of a word or of a phrase. For example, Av., cm. and Dr. for avenue, centimeter and Doctor, respectively.

Acronyms are words formed by one or a few initial letters of each word of a compound expression. Some literature makes distinctions among *acronyms*, that are pronounced as a single word, for example NATO or OVNI (in Spanish)[9], and *initialisms*, that are pronounced as individual letters, such as *USA* or *DGI* (in Spanish).[10] We will include initialisms in our acronym definition.

In written text, contractions are shortened versions of a word created by omission of internal letters (for example, *'d* for *would*).

Abbreviation and acronyms are often ambiguous, especially in the medical domain [204]. As mentioned, even the same abbreviation can be used with different meaning by distinct medicine specialties [204] or by different physicians of the same specialty in a hospital. For example, *HTP* can stand for *hipertensión pulmonar -pulmonary hypertension-* and for *hipertensión portal -or portal hypertension-*; Pakhomov et al. [204] mentions that *RA* has following eight senses:[11] *rheumatoid arthritis*, *renal artery*, *right atrium*, *right atrial*, *refractory anemia*, *radioactive*, *right aram* and *rheumatic arthritis*. Abbreviations and acronyms used in informal texts in the medical domain do not always follow a standard (consider for example VPorta, that corresponds to *vena porta* -portal vein- and VN, that refers to *valor normal* -normal value-); furthermore, the same term can be referred to with different abbreviations (consider RD, R.D and RD. and RDER for *riñón derecho -right kidney-*).

Stop words

Stop words are very common words, such as determiners and prepositions, that are considered irrelevant for some tasks, as indexing in information retrieval, and are, therefore, excluded. Some other tasks, such as *author identification* use them. There is no universal list of stop words.

Morphology

In linguistics, morphology studies the words, their structure and the mechanisms of word formation.

[9] NATO: North Atlantic Treaty Organization, OVNI: *objeto volador no identificado* (unidentified flying object).

[10] USA: United States of America, DGI: *Dirección General Impositiva* -tax administration department-.

[11] According to UMLS 2001 version AB.

Words are built up from *morphemes*, the minimal meaningful units of a language. There are two classes of morphemes: *stems* and *affixes*. The stem is the main morpheme of the word. Affixes gives further meaning to the word and can be categorized as *prefixes*, *suffixes*, *infixes* or *interfixes* and *circumfixes*. Prefixes precede the stem, suffixes follow it. Infixes are placed in the middle of the stem and are used to join two words or morphemes and circumfixes precede and follow the stem. For example, the word *snakes* is formed by the stem *snake* and the suffix *s*. The past participle of *spielen* (*to play* in German) is *ge*spiel*t* (ge- and -t form a circumfix) [137]. Morphemes are useful to help in the pronunciation of words, in their understanding and in the detection of out of vocabulary words.

Morphemes can be combined in different ways to create words. Some of the methods to do it are called *derivation*, *inflection* and *compounding*. Inflection is the combination of a word stem with a morpheme, that expresses a grammatical function or attribute as tense, person, number or gender (for example, the inflectional morpheme -s in English is used for obtaining plural nouns). Derivation is the combination of a word stem with a morpheme, in such a way that a different meaning is obtained (for example, the prefix un- negates the rest of the word) or the part of speech is affected (for example, a verb is changed to a noun, as in *mix* and *mixture*).[12] Jurafsky and Martin [137] define compounding as the combination of multiple word stems together. Compounding is very usual in German. More information about inflection, derivation, compounding and word formation can be seen in [137].

For example, in *amorphous*, the morphemes are *a-* (without), *morph* (form) and *-ous*. *morph* is the root, and the other morphemes are derivational affixes. In *organisms*, organ is the root and *-s* is an inflectional morpheme.

n-grams

n-grams are contiguous sequence of n objects (words and letters, among others) taken from a text. The examination of the n-gram frequency of a corpus gives insight about the object's use in the language, which can be used to predict what would follow a sequence. Models that predict the nth word based on the n-1 previous words of the text are called n-gram models. An n-gram of size one is referred to as a *unigram*; of size 2 as a *bigram*, of size 3 as a *trigram*, and so on. For a further reference on n-grams and current applications see Pustejovsky and Stubbs [210].

Inverted index

An *inverted index* is an *index* that maps content to its location in a document, database or set of documents [162].[13] It could for example, map a word to a list that records which multi-word terms the word occurs in.

Knowledge base

A knowledge base is a repository of information.

[12] Both definitions were adapted from Wikipedia `https://en.wikipedia.org/wiki/Inflection`, `https://en.wikipedia.org/wiki/Morphological_derivation` and [137].

[13] Wikipedia definition of inverted index `https://en.wikipedia.org/wiki/Inverted_index`.

2.3 Text Mining

The workflow of a standard information extraction system (also called an information extraction pipeline) is shown in Figure 2.1, but each implementation might involve different processing steps. A natural language processing workflow or pipeline has many steps in common.

An NLP workflow is usually composed by following phases: lexical analysis (segmentation of sentences, tokenization and normalization), morphological analysis (stemming, lemmatization), part of speech tagging, syntactic analysis (syntactic parsing, shallow parsing, dependency tree parsing) and semantic analysis (techniques, that assign meaning to sentences). It will be better defined in Section 2.3.1.

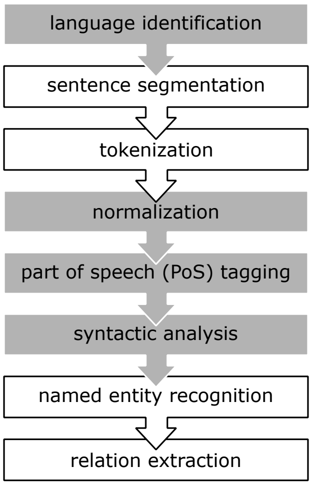


Figure 2.1: Example of an information extraction workflow. Modules in gray are optional.

A standard information extraction system generally involves sentence boundary detection (also called sentence segmentation), tokenization, part of speech (PoS) tagging, syntactic analysis, named entity recognition,[14] negation and speculation detection[15] and relation extraction.[16] Section 2.3.3 describes IE modules not described in Section 2.3.1. NLP and IE might be preceded by a language identification module.

2.3.1 Natural language processing modules

In this section we introduce the NLP modules used or referred by us.

Sentence segmentation is the process of identifying sentences in a text. Usually syntactic, semantic analysis, negation detection and relation extraction are performed sentence by sentence instead of across sentences. Therefore, a sentence-splitting process is needed.

Decisions have to be taken in order to know where a sentence ends. For example, “Seen by Dr. Charles.” is a single sentence, although the English and Spanish

[14] Named entity recognition refers to the identification of instances of a specific type of information units in a text and to its assignment of a class. It will be better defined in Section 2.3.3.

[15] Negation and speculation detection refers to the task of determining if a given finding is under the scope of a negation or a speculation term.

[16] Relation extraction is the action of identifying relations among named entities in a given text (e.g. the date of a marriage or the localization of a finding). It will be better defined in Section 2.3.3.

language rules say that a period followed by a common word (and optionally some blank spaces) might indicate the beginning of a new sentence. In this case it is not, because the period is preceded by an abbreviation ("Dr."). There are also other symbols, such as "?" that can be sometimes used to segment sentences. For example, the talk "How should you write your papers? will be given next Tuesday at 1 pm".

Hand-made rules can be used to do sentence segmentation. The use of a list of known abbreviations can reduce the introduction of sentence splitting mistakes.

Tokenization is the process of identifying the different units, named *tokens*, of an input text. Tokens are not necessarily sequences between blank spaces or punctuation marks. For example: "Mr. White told they've stolen $5,000." should be tokenized as "Mr." , "White", "told", "they", "have", "stolen", "$", "5,000" and ".". A tokenized text is the input of a syntactic analysis and of a PoS tagger.

Tokenization is language-specific. For example, in German compound nouns, such as *Bezirksschornsteinfegermeister*,[17] are used. Therefore, a compound-splitter module that looks for words that can be divided in other words can be used. In Chinese and Japanese, words are not always delimited by ideograms. In Chinese, in some cases, two ideograms might mean two different words and also only one word. For instance, 冰箱 means refrigerator, but 冰 means ice and 箱 means box and 酒店 means hotel, while 酒 means wine and 店 means shop. Other languages, such as Spanish and English, are easier to tokenize, but still have particularities (such as apostrophes for possession -*Muriels'*- and contraction -*won't*- in English and hyphenation -*short-sleeved*- and abbreviations -*Dr.*, *cm.*- for both languages), that make them need carefully designed tokenizers.

An error in the detection of abbreviations can cause a sentence to be separated into two sentences. Dates, units of measurement, chemical formulas tokenization can present difficulties.

Tokenizers are constructed by means of hand-made rules or by machine learning techniques.[18]

Normalization is the process of transforming text to a canonical form, so that their processing is easier and less error prone. Some languages, such as Spanish, have diacritic signs (such as ∼, the dash above the n, as in *niña* -girl-, called *tilde sign*, and accents-). When written above a letter they change its pronunciation. Normalization tasks can involve, among others, diacritic removal, case-folding (reducing all the letters to lower case), contractions, abbreviations and acronyms expansions and conversion of numbers to their word equivalence. Normalization usually improves information extraction and speech synthesis results, but can also lead to errors, for example in word-sense disambiguation and part-of-speech tagging,[19] when diacritics are removed and to information loss if, for instance, some semantics is given to words written in uppercase.

Stemming and lemmatization. The goal of both tasks is to reduce inflectional and derivationally related forms of a word to a common base form [162]. Stemming is the process of eliminating affixes from a word in order to obtain its stem. The

[17] Bezirksschornsteinfegermeister: district master chimney sweeper.

[18] Some references to tokenization can be seen in [162] and in the informal post *The art of tokenization* https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en (accessed Feb. 2018).

[19] Word-sense disambiguation is defined in Appendix C and part-of-speech tagging ins defined in the next paragraphs.

stemming process usually removes the end of the words. For example, the stems of *running* and *considered* are *run* and *consid*. Lemmatizing is the action of obtaining the base or dictionary form of a word, known as *lemma*. For example: the lemma of *did* is *do*. In Spanish the word *satisfaciendo* would be stemmed as *satisfac* and lemmatized as *satisfacer*. In English the word *went* would be stemmed as *went*, and lemmatized as *go*. Stemmers are easier to implement than lemmatizers.

Part of speech (PoS) tagging. Part of speech is a category of words that exhibit a similar behavior, appearing in similar contexts and having similar functions in a sentence, for instance: noun. PoS tagging is the process of assigning a single label of class of part of speech to each token of a text. In many cases there are different tags that can be assigned to a word. For example, they[PR] play[VB,NN] like [IN, VB] children[NN] .[Fp].[20] The fact that some words have more than one PoS tag can be solved with rule-based PoS-tagging, transformation-based tagging or machine learning. These methods take the word context into account.

A POS tagset is a set of labels for part of speech. Some PoS tagsets are the: *Brown Corpus*, that has one million words of texts of different domains and 87 PoS tags [91] and the most used for English, the *Penn Treebank*, a 4.5 million words corpus with texts of the Wall Street Journal, IBM computer manuals and nursing notes, among others and that has 48 PoS tags [164]. For Spanish, the tagset proposed by EAGLES[21] is used.

Syntactic analysis (or parsing) is the analysis of sentence structure according to a pre-defined grammar (given by linguists or by probabilistic parsers), that specifies valid syntactic constructions. Given a sentence, a syntactic analysis validates its correctness and specifies a parse tree for it. Sometimes, complete parsing is very expensive. Alternatively, superficial syntactic analysis techniques, such as shallow parsing and chunking can be used.

One of the main structures used for parsing are dependency trees. Dependency trees are graphs, whose nodes correspond to words and whose edges correspond to dependencies among them. It is constructed based on statistics taken from corpora.

2.3.2 Resources for general NLP

In this section we introduce some language resources for NLP, that are going to be refereed by us throughout the rest of our work. We also introduce some supporting tools for NLP.

Language resources

Some language resources for NLP, that will be used or referred by us are:
Wikipedia[22], a free online encyclopedia, available in different languages; **DBPedia**[23], a knowledge base that contains Wikipedia content in structured format (the structured information is publicly available), and **WordNet**, a lexical database, that has words grouped in set of synonyms (called synsets), short definitions, some sample uses and one level hypernym-hyponym relations.[24] There are semantic relations

[20] IN means preposition, FP means punctuation sign, NN means noun, PR means pronoun and VB means verb.

[21] EAGLES (Expert Advisory Group on Language Engineering Standards) Guidelines http://www.ilc.cnr.it/EAGLES/browse.html (accessed Feb. 2018).

[22] Wikipedia https://www.wikipedia.org/ (accessed Feb. 2018).

[23] DBPedia: http://wiki.dbpedia.org/ (accessed Feb. 2018).

[24] See definitions of hypernym, hyponym and other terms in Appendix C.

between synsets. WordNet can be used for word sense disambiguation, IR, semantic analysis, automatic text classification and automatic translation, among others. There are extensions of WordNet to other languages. One of them, EuroWordNet, includes Spanish.

Other existing language resources include AnCora[25] and pretrained word embeddings.[26]

Supporting tools

We describe briefly some NLP supporting tools, with emphasis put on those working for Spanish.

There are several tools that provide a suite of text processing functionalities, such as tokenization, PoS tagging, parsing, named entity recognition for general domain entities and coreference resolution. Some of them are Stanford CoreNLP [163][27], OpenNLP, Google Cloud Natural Language[28], Freeling [35] and LingPipe[29]. The first four support Spanish. Freeling PoS tagger is based on the proposal of EAGLES.[30,31] NLTK (Natural Language Toolkit)[32] is an open source platform for working in Python with natural language. It provides access to copora, lexical resources and a suite of text processing libraries for NLP tasks, including classification [25]. SpaCy[33] is also an open source platform for NLP, but unlike NLTK it is focused in production environments. It supports deep learning workflows. The Snowball stemmer supports Spanish[34] and is implemented in NLTK.[35]

MATE parser [27] and MALT parser are data-driven dependency parsers. The applied linguistics department of *Universidad Pompeu Fabra*, Barcelona (IULA) has trained both dependency parsers for Spanish [15].[36]

There are some frameworks, that help to implement the different stages of NLP and IE systems. They usually implement the traditional NLP modules (sentences segmentation, tokenization, PoS tagging and general domain named entity recognition among others). They provide a graphical user interface that allows to access and to make the composition of different components. Some of them have corpora annotation functionalities. Examples of these frameworks are GATE (General Architecture for Text Engineering) [64, 65] and UIMA (Unstructured Information Management applications).[37] GATE provides resources for the processing biomedi-

[25] AnCora corpus include Spanish and Catalan corpora, http://clic.ub.edu/corpus/en (accessed Mar. 2018).

[26] Word embeddings constitute a technique that expresses words as vectors of real numbers, in such a way that words with similar meanings have similar vectors. The typical example is the result king-man+women=queen, obtained by adding and subtracting the vectors corresponding to those words.

[27] CoreNLP: https://stanfordnlp.github.io/CoreNLP/ (accessed Feb. 2018).

[28] Google Cloud Natural Language https://cloud.google.com/natural-language/ (accessed Feb. 2018).

[29] Alias-i. 2008. LingPipe 4.1.0. http://alias-i.com/lingpipe (accessed Jan. 2018)

[30] Freeling tagset for Spanish can be seen in following url https://talp-upc.gitbooks.io/freeling-user-manual/content/tagsets/tagset-es.html (accessed Feb. 2018).

[31] There is a Freeling module for the Spanish spoken in Argentina http://habla.dc.uba.ar/gravano/freeling-ar-es.php?lang=esp (accessed Feb. 2018).

[32] http://www.nltk.org/ (accessed Feb. 2018).

[33] SpaCy https://spacy.io/ (accessed Feb. 2018).

[34] The Snowball stemming algorithm for Spanish is described in http://snowball.tartarus.org/algorithms/spanish/stemmer.html (accessed Feb. 2018).

[35] http://www.nltk.org/_modules/nltk/stem/snowball.html (accessed Feb. 2018).

[36] MALT parser trained for Spanish by IULA http://www.iula.upf.edu/recurs01_mpars_uk.htm (accessed Feb. 2018).

[37] Apache UIMA https://uima.apache.org/.

cal documents, by means of populating gazetteers with ontologies of the biomedical domain. ABNER and MetaMap,[38] among other biomedical tools, are available to use from GATE.[39]

Available resources for BioNLP, including corpora, knowledge sources and supporting tools, are presented in Section 2.4.

2.3.3 Information extraction tasks

In this section we introduced the information extraction task mentioned throughout the work.

Named Entity Recognition (NER)

The goal of named entity recognition is to identify instances of a specific type of information units in a text and assign them a class. The discovery of those information units and their class is called *named entity recognition* (NER) or *named entity recognition and classification* (NERC). Besides having as output the predicted type of an entity, a score designating the confidence of it belonging to this entity type could be returned. Sometimes, entity normalization (i.e. linking the entity to an entry of an ontology, terminology or database, for example, UMLS)[40] is included in the NER task (this is called *entity normalization*, *entity linking* (EL) or *entity disambiguation*). Other task related to NER, *slot filling*, is about filling attributes of an entity with values found in a text.

Some examples of named entity recognition are to identify gene names within a collection of MEDLINE abstracts, names of people and organizations in newspaper articles or names of clinical findings in clinical reports. If, for example, the UMLS CUI[41] is assigned to the findings, we would be talking of the normalization task.

An entity mention is a textual reference to an entity. Entities can be referenced by name, pronoun or nominally.[42]

Once entities are recognized, they can be linked to external resources and also relation extraction can be performed.

More details about named entity recognition, its challenges in the biomedical domain and in Spanish will be given in Chapter 4, which is entirely dedicated to this problem.

Relation extraction

The goal of relation extraction is to detect a specific type of relation among entities. Relations can be binary or n-ary with n greater than 2. Examples are interaction among drugs (drug-drug interaction -DDI-), the marriage of two people or the location of a medical finding. An example of a ternary relation could be the two members of a married couple, and the date when they married. The location would add a fourth component to the relationship. See [50] for more detail about relation extraction in the biomedical domain. Named entity recognition is a previous step needed for relation extraction.

[38] ABNER and MetaMap are described in Section 2.4.3.

[39] GATE biomedical specific resources: https://gate.ac.uk/sale/tao/splitch16.html#x21-41000016.1.

[40] UMLS is a resource that contains biomedical vocabularies, software tools and standards and will be further defined in Section 2.4.2.

[41] UMLS CUI is a UMLS concept identifier and will be futher defined in Section 2.4.2.

[42] Definition of entity mention of the ACE 2003 EDT Task guidelines https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/edt-guidelines-v2.5.pdf (accessed Feb. 2018).

2.3.4 Other text mining tasks

Text classification

Text classification is the task of classifying texts in different classes. For example, in a newspaper article if it has left or right political tendency or in a medical report if it has clinical findings or it does not.

Terminology extraction (synonym, acronym and abbreviation extraction)

Terminology extraction in a domain deals with the extraction of concepts of a text belonging to a particular domain. In the biomedical domain multiple entities can have alternative names (synonyms). Also, terms can be expressed as acronyms or as abbreviations. A mapping of all synonyms and abbreviations of an entity to a single concept improves the results of the information extraction tasks [50].

Other tasks

Other text mining tasks applied to the biomedical domain include summarization, question answering, event extraction, hypothesis generation and medical and biological literature-based discovery.

The summarization task goal consists in making a summary of a text. It could, for example, be a summary of a paper or of a drug package leaflet. The question answering task goal is to be able to answer a question with sentences instead of with documents. Hypothesis generation deals with the discovery of relations that are inferable but not explicit in texts. For more detail about these tasks in the biomedical domain see [50, 229].

2.4 Resources for biomedical text mining

In this section we will present existing resources for biomedical text mining. We will focus on resources of the medical domain, but also mention some available resources of the biological domain. We will present existing corpora, knowledge sources and supporting tools focusing on those related to the present work.

2.4.1 Corpora

Unless otherwise stated, the mentioned corpora are for English language.

MIMIC II and **MIMIC III**, its update, provide de-identified health data of intensive care patients, including medical reports, physiologic and vital signs data obtained from patient monitors, among others. For more information see [135].[43]

The **ShARe corpus** contains de-identified clinical reports from MIMIC II database version 2.5. It contains discharge summaries, electrocardiogram, echocardiograms, and radiology reports.

The **QUAERO corpus** is comprised of documents with information about commercialized drugs of the European Medical Agency (EMEA), titles of articles indexed in MEDLINE and patents registered with the European Patent Office (EPO). Documents are written in French, but are available also in English (MEDLINE), English and German (EPO) and other European languages, including Spanish (EMEA). Concepts were annotated based on a subset of UMLS, including anatomical entities and disorders [196]. The corpus was originally developed as a resource for named entity recognition and normalization.

[43] MIMIC https://mimic.physionet.org/ (accessed Feb. 2018).

Mantra [144] is a multilingual parallel corpus comprised of titles of papers indexed by MEDLINE, by EPO patents and by documents about commercialized drugs of EMEA in English, French, German, Spanish and Dutch. Texts in the QUAERO corpus constitute a subset of the Mantra corpus. Biomedical concepts were annotated based on a subset of UMLS including anatomical entities and disorders.

MEDLINE is a database, that contains bibliographic information, such as title, authors, abstracts and journal of scientific articles in the biomedical domain. It is maintained by the United States National Library of Medicine (NLM).[44] The 2017 MEDLINE contains over 24 million references published from 1946 to the present in over 5,600 journals worldwide in about 60 languages (92.1% of the references are in English). MEDLINE is freely available. MEDLINE records are indexed with NLM Medical Subject Headings (**MeSH**).[45] **PubMed** is a search engine for texts in MEDLINE, online books and journals in the biomedical domain. The result of a search is a list of citations (authors, titles, source and abstract) to journal articles, and, if available, a link to the full-text. PubMED searches MeSH terms. **PubMed Central** (PMC) is an open access repository for peer reviewed accepted papers.

Some subsets of MEDLINE have been annotated and are distributed to the community through shared tasks. One of them composes the **GENIA corpus** (MEDLINE abstracts, related to MeSH terms *Human*, *Blood Cells* and *Transcription Factors*, annotated for part of speech, syntax, coreference, named entities, events, among others) [254]. Another is **BioText**, composed, among others, of a reduced set of MEDLINE titles and abstracts labeled for diseases and treatments and relations among them -used in [216]-[46], a reduced set of MEDLINE abstracts with annotated abbreviations that was used in [223], and another dataset labeled for protein-protein interactions, used in [215]. **BioCreative Gene II** tasks data consists of MEDLINE abstracts annotated for gene names and related entities [6, 234].

Other public datasets are **TREC Genomics Track dataset** (full-text articles from genomics-related journals, some of them annotated for different tasks), **ImageCLEF evaluations corpora** (texts of patients' case descriptions), and the **BioScope corpus**, composed of medical reports, abstracts and full papers annotated for uncertainty, negation and their scopes [267].

BIREME *Centro Latinoamericano y del Caribe de Información en Ciencias de la Salud*[47] (Latin American and Caribbean Center on Health Sciences) has the goal of facilitating the access of medical literature produced in Latin America and the Caribbean and provides abstracts of scientific articles of the biomedical domain, some of them written in Spanish.

SciELO (Scientific Electronic Library Online) is a virtual platform, that provides open access to scientific and technical electronic journals in Latin American and Caribbean countries [202].

Table 2.1 presents a summary of the available corpora, that we are going to mention in the rest of the work.

2.4.2 Knowledge sources

There exist various terminologies that serve as a basis to detect relevant terms in medical reports, to annotate data and for data integration. The main knowledge

[44] NLM `https://www.nlm.nih.gov/about/index.html` (accessed Feb. 2018).

[45] Medical Subject Headings are introduced in Section 2.4.2.

[46] BioText labeled data `http://biotext.berkeley.edu/data/dis_treat_data.html` (accessed Feb. 2018).

[47] BIREME was formerly called *Biblioteca Regional de Medicina*.

name	description	genre	domain
BioCreAtiVe I & II (GENETAG) [234]	MEDLINE abstracts annotated for gene and protein NER	AB	biomedical
BioScope [267]	annotations for uncertainty, negation and their scopes	MR, AB, FT	biomedical
BioText	subsets of MEDLINE entries annotated for diseases and treatments and relations among them, for abbreviations, and for protein-protein interactions	AB, TI	biomedical
BIREME[a]	abstracts of scientific articles of Latin America and the Caribbean of health sciences	AB	biomedical
GENIA [140]	MEDLINE abstracts retrieved using MeSH terms "human", "blood cells" and "transcription factors".	AB	biological
IMAGE Clef	patient case descriptions	DO	medical
Mantra [144]	information about commercialized drugs of EMEA, titles of MEDLINE articles and patents registered in EPO	FT, AB	medical
MEDLINE[b]	database with bibliographic references to journal articles in the life sciences with a concentration in biomedicine[c]	AB, TI	biomedical
QUAERO Corpus [196]	information about commercialized drugs of EMEA, titles of MEDLINE articles and patents registered in EPO. Subset of Mantra corpus.	FT, AB	medical
ShArE Corpus	de-identified medical reports from MIMIC II	DS, EC, ECC, RR	medical
TREC Genomics Track dataset[d]	articles from genomics-related journals	FT	biological

[a] BIREME: http://www.paho.org/bireme (accessed Feb. 2018).
[b] MEDLINE https://www.nlm.nih.gov/pubs/factsheets/medline.html (accessed Feb. 2018).
[c] Definition taken from https://www.nlm.nih.gov/pubs/factsheets/medline.html (accessed Feb. 2018).
[d] TREC Genomics Track dataset http://trec.nist.gov/data.html (accessed Feb. 2018).

Table 2.1: Summary of public datasets of the biomedical domain. References: AB: abstracts, DO: formal documents, DS: discharge summary, EC: electrocardiogram, ECC: echocardiogram, FT: full text scientific articles, MR: medical reports, RR: radiology reports and TI: paper titles.

sources of the biomedical domain can be seen in Table 2.2.

The **Unified Medical Language System (UMLS)** [155] is a resource maintained by the NLM, that contains biomedical vocabularies, software tools and standards to enable data integration. It has among its tools the *Metathesaurus*, that contains terms and codes from nearly 200 dictionaries, terminologies and ontologies. Some of them are ICD-10, LOINC©, MeSH©and SNOMED CT©. It also has a *Semantic Network*, to represent relations among Metathesaurus concepts and lexical tools that can be used for natural language processing. UMLS concepts have unique identifiers called CUI (concept unique identifiers). One concept might have different terms in the different UMLS vocabularies. The CUI is a unique code that identifies all the terms of a same concept of the different vocabularies.

The **ICD-10** (International Statistical Classification of Diseases and Related Health Problems 10th Revision),[48] is a standard diagnostic terminology for epi-

[48] ICD-10 http://apps.who.int/classifications/icd10/browse/2016/en (accessed Feb.

demiology, health management and clinical purposes.

SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms)[49] is a multilingual controlled vocabulary of clinical terminology, that allows the standardization of terms. It contains more than 300,000 concepts organized into 19 hierarchies. Some of them are *body structure*, *clinical findings* and *substance*. Concepts are identified by a unique id. Each concept has a description and may have synonym concepts associated and relations to other concepts (for example, *is a* represents the hierarchical structure and *has a* and *finding site* represent causative and location relations). For a description of SNOMED CT see [240] and for some reviews of its use see [152, 153].

MeSH (Medical Subject Headings) is a controlled vocabulary thesaurus, used, among others to index MEDLINE articles. MeSH terms are added to bibliographic citations during the indexing of MEDLINE and for the description of books and other documents acquired by NLM.

LOINC (Logical Observation Identifiers Names and Codes)[50] is a vocabulary of medical laboratory observations and nursing diagnosis and interventions, among others. It was developed by the Regenstrief Institute and is publicly available.

RadLex (Radiology Lexicon) is an ontology produced by the Radiological Society of North America (RSNA)[51] specific to the radiology domain and written in English. It has been specifically developed to satisfy standardized indexing and retrieval of radiology information. It satisfies the needs in this domain by adopting features of existing terminology systems as well as producing new terms to fill critical gaps. It unifies and supplements other lexicons while it also has mappings to them. It has over 75,000 terms, classified -among others- in *imaging modality*, *procedure*, *object*, *imaging observation*, *non-anatomical substance*, *anatomical entity* and *clinical finding* (see RadLex Tree Browser).[52]

The **WHO-ART** (WHO Adverse Drug Reaction Terminology) is an adverse drug reactions terminology included in UMLS.[53]

SNOMED CT, MeSH, UMLS, ICD-10 and WHO-ART are available in Spanish. There is an Argentine Spanish edition of LOINC.[54] RadLex is only available in English and in German. As far as we know, there is no complete RadLex translation to Spanish.

The US National Center for Biomedical Ontology (NCBO) maintains ontologies that can be accessed and used through BioPortal.[55] As far as we know, there are no Spanish ontologies available. BioPortal also provides the *Annotator Tool*, that looks for terms in ontologies by doing an exact string match.[56] For more information about BioPortal see [195]. The British National Center for Text Mining (NaCTeM)[57] also provides resources for biomedical text mining.

2018).

[49] SNOMED CT http://www.ihtsdo.org/snomed-ct (accessed Feb. 2018).

[50] LOINC https://loinc.org/ (accessed Mar. 2018).

[51] RSNA: Radiological Society of North America, http://www.rsna.org/ (accessed Feb. 2018).

[52] RadLex, Radiology Lexicon: http://rsna.org/RadLex.aspx (accessed Feb. 2018).

[53] WHO adverse drug reaction terminology https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/WHO/index.html (accessed Feb. 2018).

[54] Argentine Spanish LOINC edition https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/LNC-ES-AR/ (accessed Feb. 2018).

[55] BioPortal https://bioportal.bioontology.org/ (accessed Feb. 2018).

[56] BioPortal Annotator https://bioportal.bioontology.org/help?pop=true#Annotator_Tab (accessed Feb. 2018).

[57] British National Center of Text Mining http://www.nactem.ac.uk/(accessed Jun. 2017).

name	description	provider
ICD-10[a]	diagnostic terminology for epidemiology, health management and clinical purposes	WHO
LOINC[b]	vocabulary of medical laboratory observations and nursing diagnosis and interventions, among others	Regenstrief Institute
MeSH[c]	controlled vocabulary thesaurus, used among others to index MEDLINE articles and to describe book contents	NLM
NCBO[d]	maintains ontologies that can be accessed and used through BioPortal	NCBO
RadLex[e]	English ontology specific of the radiology domain	RSNA
SNOMED CT [240]	multilingual controlled vocabulary of clinical terminology. Concepts have synonyms and relations	SNOMED International
UMLS [155]	nearly 200 dictionaries, terminologies and ontologies	NLM

[a] ICD-10 http://apps.who.int/classifications/icd10/browse/2010/en (accessed Jan. 2018).
[b] LOINC https://loinc.org/ (accessed Mar. 2018).
[c] MeSH Browser https://meshb.nlm.nih.gov/search (accessed Feb. 2018).
[d] NCBO BioPortalhttps://bioportal.bioontology.org/ (accessed Mar. 2018).
[e] RadLex http://www.radlex.org/ (accessed Mar. 2018).

Table 2.2: Summary of knowledge sources of the biomedical domain.

2.4.3 Supporting tools

An overview of some biomedical text mining supporting tools is presented next. For some of them more detail is given in the subsequent chapters.

BeCalm (Biomedical Annotation Metaserver)[58] is a platform that offers visualization of biomedical texts annotations and benchmark possibilities of biomedical entity recognition systems. It lists and describes NER resources in the biomedical domain, such as ABNER, BANNER and MetaMap.[59]

MetaMap is an open source program that maps biomedical text to concepts in the UMLS Metathesaurus (exact and partial matches are provided). It supports information retrieval, text mining, literature-based discovery, document indexing, classification and question answering. MetaMap includes a NegEx [43] implementation, an abbreviation and expansion module (based on [223]), an acronym disambiguation module and a word sense disambiguation functionality. It also provides an API (application programming interface) [16]. MetaMap is not appropriate for real time processing. A new version, MetaMap Lite, is available since 2017 and is promoted as less rigorous but much faster than MetaMap [73]. MetaMap works with English texts. A work in progress to translate Spanish text to English in order to be available to use MetaMap for Spanish has been published in 2008 [36].

MedLEE (Medical Language Extraction and Encoding System) [94, 93, 95] extracts information from medical reports. It was designed for chest radiology reports and later extended to other radiology reports and to discharge summaries. It includes a functionality for asserting the relation between a disease or disorder, a sign or symptom, or a procedure and an anatomical site.

ABNER (A Biomedical Named Entity Recognizer) [224, 225] and **BANNER**[60]

[58] BeCalm: http://www.becalm.eu/ (accessed Feb. 2018).
[59] BeCalm NER resources: http://www.becalm.eu/NerResources (accessed Feb. 2018).
[60] BANNER: http://banner.sourceforge.net/ (accessed Feb 2018).

[150] are named entity recognition systems for biological entities based on conditional random fields.[61] ABNER looks for genes, proteins, cell types, RNA and DNA. BANNER was primarily thought for biomedical text but is designed for domain independence. It provides an application programming interface, that allows users to incorporate ABNER into other systems.

cTAKES$^{\text{TM}}$ (Clinical Text Analysis and Knowledge Extraction System) [222] is a system based on NLP for information extraction from medical reports. It is built on UIMA and on OpenNLP and contains modules of sentence segmentation, tokenization, PoS tagging, named entity recognition and normalization, shallow parsing and negation detection. **LingPipe** includes a set of tools for processing MEDLINE data. **JULIE lab** at Jena University offers NLP tools suitable for processing biological text data, that include semantic search and information extraction functionalities, among others.[62]

Freeling-Med [197, 103] is an adaptation of the lexica of Freeling linguistic analyzer with medical terms extracted from Spanish dictionaries and ontologies. The medical resources used to enhance Freeling are medical abbreviations and acronyms for Spanish [147], SNOMED CT terms,[63] a database with names of drugs commercialized in Spain and ICD-9.[64] References are added to help decisions at a semantic level in case of ambiguity. To the best of our knowledge FreelingMed is not publicly available.

PROSA-MED[65] is a project carried out by a consortium of Spanish universities and institutions of the sanitary field in order to extract information from medical reports. Its goal is to perform NER of drugs and diseases and to extract adverse drug interaction relations in Spanish, Catalan and Basque texts.

Savova et al. [222] describe other supporting tools for biomedical text mining. **ORBIT** (Online Registry of Biomedical Informatics Tools)[66] maintains a registry of corpora, knowledge bases and software for BioNLP.

2.5 Evaluation metrics

To judge the correctness of system outputs, they must be compared with the *ground truth*. The ground truth is created by manually generating the appropriate annotations in a test set. Due to human factors and to the vagueness and ambiguity of texts, ground truth is not necessarily *the truth*.

The most usual metrics of performance applied to information extraction are called *precision*, *recall* and *F1*. These metrics are calculated based on the number of true positives (TP: instances correctly classified as positive), false positives (FP: instances wrongly labeled as positive) and false negatives (FN: instances wrongly identified as negative) results. True negatives (TN: instances correctly classified as negative) are taken into account for other metrics, such as *accuracy*. For an explanation of these values see Table 2.3.

[61] Conditional random fields (CRF) are introduced in Section 4.4.4.

[62] `http://julielab.github.io/` (accessed Feb. 2018).

[63] SNOMED CT terms added are from October 2011 edition. The corresponding SNOMED CT and UMLS identifiers were also included.

[64] ICD-9. International Statistical Classification of Diseases 9th edition.

[65] PROSA-MED `http://ixa2.si.ehu.eus/prosamed/` (accessed Mar. 2018).

[66] ORBIT `https://orbit.nlm.nih.gov//` (accessed Mar. 2018).

		actual values (GS)	
		positive	**negative**
prediction	**positive**	TP	FP
	negative	FN	TN

Table 2.3: Confusion matrix. GS refers to gold standard.

Precision: of the instances classified by the system as positive the fraction that is actually positive.

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

Recall: of the positive instances, how many were classified by the system as positive. It is also called *coverage*.

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

Precision can also be defined as "how useful the search results are" and recall as "how complete the results are".[67]

Accuracy: fraction of correctly classified instances over the total number of instances.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

F-measure: The F-measure is the harmonic mean of precision and recall. When β=1 in Formula 2.4, precision and recall are given the same weight and the measure is called F1 (Formula 2.5).

$$F\beta = (1 + \beta^2)\frac{precision * recall}{\beta^2 * precision + recall} \tag{2.4}$$

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{2.5}$$

There is a trade-off between precision and recall. If more instances are retrieved the recall will be higher but a higher percentage of false positives will be retrieved. Depending on the user needs or on the system application, higher precision or higher recall might be preferred. For example, in the case of a very contagious disease, it might be more important to know all the possible infected people of a group and then double check if they really have the illness (higher recall is desired). On the other hand, in the case of a pregnancy test, it is more important that if a woman is detected as pregnant she really is (higher precision).

2.6 Previous work

In this section we will introduce previous works related to BioNLP. The goal is to present other areas of work in the same domain, some of which could be considered for future work. We will describe previous surveys in the area and preceding works

[67] Definitions taken from Wikipedia, https://en.wikipedia.org/wiki/Precision_and_recall (accessed Nov. 2017).

in acronyms and abbreviation identification. Next, we will present a review of the anonymization (also called de-identification) task. Afterwards, we will present some research published recently on medical records spelling correction and punctuation restoration and a summary of other research related to the field. Then, we introduce challenges related with our work but in the general domain, with emphasis put in non-English texts and finally, we introduce challenges in the BioNLP domain.

Related work specific to annotation, to named entity recognition and to negation and speculation detection will be presented in Chapters 3, 4 and 5 respectively. Works and challenges devoted to more than one of those areas will be briefly presented in this section.

2.6.1 Previous surveys

Cohen and Hersh [50] make a survey of the state of the art of biomedical text mining from ca. 2003 until ca. 2005. They present the areas of research of the area in 2005: NER, document classification, terminology extraction, relation extraction and hypothesis generation and propose that the major challenge of researchers in the BioNLP domain until ca. 2015 was to make text mining systems useful for biomedical researchers. They explain the growth of the scientific information available, putting as example the growth of MEDLINE.

Zweigenbaum et al. [293] define concepts and areas or work of BioNLP and make a survey of the state of the art of from 2005 until 2007 with focus on biological literature.

Demner-Fushman et al. [72] present the state of the art of clinical NLP in 2009 and how it contributes to clinical decision support systems by extracting facts.

Zweigenbaum and Demner-Fushman [291] wrote a book chapter presenting methods to help researchers access the contents of biomedical literature. It defines terms of the area and existing tools at the moment (2009).

Chapman and Cohen [41] wrote the editorial of a special issue about research topics in biomedical text mining and natural language processing in 2009. They describe the areas of interest, based on the submissions and highlight the high number of submissions received.

Simpson and Demner-Fushman [229] provide an overview of the state of the art in year 2012 in text mining in the biomedical domain with an emphasis in the resources and tools available to biomedical researchers and in the major text mining tasks of interest to the community, that comprise the recognition of facts from biomedical literature (explicit or implicit), document summarization and question answering.

Huang and Lu [124] review the different community challenge evaluations held in biomedical text mining between 2002 and 2014. They present graphics showing the main biological and medical problems, the organization of challenges through the years and their subtasks.

2.6.2 Acronyms and abbreviations

In this section we will make reference to abbreviations, although we are referring to abbreviations and acronyms.

The detection of abbreviations and their expansions, and the normalization of abbreviations is an important constituent in named entity recognition and word sense disambiguation tasks. Unlike in biomedical literature, in medical reports abbreviations are commonly used without any reference to their definitions or long forms. This, added to their ambiguity and to the lack of naming conventions or the avoidance to stick to them, makes the problem of adequately interpreting abbreviations more challenging.

In the problem of extraction of biomedical abbreviations and their corresponding definition, usually methods rely on the proximity of short forms (abbreviations) and their expansion and in the fact that abbreviations are written between parenthesis [50, 81]. For example, in the sentence “The cardiologist assessed what the risk would be if each patient used an oestrogen containing contraceptive pill (OCP)”, *OCP* is the abbreviation or short form and *oestrogen containing contraceptive pill* is its definition or long form. The short form (SF) -i.e. the abbreviation- might precede or follow the long form (LF). Furthermore, abbreviations can be formed in different ways. Some examples can be seen below:

pattern	**example**
-LF (SF)	Peripheral myelin protein 22 **(PMP22)**
-(SF) LF	**(SNURF)** small nuclear RING finger protein
-SF composed of different words	capillary zone electrophoresis **(CZE)**
-SF composed by only one word	lipopolysaccharide **(LPS)**
-SF composed by different words separated by -	non-rapid-eye-movement **(NREM)**

Cohen and Hersh [50] report in 2005 the abbreviation definition problem as close to be a solved problem. There are two main approaches for identifying abbreviation definitions: rule-based and machine learning methods. Below are some implementations. ALICE [110] uses heuristic pattern-matching rules, BIOADI [145] uses different machine learning approaches, AB3P [236] develops a machine learning algorithm that does not require to manually label data and J. Pustejovsky and Morrell strategy [129] incorporates natural language processing techniques, such as POS tagging and shallow parsing into the acronym recognition algorithm. The copora used by the supervised learning methods is taken from MEDLINE. Other strategies for identifying abbreviations definitions or mapping abbreviations to full forms can be seen in [157, 223, 289, 287].

Regarding abbreviations that are used without mentions to their definition, efforts are done in creating abbreviations databases and in normalizing short forms, by disambiguating them. Therefore, usually the context has to be taken into account.

The *ShARe/CLEF eHealth* challenge was organized in 2013 in order to normalize short forms with the goal of improving patients understanding of reports [184]. Pakhomov et al. work on abbreviation normalization [203] and on abbreviation disambiguation in clinical notes, based on their context [204].

Xu et al. [286] describe how they built a clinical abbreviation database containing abbreviations and expansions proceeding from UMLS and ADAM, a MEDLINE abbreviation database. Moon et al. [172] created an inventory of abbreviations and acronyms occurring frequently in clinical notes and mapped them to long forms of UMLS and to medical abbreviations dictionaries. These abbreviations databases are in English. There also exist some compilations of Spanish medical abbreviations and acronyms [147].[68]

[68] Compilation of medical acronyms and abbreviations from the National Academy of Medicine of Colombia http://dic.idiomamedico.net/Siglas_y_abreviaturas (accessed Mar. 2018).

2.6.3 Anonymization or de-identification

In the clinical domain, anonymization or de-identification is the process of removing from medical records all information that could identify a patient. In some cases, names and patients ids (identification codes from a knowledge base) have to be removed, in others even the diseases have to be changed by others, since otherwise, they could help identify the patient.[69] Also identifiers of the physicians that made the studies have to be removed usually. Institutions and countries might have data sharing policies and legislation, stating to which degree information has to be anonymized in order to be shared. For that reason, de-identification is a very important task in BioNLP. US HIPAA (Health Insurance Portability Accountability Act), the European Data Protection Directive 95/46/EC and Argentine laws 26,529, 17,132 and 25,326, mentioned in previous chapter, are examples of data sharing policies. In many cases, it is not possible to share the medical records, even if they have been anonymized, for instance, when patient consent is required.

Many factors have to be taken into account when anonymizing medical records. See [100, 85] to read about perturbative and non-perturbative methods.

Emam [84] wrote a technical report telling how patient health data was being anonymized in Canada in year 2006 and discusses the adequacy of this practices. Some de-identification challenges have been organized, for example i2b2 *2006 De-identification and Smoking* and *2014 De-identification and Heart Disease Risk Factors* challenges. The annotation process performed for the latter is explained in [245]. Gkoulalas-Divanis and Loukides [100] describe algorithms to anonymize while preserving patient demographics (non-perturbative anonymization) and to anonymize diagnosis codes. A tutorial on *Privacy Challenges and Solutions for Medical Data Sharing* was organized by IBM Research Zurich and Cardiff University in 2011.[70] Gkoulalas-Divanis et al. [101] present a survey of more than 45 algorithms that have been proposed for publishing data of EHRs preserving patient's privacy. Emam et al. [85] mention the ambiguity of the concept of anonymous data, three forms of sharing data (*public*, *quasi-public*, *non-public*) and different ways of data perturbation with de-identification goals and their effects with regards to the possibility of a meaningful analysis. Many other analysis and surveys on the subject [258, 278, 161] and anonymization implementations [250, 259, 265, 66, 98] have been published.

2.6.4 Text correction

Some works about medical reports pre-processing have been published recently. They include spelling correction [148, 90] and punctuation restoration [217], initially thought for dictation transcription of medical reports.

2.6.5 Summary of other works in the area

A brief mention of previous work for other subjects is presented next.

A survey of coreference resolution for EHR is presented in [262].

Some publications that consider both, image and text retrieval, can be seen in CLEF 2012 proceedings.[71] Their main focus is image retrieval based on visual and textual information [288, 186, 48, 33, 74, 74].

[69] See European Data Protection Directive 95/46/EC, article 2a http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A31995L0046 (accessed Feb. 2018).

[70] Tutorial on Privacy Challenges and Solutions for Medical Data Sharing. Slides. https://www.zurich.ibm.com/medical-privacy-tutorial/ (accessed Jan. 2018).

[71] CLEF 2012 proceedings. http://www.imageclef.org/publications#proceedings (accessed June 2017).

Another active area of research is the enhancement of public health based on social media data. For example, an application of information extraction from informal texts is presented in [143]. In this work an annotated corpus, that can be used to check if the mention of a medication in a tweet indicates that the writer has taken it, is described.

Goldstein and Uzuner [104] automatically classify diseases based on discharge summaries. Therefore, they check if the disease is mentioned or not, and if its mentioned if it is under the scope of a negation or speculation term.

Faessler and Hahn [88] recently presented SEMEDICO, a semantic search engine for biomedical literature search. They work with abbreviations, relations -whose retrieval is ranked according to the degree of factuality mentioned in the text-, and events.[72]

2.6.6 Challenges in the general domain mainly for languages other than English

We present here some BioNLP challenges in the general domain putting emphasis in languages other than English.

MET-1 (1996), MET-2 (1997), CONLL-2002 and CONLL-2003 organized, among others, NER tasks in languages different than English. In all cases named entities were names of persons (PER), geographical locations (LOC) and organizations (ORG). CONLL-2002 and CONLL-2003 also included a miscellaneous (MISC) class, that considered entities not belonging to any of the before mentioned entity types, and MET-1 and MET-2 included dates, times and numbers. Handled texts were newspaper articles. More information can be seen in Section 4.3.

The Text Analysis Conference (TAC) (2008-ongoing)[73] provides challenges in order to evaluate NLP and IR tasks. Some of TAC tasks were: *Cold Start Knowledge Base Population (KBP)*, whose goal is NER, slot filling and entity linking from PER, ORG and LOC; *Tri-lingual Entity Discovery and Linking (EDL)*, that aims to do NER and entity linking from text written in English, Chinese and Spanish, and *Extraction from Drug labels*.[74]

Other challenges that organized named entity recognition tasks in the general domain were MUC (Message Understanding Conference) [47, 107, 46], ACE (Automatic Content Extraction Program)[75], HAREM for Portuguese [92] and IREX (Information Retrieval and Extraction Exercise for Japanese)[76].

2.6.7 BioNLP Challenges

We present in this section a review of challenges that have been created in the BioNLP domain. Usually, their aim is to improve the state of the art of a given biomedical information extraction task. Texts genres considered include scientific literature, drug leaflets and different kinds of clinical reports (e.g. radiology reports and discharge summaries). The outputs of the proposed systems are evaluated using pre-defined evaluation metrics. Besides, in some cases, datasets are published also for non-participants. This data is useful for the development and evaluation of biomedical text mining systems.

[72] SEMEDICO http://semedico.org/ (accessed Feb. 2018).

[73] TAC: https://tac.nist.gov/about/index.html (accessed Feb. 2018).

[74] TAC 2017 - Adverse Drug Reaction Extraction from Drug Labels https://bionlp.nlm.nih.gov/tac2017adversereactions/ (accessed Feb. 2018).

[75] ACE https://www.ldc.upenn.edu/collaborations/past-projects/ace (accessed Feb. 2018).

[76] IREX Program http://nlp.cs.nyu.edu/irex/index-e.html (accessed Feb. 2018).

More importance will be given to the areas of work of this thesis. Table 2.4 provides a summary of the BioNLP challenges described. Specific details about the challenges results in areas related with this work will be given in next chapters.

BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) challenges[77] consist in the evaluation of IE systems applied to the biomedical domain. BioCreAtivE challenges have been organized in conjunction with workshops since 2004.

The main issues addressed at BioCreAtivE are concerned with IE from scientific literature of the biological domain. The main tasks consist in NER, entity normalization and entity linking (for instance, recognition of genes, proteins, chemical compounds and drugs, normalization of gene names and linking of gene and protein names to existing database entries) and relation recognition (such as protein-protein interactions -PPI-, chemical-protein interactions, chemical-disease and drug-disease relations). Binary document classification has also been evaluated.

CLEF[78], now Conference and Labs of the Evaluation Forum, formerly Cross-Language Evaluation Forum, promotes the access and retrieval of multilingual information. It runs yearly conferences since year 2000 and since year 2010 it also organizes challenges, called *Evaluation Labs*. Some of the Evaluation Labs organized between 2010 and 2017 are:

- **CLEFeHealth - CLEF eHealth Evaluation Lab** (2012-). Evaluation of IE and IR in medical reports written in different languages, speech recognition. In 2012 a workshop has been organized [248]. It was followed from 2013 on with following tasks, among others, whose classification can be seen in Figure 2.2: acronym normalization, NER and named entity recognition and normalization of disorders (2013), information extraction from English medical reports, including slot filling and negation and speculation detection (2014), named entity recognition and normalization from French scientific articles and drug inserts (2015 and 2016), named entity recognition, normalization and text classification in French clinical texts and English death certificates. In all cases normalization consisted of a mapping to UMLS CUIS.
- **ImageCLEF - Cross Language Image Annotation and Retrieval** (2003-). Evaluation of image automatic annotation, classification, analysis and multilingual retrieval, combining texts and images. Since 2004 there is a medical retrieval track, that includes different tasks associated with medical images.
- **QA4MRE - Question Answering for Machine Reading Evaluation** (2011-2013).
- **QALD-3 - Question Answering over Linked Data** (2013).
- **QA Track — CLEF Question Answering Track.** (2014-2015). It included BioASQ: Biomedical semantic indexing and question answering.
- **CLEF-ER - Entity Recognition** (2013). multilingual automatic annotation of named entities and normalization (attribution of CUIs) in corpora in English, French, German, Spanish, and Dutch. The corpora were composed by the documents that were later part of the QUAERO corpus. Texts in Spanish stemmed from the EMEA corpus.
- **LifeCLEF** (2014-2016), evaluation of multimedia information retrieval on biodiversity data for identification of species.

[77] BioCreAtIvE: http://www.biocreative.org/ (accessed Feb. 2018).
[78] CLEF http://www.clef-initiative.eu/ (accessed Feb. 2018).

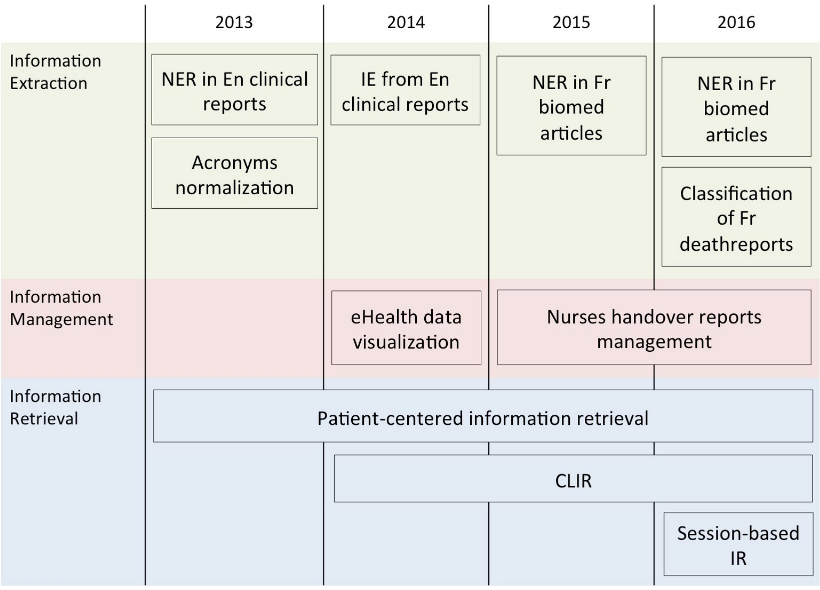

2013
2014
2015
2016
Information Extraction
NER in En clinical reports
Acronyms normalization
IE from En clinical reports
NER in Fr biomed articles
NER in Fr biomed articles
Classification of Fr deathreports
Information Management
eHealth data visualization
Nurses handover reports management
Information Retrieval
Patient-centered information retrieval
CLIR
Session-based IR

Figure 2.2: CLEF eHealth tasks. Taken from CLEF eHealth site (https://sites.google.com/site/clefehealth/).

BioASQ[79] is a challenge on biomedical semantic indexing and question answering. It has been organized yearly since 2013. The first three editions were organized by CLEF and since 2016 they are part of the BioNLP workshop of the ACL conference.

i2b2 (Informatics for Integrating Biology and the Bedside)[80] provides software, research datasets and organizes shared tasks. One of its goals is to provide software tools to extract clinical information from unstructured medical reports with the use of NLP. Several different challenges have been organized between 2006 and 2014. Among their goals are NER, de-identification of reports, identification of assertion and relations in clinical texts. Approximately 1,500 de-identified notes from the *Research Patient Data Repository at Partners HealthCare*, that have been used in the first four i2b2 Challenges have been released. The rest of the notes is planned to be released at each one-year anniversary of the corresponding challenge. i2b2 challenges include, the:

- **De-identification and Smoking Challenge (2006)**, whose goal is to evaluate the state-of the-art in automatic de-identification and the identification of smoking status in discharge summaries,
- **Obesity challenge (2008)**, that points at recognizing obesity and co-morbidities,
- **Medication challenge (2009)**, where medication information from clinical text had to be extracted,
- **Concepts, assertions, and relations challenge (2010)**, whose goal was to do NER of medical problems, treatments and tests, assertion classification on given findings (as present, absent, speculated, conditionally present in the patient at some future moment and mentioned, but associated with someone else) and relation classification of pairs of given concepts in clinical reports [201],

[79] BioASQ challenge: http://www.bioasq.org/ (accessed Jan. 2018).
[80] i2b2 https://www.i2b2.org/ (accessed Jan. 2018).

- **Coreference challenge (2011)**, where coreference resolution for electronic medical records has been addressed,
- **Temporal relations challenge (2012)**, where the extraction of temporal relations in clinical text has been evaluated, and, finally, the
- **De-identification and heart disease risk factors challenge (2014)**, that looked for automated systems for the de-identification of clinical reports and for the identification of risk factors for heart disease over time.

Additionally, a community annotation experiment, was performed for the edition of 2009 and publications about the annotation process for obtaining the datasets of years 2012 and 2014 competitions were developed.

TREC (Text REtrieval Conference)[81] was created in 1992 as one of the metric-based evaluations within the TIPSTER DARPA-sponsored[82] projects. The project concluded, but TREC still continues.[83] Nowadays TREC is co-sponsored by the United States National Institute of Standards and Technology (NIST) and the U.S. Department of Defense (DOD).

TREC's goal is to provide the necessary infrastructure (i.e. text collections and evaluation methodology, among others) for large-scale evaluation of text retrieval methodologies. TREC has yearly workshops, consisting of a set of tracks. The test collections and evaluation software are available to the community. It contains the first-large scale evaluation of non-English documents (including Spanish) and retrieval across multiple languages.

A presentation about the past and present TREC biomedical tracks done in occasion of the 25 years of TREC in 2016 is available online.[84] TREC Biomedical tracks include, the:

- ***Genomics Tracks* (2003-2007)**,[85] that comprised following tasks: ad-hoc IR and IE (2004), ad-hoc IR and categorization of full-text documents (2005), and retrieval of passages of biomedical documents, that contained answers to questions (2006 and 2007). A detailed description about the tasks and the corpora provided can be seen in the Genomics Track Overview paper [117] and details about each track can be studied in [116, 118, 119, 120, 121].
- ***Clinical Decision Support Track* (2014-2016)**,[86] whose goal was the retrieval of full-text biomedical articles of PUBMED Central relevant for answering a question in a set of given summaries of medical records. An example of a question is *What is the patient diagnosis?* 2014 and 2015 editions used synthetic summaries of medical records and 2016 edition used actual electronic health records (obtained from MIMIC-III database).
- ***Entity Track (2009-2011)***,[87] whose goal was to discover entities in web search [19].
- ***Medical Records Tracks* (2011-2012)**,[88] whose goal was to retrieve electronical health records and information within them to identify patients who might be candidate for clinical studies [273, 272], and the

[81] TREC conference http://trec.nist.gov (accessed Jan. 2018).

[82] DARPA stands for United States Defense Advanced Research Projects Agency.

[83] MUC was another evaluation created within the TIPSTER project.

[84] Past and present TREC biomedical tracks done in ocasion of the 25 years of TREC https://dmice.ohsu.edu/hersh/trec-25th-biomed.pdf (accessed Jan. 2018).

[85] TREC Genomics Tracks https://trec.nist.gov/data/genomics.html (accessed Jan. 2018).

[86] Clinical Decision Support tracks. See http://www.trec-cds.org/ 2014.html, 2015.html and 2016.html. (accessed Jan. 2018).

[87] TREC Entity Tracks http://trec.nist.gov/data/entity.html (accessed Jan. 2018).

[88] TREC medical tracks http://trec.nist.gov/data/medical.html (accessed Jan. 2018).

- ***Precision medicine track (2017)***,[89] that had the aim of providing precision medicine information to clinicians treating cancer patients. The tasks were to retrieve 1) scientific abstracts addressing relevant treatments for a given patient and 2) clinical trials for which a given patient is eligible to enroll.

Data released by TREC and TREC proceedings are available online.[90]

bioCADDIE's (biomedical and healthCAre Data Discovery Index Ecosystem dataset retrieval challenge (2016)[91] goal is to create ways to facilitate the access of biomedical researches to relevant datasets, in the cases where information retrieval queries cannot be answered by metadata associated with these datasets. Answer to these queries may involve analysis of structured datasets, unstructured texts and links to scientific articles.

2.7 Resumen

Entre los textos disponibles digitalmente los hay de distintos géneros (por ej. periodístico, científico, informes médicos) y dominios (por ej. legal, biomédico y de entretenimiento). Los textos pueden ser de naturaleza formal o informal, de acuerdo con la corrección en su escritura (considerando errores ortográficos y gramaticales). Los informes médicos son ejemplos de textos de carácter informal, debido a que generalmente se escriben con escaso tiempo y en ocasiones se requiere su brevedad.

Dentro del dominio biomédico, trabajaremos con informes médicos del área de la radiología (RR). En los informes médicos existe gran cantidad de abreviaturas, muchas de ellas ambiguas [204] y lenguaje especializado. Además, hay abundancia de términos de negación y especulación [42, 267, 61]. En particular, los RR, tienen lenguaje especializado, dado que estos informes constituyen una forma de comunicación entre el radiólogo y el médico que solicitó el estudio. La comunicación entre especialistas tiene que ser clara y oportuna [24, 37, 23]. Hay poca información acerca de cómo escribir un RR, pero las publicaciones existentes recomiendan utilizar textos breves, incluso con expresiones gramaticalmente incorrectas [111, 282, 49, 166, 235].[92]

En este capítulo presentamos, entre otras, las definiciones de procesamiento del lenguaje natural (NLP), extracción de la información (IE) y minería de textos (TM) y describimos algunas de las etapas de un proceso de IE. Introducimos la definición de entidad anatómica (AE) y hallazgo clínico (FI) y algunos tipos de informes médicos y terminología lingüística. Entre otros, definimos: *corpus* (plural *corpora*): una colección de textos, que puede estar enriquecida con algún tipo de anotación y *gold standard*: un corpus anotado por especialistas y que se utiliza como criterio de referencia para determinar cuáles son los resultados correctos de una tarea determinada. Mencionamos recursos existentes para NLP en general y para BioNLP e introducimos las métricas que se utilizan habitualmente para determinar la corrección de los sistemas informáticos relacionados con la temática. Finalmente, presentamos una revisión de trabajos previos y competencias realizadas en el área. En capítulos posteriores se proveerá información detallada de los trabajos previos de las distintas temáticas tratadas en esta tesis.

[89] 2017 precision medicine track: http://www.trec-cds.org/2017.html (accessed Jan. 2018).

[90] Data released by TREC http://trec.nist.gov/data.html. TREC proceedings and presentations http://trec.nist.gov/pubs.html (accessed Jan. 2018).

[91] bioCADDIE dataset retreival challengehttps://biocaddie.org/biocaddie-2016-dataset-retrieval-challenge-registration (accessed Jan. 2018).

[92] http://www.chestx-ray.com/index.php/practice/how-to-compose-a-radiology-report-guidelines (accedido Jan. 2018).

name	description	lang.	main tasks	genre	domain
BioASQ[a]	biomedical semantic indexing and question answering	EN	QA, SI	AB, SA	biomedical
bioCAddie	biomedical and healthCAre Data Discovery Index Ecosystem	EN	TR, DC		biomedical
BioCreAtiVe[b]	Critical Assessment of Information Extraction in Biology	EN	NER, EL, EN, RE, DC	AB, PT, SA	biological
CLEF[c]	Conference and Labs of the Evaluation Forum	EN, FR	AN, NER, IE, IR, QA, DC	ClR, SA, DI	biomedical
CoNLL 2010 Shared Task	Learning to Detect Hedges and Their Scope in Natural Language Text [89]	EN	ND, UD	SA,WP	biological
i2b2[d]	Informatics for Integrating Biology and the Bedside[199, 201]	EN	DEID, IE, CR , NER, RE, ND, UD	DS, EHR, ClR	medical
TREC	Text retrieval conference	EN, SP	TR, IE, IR, DC	EHR*	biomedical*

[a] BioASQ http://www.bioasq.org/workshop (accessed Mar. 2018).
[b] BioCreAtiVe http://www.biocreative.org/events/ (accessed Mar. 2018).
[c] CLEF http://www.clef-initiative.eu/ (accessed Mar. 2018).
[d] i2b2 https://i2b2.cchmc.org/faq#data1 (accessed Mar. 2018).

Table 2.4: BioNLP Challenges summary. References of the tasks: AN: acronym normalization, CR: coreference resolution, DC: document classification, DEID: de-identification, EL: entity linking, EN: entity normalization, IE: information extraction, IR: information retrieval, ND: negation detection, NER: named entity recognition, QA: question answering, RE: relation extraction, SI: semantic indexing, TR: text retrieval and UD: uncertainty detection. References on the languages: EN: English, FR: French, SP: Spanish. References on the type of documents: AB: abstracts of scientific papers, ClR: clinical reports, DI: drug inserts, DS: discharge summaries, EHR: electronic medical records, PT: patents text, SA: scientific articles, WP: Wikipedia publications; *: among others.

Part II

Development

CHAPTER 3

Annotation of Spanish radiology reports

This chapter describes the annotation process, schema and guidelines followed to obtain an annotated dataset for entity recognition, negation detection and relation extraction in Spanish radiology reports. First, we introduce the importance and difficulty of annotation processes. Then, we describe the dataset chosen to perform the annotation and the preprocessing performed to it. After that, we describe the annotation process, annotation schema and annotation guidelines followed. Next, an analysis of the resulting annotation is shown. Finally, previous work, discussions and conclusions are presented.

3.1 Introduction

Pustejovsky and Stubbs [210] define an annotation over an input as "any metadata tag used to mark up elements of the dataset".

Annotated corpora are required to evaluate information extraction algorithms and for training supervised machine learning methods.

As already mentioned, there is scarcity of publicly available annotated corpora in the biomedical domain, in particular for non-English texts.

There are two main reasons for that: first, the generation of new annotated data has high associated costs due to the need of expert knowledge in order to correctly interpret the specialized vocabulary texts, and, second, the ownership of the data is very discussed, especially when it refers to information that might identify the patient. Each country and institution has different regulations and some tasks - e.g. anonymization and sometimes authorization from the patients and from the institutions- have to be performed before publishing the data. So, although the availability of annotated data is a highly valuable asset for the research community, it is very difficult to access it. As Neves and Leser [194] mention, "the lack of gold standards is considered as one of the main bottlenecks for developing novel text mining methods".

Furthermore, annotation guidelines have to be carefully designed and reviewed in an iterative process. They have to be clear enough so as to be followed by different annotators with a high annotation agreement.[1] In this regard, it is important to

[1] Inter-annotator agreement (IAA) is a measure that indicates the coincidence in annotation criteria among different annotators and will be defined in Section 3.4.1.

notice that in the biomedical domain there are often differences in the annotation criteria among different annotators [270] or even among the same annotator (because of incompleteness of annotation guidelines -often originated by the complex nature of texts-, the allowance of multiple tagging and human errors) [255, 238]. Common differences in annotations are boundaries of entities and the classification of entities in their type. Inconsistencies in the annotation may affect training and evaluation of machine learning techniques as well as the evaluation of other techniques. Finally, it does not exist an annotation standard.

The aforementioned situations make the guidelines definition a difficult and time-consuming task.

Among others, following tasks are described by Ide and Pustejovsky [127] as needed to do an annotation process: create files in a standard file format, write annotation guidelines, define needed annotator skills and knowledge, train annotators in the annotation schema until reaching an acceptable inter-annotator agreement (IAA), plan the annotation order and assignments, distribute documents to the annotators, monitor annotator's progress, collect annotations from the annotators, track inter-annotator agreement to ensure the quality of the annotations, schedule meetings, track worker hours and project budget.

Many tools that have been created to support experts in annotating texts. The intuitive use, the support of different text formats, the availability of automatic pre-annotation functionalities, the access to background knowledge such as ontologies, and a comprehensible visualization of annotations are important aspects of them [194].

Instead of manual annotation, distant supervision can be used. In the distant supervision setting, the corpus is usually automatically annotated using an external knowledge base. Distant supervision reduces effort in building training data but introduces noise. There are studies about how to reduce the uncertainty in distant supervision methods [160, 113]. Adding a small set of human-annotated training data to a distant supervision dataset, usually increases significantly the precision of the system [123]. This human-annotated data could be obtained through crowdsourcing. In particular, crowdworkers could be trained by high-quality labeled data [156].

We are interested in supporting physicians with information extraction methods, such as named entity recognition (NER), relation extraction (RE) and negation and uncertainty detection in Spanish radiology reports. This could help to detect the main illnesses present among the patients, the patients evolution and to detect problems not expressed in an explicit way. To train algorithms and to evaluate our methods we need a gold standard. To the best of our knowledge, there are no publicly available annotated datasets of Spanish medical reports. Neither are there annotated datasets for relation extraction between this type of entities.[2]

For this reason, we worked on the creation of an annotated corpus of Spanish radiology reports for named entity recognition, negation and speculation detection and relation extraction. This chapter describes the process, the annotation schema and decisions taken in a way that it is possible to re-use by other researchers working with the same goal. We plan to publish our annotated dataset, in what would constitute, to the best of our knowledge, the first Spanish corpus publicly available for named entity recognition and for relation extraction in clinical reports. The corpus also includes annotations for negation detection.

[2] Recently, briefly before publishing the results of our annotated corpus [59], a dataset annotated for negation in Spanish medical reports has been put publicly available [165]. Others, that include radiology reports, are being developed [61].

After describing the dataset used, and the preprocessing done to it in Section 3.2, Section 3.3 describes the annotation process, schema and guidelines. Then, in Section 3.4, an analysis of the resulting annotated dataset, that includes the number of entities and relations discovered by the annotators and the inter-annotator agreement among others, is presented. Next, Section 3.5 describes previous related work. Finally, Section 3.6 discusses the results of the dataset analysis and Section 3.7 presents conclusions. The chapter includes part of the content of following co-authored publications [58, 59].

3.2 Data

In this section we describe the characteristics of our dataset and provide some examples. We also explain how we selected and anonymized the reports to be annotated.

We have 85,621 different kinds of ultrasound reports (e.g. kidney, abdominal, small parts, Doppler) provided by one of the most important public health hospitals in Argentina. Reports are short, there is an average of 8 sentences per report and an average of 9 words per sentence. They begin with a report number, the age of the patient at the time of the imaging study, the date of the study and the patient's identification number. They contain only one section, that includes observations, conclusions and suggestions. In some cases, they have information about the doctor or doctors who performed the ultrasonography. There is abundance of abbreviations and acronyms and also grammatical and lexical errors exist.

Below, we show examples of two radiology reports with their translation to English. Some particularities and orthographic errors are highlighted with bold and are also explained.[3]

Examples

Report examples

53222 —12a —20070503—A12402 HIGADO: tamano y ecoestructura normal. VIA BILIAR intra y extrahepatica: no dilatada. VESICULA BILIAR: alitiasica. Paredes y contenido normal. PANCREAS: tamano y ecoestructura normal. BAZO: tamano y ecoestructura normal. Diametro longitudinal: **(cm)6.3** RETROPERITONEO VASCULAR: sin alteraciones. No se detectaron adenomegalias. No se observo liquido libre en cavidad. Ambos riñones de características normales. **RD**: **6.5cm** **RI**: **7.7cm** Vejiga **s/p**. Visto con Dra. Galarza Martha M.N. 93247

53222 —12a —20070503—A12402 LIVER: normal size and echotexture. BILIARY TREE: intra and extrahepatic: not dilated. GALLBLADER: no gallstones were seen. Walls and content normal. PANCREAS: normal size and echotexture. SPLEEN: normal size and echotexture. Longitudinal diameter: (cm)6.3 RETROPERITONEAL COMPARTMENT: unremarkable. No lymphadenopathy was detected. No free fluid in the peritoneal cavity was observed. Both kidneys unremarkable. RK: 6.5cm LK: 7.7cm Bladder w/p. Seen with Dr. Galarza Martha N.L.N. 93247

Notice that some words are sometimes written in upper case (in these reports

[3] Data regarding physicians, that performed the study was changed.

many of the mentioned organs; in some cases, procedures, such as *ECOGRAFIA RENOVESICAL* -renovesical ultrasound- are written in upper case; usually abbreviations and acronyms are also written in upper case). This characteristic differs from report to report and also in the same report. Some accents and tilde signs are present (e.g. *características* and *riñones*), but others are missing (e.g. *extrahepatica*, should be *extrahepática* and *tamano* should be *tamaño*). There is presence of abbreviations and acronyms (e.g. *RD* and *s/p*) and measures are not separated from measure units (6.5cm should be 6.5 cm). Finally, in Spanish decimal separators are commas instead of dots (6.5 should be 6,5).

38242 —10a —20070803—A12540 Vasos abdominales permeables, marcada ascitis con moderada esplenomegalia homogenea, **estan** marcada la ascitis que no se pueden valorar asas intestinales para descartar tiflitis o apendicitis. . Vesicula con contenido (**sludge**) que no se moviliza, por ayuno **?** '. pancreas sin alteraciones. Higado homogeneo..Liquido pleural bilateral, deseariamos evaluar RX de torax.

38242 —10a —20070803—A12540 Patent abdominal vessels, marked ascites with moderate homogenous splenomegaly. Ascites is so noticeable that intestinal loops cannot be evaluated to rule out typhlitis or appendicitis. . Gallbladder with content (sludge) that does not mobilize, because of fasting ? '. pancreas unremarkable. Homogeneous liver.. Bilateral pleural effusion, we would like to evaluate chest X-ray.

Notice that *sludge* is an English term. Additionally, the question mark indicates that the cause *por ayuno* -because of fasting- is not sure. *estan* (are) should be *es tan* (is so).

As previously mentioned, Spanish has diacritics: accents and the *tilde symbol.*

Informal texts, such as chats, social media messages, instant messages (such as whatsapp and SMS -short message service-) are usually written without paying attention to non-ASCII characters or to punctuation signs. This is mainly due to time constraints, to lack of linguistic knowledge (orthography and syntax) and to lack of training in typing.

Clinical reports and in particular radiology reports constitute no exception in the avoidance or misplacement of accents and the tilde symbol. They are usually also written in a telegraphic style. Sometimes verbs are missing. There are many time constraints that help to a poor writing. Furthermore, physicians might not have always access to the same computer, which might increase typing errors. Finally, the focus is put on being able to understand the texts by themselves or by other physicians -often from the same institution- and not by NLP techniques.

Some common errors and characteristics appearing in our radiology reports are explained next:

Some orthographic and grammatical errors:

- **spelling errors:** e.g. *hetrogeene* instead of *heterogenea* (heterogeneous), *rinonoes* for *riñones* (kidneys), *híígado* instead of *hígado* (liver) and *dilataci{on* and *dilatación* instead of *dilatación* (dilation),
- **grammatical errors:**, such as *contorno irregulares* instead of *contorno irregular* (irregular contour),

- **missing diacritics:** e.g. accents in *formacion* instead of *formación* (formation) and tilde sign in *rin'on* or *rinon* instead of *riñón* (kidney),
- **missing punctuation marks:** e.g. "No visualizo apendice Se visualiza coledoco 7mm" ("I do not visualize appendix Choledochus 7mm is visualized") should be "No visualizo apendice**.** Se visualiza coledoco 7mm" and "pancreas sin alteraciones bazo de tamaño normal" ("pancreas unremarkable spleen normal sized") should be "Pancreas sin alteraciones**.** Bazo de tamaño normal",[4]
- **missing verbs:** e.g. *no dilatación de via biliar -no bile duct dilatation-* instead of *no se detectó dilatación de la vía biliar.*
- **dots do not appear right after an abbreviated term.** e.g. "R .Derecho," (right k.)[5] should have been written as R. Derecho, and tokenized as "R." "Derecho" and ",", but tokenizers tokenize it as "R", "." ,"Derecho",",".
- **dots separating tens and units are located erroneously.** e.g. in **RD: 6 .6 mm** there is an incorrect space between 6 and 6mm. The tokenizer will tokenize this measure as "6", "." , "6" and "mm", where it should have been tokenized as "6.6" mm.
- **sometimes a number and the measure unit are not separated by a space** e.g. **6.2cm** will be tokenized as "6.2cm" instead of "6.2" and "cm", result that would have been obtained if the input would have been "6.2 cm". An example of two previous problems is **1 .4cm**.
- "**Art** hepatica mide 2.7 **mm .Vci** retroehepatica normal" -*Hepatic art measures 2.7 mm. Normal retrohepatic IVC-* . *mm .Vci* introduces tokenization issues.
- ambosmediastinos for ambos mediastinos *both mediastinum* will not be separated into two tokens.

Existence of abbreviations and acronyms

- **lack of standards** In s/p: sin particularidades (w/p, without particularities), VN: valor normal (usual value), v biliar: via biliar (bile duct) and art: arteria (artery) mm: milímetros (milimeters), only s/p and mm are standard abbreviation.
- **existence of multiple abbreviations or acronyms for the same concept**. e.g. RD and RDER: riñón derecho (right kidney) .

Measures and units of measure

- **the same unit of measure is written in a variety of ways**, e.g. cm, cm. c. for centimeters. Sometimes it is written between brackets with or without spaces -e.g. 2,6 (cm) and 3,8 (cc)-,[6]
- **characteristics of the measures**. there might be more than one measure corresponding to different type of measures (longitudinal, transversal, etc.). Measures can be written with or without units of measure. e.g. 2,6 (cm) x 1,9 (cm) x 1,4 (cm) and 15 x 4 x 3,5 cm,
- **lack of standard in the way of writing**. usually the measure is written before the unit of measure, but sometimes it is written after it. eg: (cm) 7; the measure and unit of measure can appear between brackets (5,7 cm), measures are frequently missing, eg: Rinon derecho: diametro longitudinal (cm).;. or , can be used as tens and decimal separator (e.g. 3.1 cm and 3,1 cm),

[4] A comma could have been used instead of a period.

[5] k. means kidney.

[6] Commas are used as decimal separators in Spanish.

- a combination of many of the above can also appear. e.g. 13x8mm

We will explain next how we selected the dataset to be annotated.

3.2.1 Selection of the dataset to be annotated

We discovered that some of the 85,621 reports had exactly the same content. We removed all the repetitions. There were also reports that had no content or that had less than three words and some that where test reports (with nonsense content). All of them were also removed. After this filtering process we obtained 79,123 reports.

Reports were anonymized and a subset of these 79,123 reports were annotated. We will next explain how we selected the dataset to be annotated.

Since we are interested in examining the existence of different health problems, we performed a selection of the reports to be annotated defining four sets. The first, called **hypertrophic pyloric stenosis**, involves reports containing information about the pyloric muscle and pyloric canal, that might refer to pyloric stenosis (pyloric obstruction); the second, called **splenomegaly**, contains reports referring to the spleen, whether of normal size or enlarged; the third, called **appendicitis**, has reports that mention the appendix and that might or might not refer to appendicitis and the fourth, called **generic**, comprises a set of ultrasound reports corresponding to different body parts and possibly involving different findings or diseases not included in the previous cases.

The first three sets are particularly interesting because the extraction of entities and relations among them could suggest possible medical problems, that might lead to surgical interventions. The fourth set is useful for studying entities and relation extraction in general terms. All sets were put together and one final dataset to be annotated was built.

For instance, taking into account the age of the patient and the size of the spleen it is possible to determine whether the patient has splenomegaly (enlargement of the spleen) or not according to normal reference values. Furthermore, the visibility of the appendix, its maximal outer diameter not exceeding 6 mm. and its non-compressibility are the most reliable criteria used in the diagnosis of acute appendicitis. And, finally, a thickness of 3 mm or greater in the pyloric muscle, a length of the pyloric canal of 15-17 mm or more and a patient less than 3 months old, are useful parameters for diagnosing pyloric stenosis. Thus, information is not always explicitly written in medical records (e.g. in the cases described above, splenomegaly and appendicitis can be detected through indirect information like the measure of the spleen or the visibility of the appendix, instead of explicitly). The automatic detection of critical issues, such as appendicitis and pyloric stenosis, is of interest and is being studied (see e.g. [78] and [183]) and could allow their communication by pager or alternatives methods, as is described in [149].

Two native Spanish speakers annotated the texts. One of them with a medical background and the other with a technical background. A subset of the reports was annotated by both annotators. Table 3.1 presents the number of files processed by each annotator. Overall 513 different files have been annotated.

	ann. 1	**ann. 2**	**both**	**total**
Total	364	210	61	513

Table 3.1: Number of annotated files annotated by annotator 1 (ann. 1), annotator 2 (ann. 2) and by both of them.

3.2.2 Presence of diacritics in the selected dataset

In the reports to be annotated, we noticed the previous mentioned problem of missing accents and tilde signs and of no standard use of them even in the same report. The same word is written sometimes with accents and sometimes without. Some words never appear with accents, although needing them. Some words have accents, but not in the right letter (due to typographical errors), and others have a *grave accent* (`) instead of an *acute accent* (´). In Table 3.2 we present the percentage of accentuated vowels in our 513 annotated radiology reports compared to the number of accentuated vowels in a set of abstracts of scientific articles of the medical domain written in Spanish, that has the same amount of words. It can be seen that in the reports, the amount of accentuated vowels is in average 11.26 times less than in a formal article written in the same language.[7]

	radiology reports # with diacritics without diacritics (%)	formal text # with diacritics without diacritics (%)	relation among formal and informal texts in %
a (á vs. a)	74 /21,134 (0.35%)	834/21,190 (3.94%)	11.24%
e (é vs. e)	54 /19,681 (0.27%)	608/23,087 (2.63%)	9.60%
i (í vs. i)	116/14,627 (0.79%)	927/15,139 (6.12%)	7.72%
o (ó vs. o)	126/16,864 (0.75%)	1,884/16,534 (11.39%)	15.25%
u (ú vs. u)	10/4,641 (0.22%)	168/6,243 (2.69%)	12.49%
n (ñ vs. n)	339/12,262 (2.76%)	455/14,163 (3.21%)	1.16%

Table 3.2: Presence of diacritics in the 513 reports prior to the normalization compared with the presence of diacritics in a set of abstract of scientific articles in the medical domain written in Spanish. The last column represents how more frequently accents appear in formal texts in comparison with radiology reports.

3.2.3 Report anonymization

In some cases, medical or external information about patients, other than their id, could help reveal their identity. Consider for example an indication of the place where the patient lives or an uncommon illness that occurs in a low population area. This is not our case, since reports belong to a hospital that treats daily more than 1,500 patients and that is located in a city with 2,891,000 inhabitants.

As we previously mentioned, reports contain a report number, a patient identification number, the date of the study and the age of the patient at the time of the imaging study. In some cases, they also have information about the doctor or doctors who performed the ultrasonography and their medical license number. Names of physicians might be preceded with the title *Dr.* or *Dra.* (male or female doctor), e.g. *Dra. Suarez.* In occasions, the medical license number is written after the doctor's name, sometimes it can appear without the doctor's name (e.g. Dr. Heinz MN: 24,317 or MN:28,317). The license number can have been issued at a national (*MN*) or at a state level (*MP*).[8] Some reports are followed by the name of one or more doctors who also discussed the study. Their names are preceded by the words *Visto con* (seen with), for example: *(...) Visto con Pedro Chaves* (seen

[7] The average was taken from the last column of the table.

[8] *MN* stands for *matrícula nacional -national license number-* and *MP* stands for *matrícula provincial -state license number-*.

with Pedro Chaves). In some cases, names appear at the end of the report, without being preceded by any title.

Before performing the automatic and manual annotations reports had to be anonymized. Therefore, regular expressions were used considering the different ways of writing the title of the doctors (e.g. *DR*, *Dr.*, *doctor*, *Dra.*), the doctor's names, the enrollment numbers and the order among them. Also, names of the doctors appearing with titles or enrollments were searched to see if they appeared without titles and without enrollments and were removed. Patient and report identification were changed in a way that it is not possible to identify a patient. The date of the study was removed. Finally, strings indicating the name of other physicians who analyzed the ultrasonography were also removed.

In terms of the different types of anonymization processes that were reviewed in Section 2.6.3, we do a non-perturbative anonymization. We did not remove the age of the patient, so we are preserving the patient demographics.

3.3 Annotation Process

Our annotation guideline was improved within three iterations consisting of annotation and revision. We followed a procedure similar to the MAMA cycle (Model-Annotate-Model-Annotate) proposed by Pustejovsky and Stubbs [210], that involves iterating between specifying the annotation schema and doing the pilot annotations. Once annotators are trained and have annotated a small amount of data, data is inspected, and IAA is calculated. The model is modified until the annotations are stabilized.

As Simpson and Demner-Fushman [229] mention, there are three possible approaches to annotate biomedical texts. 1) a *manual annotation* based on annotators knowledge, 2) an *assisted annotation*, in which the output of an annotation tool is manually corrected, and 3) an *ontology-based annotation* (manual or assisted), where only terms and relations present in an existing knowledge source are annotated. Each approach has its advantages and disadvantages.

In order to decrease the annotation time, we used an assisted annotation. Entities, negation and uncertainty terms were pre-annotated automatically. Then, based on the annotation guideline, the two annotators annotated the pre-annotated reports, making corrections and adding relations.

The selected annotation approach has as disadvantage that the annotations could be biased by the pre-annotation. The advantage is that annotations were performed much faster that it would have been without a pre-annotation.

In this section we present the annotation schema and guidelines, and the automatic and manual annotation process.

3.3.1 Annotation schema

The following entities and characteristics were considered for the annotation:

- **findings (FI)**: entities corresponding to a pathological finding or diagnosis, e.g. cyst, gallstone, abscess,

- **anatomical entities -or body parts- (AE)**: e.g. breast, right thyroid lobe, liver,

- **location (LO)**: location in the body, e.g. medial, distal, peripheral, unilateral, apical, adjacent,

- **measure (ME)**: e.g. 0.3 mm, 0.5 cc, 2 cm., 0.8 (cm.), large, small, scarce, minimum,
- **type of measure (TM)**: indication of the kind of measure that a number is referring to. e.g. in *longitudinal 3 (cm) and transversal 1 (cm)*, *longitudinal* and *transversal* will be annotated as two type of measures and *3 (cm)* and *1 (cm)* as two measures, and
- **texture (TE)**: e.g. homogeneous or heterogeneous.

Other annotated concepts were the following ones:

- **negation (NT) and uncertainty terms (UT)**: We call them modifiers. e.g. *no evidence of* and *might correspond to*,[9]
- **abbreviations and acronyms**: e.g. *RI* for *riñon izquierdo* (*left kidney*), and
- **temporal terms (TT)**: includes two types of terms. Terms that denote mentions to the past (e.g. *history of*, *preoperative* and *previous*) and terms that express conditionals (e.g. *in the eventuality*, *if* -consider *if the patient has fever again (...)*-).

The following binary relations were annotated:

- **occurs in**: among findings and the part of the body where they occur (AE or LO). e.g. in *vescícula biliar de paredes engrosadas*, -*thick(FI)-walled(LO) gallbladder(AE)*-, the finding (*engrosadas*, -*thick*-) occurs in the wall (LO),
- **located in**: between location and an anatomical entity. The goal is to know where in an anatomical entity a finding is located. e.g. in the example shown above, the walls (LO) are located in the gallbladder,
- **area of**: associates an anatomical entity with a location. e.g. in *kidneys without enlargement of the excretory pathway*, there is an *area of* relation among excretory pathway (LO) and kidney (AE),
- **has measure type**: associates a type of measure with a measure. e.g. in *longitudinal 3 (cm), anteroposterior 0.54 (cm)*, the measure *3 (cm)* has measure type longitudinal,
- **measure of**: associates a measure or a type of measure with an anatomical entity, a location or a finding. See example in Figure 3.2. Also, in *pyloric muscle thickness: 3.5 cm.*, there is a relation *measure of* from thickness (TM) to the pyloric muscle (AE) and a relation *has type* from thickness to 3.5 cm. (ME),
- **texture of**: associates an entity of *texture* type to an anatomical entity, a finding or a location. e.g. in *[kidneys](AE) of [conserved](TE) echotexture*, the conserved echotexture is related to the AE kidneys,
- **negates**: relates a negation term with a finding. e.g. in *without enlargement*, the NT *without* is related with the FI *enlargement*,

[9] Negation and uncertainty terms are sometimes called negation and uncertainty cues in the literature.

- **speculates**: relation among an uncertainty term and a finding. e.g. in *compatible (UT) with [fatty liver] (FI)* the uncertainty term compatible is related with the fatty liver, and

- **not present**: relates terms referring to the past or conditional terms and a finding. e.g. in *gallbladder(AE): history(TT) of cholecystectomy (FI)*, the cholecystectomy does not necessarily exist at the present moment and is related as not present with the temporal term.

Entities and relations to annotate were selected based on the named entities and relations that are interesting for physicians.

3.3.2 Annotation guidelines

During the annotation process and discussion rounds with the annotators, the original annotation guideline was adjusted. The main final annotation guidelines were following:

- **largest possible term**: as in MUC NER task definition[10] (see Appendix A.1.3), the largest possible term of a particular entity type (that contains as substring terms of the same entity type) has to be annotated (e.g. [[[retro[peritoneo]] vascular][11] should be annotated as retroperitoneo vascular -*vascular retroperitoneum*-),

- **use of lexicons as resources**: doubts about the category of an entity (sometimes it is not clear whether a term is, for example, an anatomical entity or its boundaries are not clear) should be solved using RadLex (it needs a translation into English) or UMLS (that exists for Spanish),[12]

- **existence of spelling errors**: terms with spelling errors should also be annotated,[13]

- **multi-name expressions/terms**: unlike MUC-7 NER task definition, when there is elision of the head of one conjunct, the expression should be annotated as different terms (discontinuous expressions can be annotated by our annotation tool, e.g. in the construction *intra and extrahepatic*, annotators were asked to annotate *intrahepatic* and *extrahepatic* as entities). The decision to annotate the different terms that form multi-word expressions was taken -after some discussion- because we want the gold standard to be correct and representative of the entities existing in the real world. However, to avoid too complex annotations, cases with more than three terms were annotated as a single *term.*

- **relations across sentences**: they have to be annotated (e.g. in *orthotopic left kidney. Size diminished and (. . .)*, *size* refers to the *orthotopic left kidney* and a relation among them has to be annotated),

[10] http://itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html (accessed Jan. 2018).

[11] Words between brackets show valid anatomical entities.

[12] RadLex is more appropriate for the radiology domain, but has the disadvantage of not being translated into Spanish.

[13] We have taken this decision, although it will worsen the results, in order to have an annotation that reflects reality.

- **abbreviations**: the abbreviations and acronyms correspondig to some entity types (e.g. AE or FI) should not only be annotated as abbreviations, but also as entities (e.g. in *RD* -riñon derecho, *right kidney*-, *RD* should be annotated as abbreviation and as anatomical entity as well),

- **segmentation of annotations**: if it is not clear if part of the term corresponds to a LO and part to an AE, if from the term it is clear where the AE is located, the whole term should be labeled as anatomical entity, else a segment might be labeled as LO and another as AE (e.g. lymph node should be labeled as an AE, since it is possible to identify the location of lymph nodes in the body. The same occurs with *right iliac fossa*).[14] In *upper part of the head* it is not clear what exactly the upper part is. So *upper part* should be annotated as location and *head* as AE and *the tumor is located in the upper left part of the liver* should be annotated as follows: *the [tumor](FI) is located in the [upper](LO) [left part of the liver](AE)*,

- **prioritize anatomical entities over locations**: if there is a doubt as whether a term corresponds to an anatomical entity or to a location, it should be annotated as anatomical entity,

- **prioritize findings over locations**: if a concept referring to a finding includes a location, the largest possible concept that refers to a finding has to be annotated (e.g. *pyloric stenosis* refers to a FI, that includes a location. Therefore, [pyloric stenosis](FI) should be chosen over [pyloric](AE) [stenosis](FI)),

- **annotation of negation and uncertainty terms**: negations and uncertainty terms should be annotated only if there is a relation among them and a finding, and finally,

- **annotation of anatomical entities**: anatomical entities have to be annotated although there is no relation among them and a finding (e.g. in *right lobe of the liver has the usual size*, *right lobe of the liver* should be annotated, although it is not associated to any finding).

Some other decisions that had to be taken were how to annotate certain frequently occurring concepts in the best way. For example, we decided to always annotate *kidney implant* as an AE. Furthermore, *ovarian cyst* and *cyst in the ovary* should be annotated as [ovarian cyst](FI) instead of [ovarian](AE) [cyst](FI), and as cyst(FI) in the ovary(AE).

The annotated dataset can have embedded entities and multilabeled entities.[15] For example, "esplenomegalia homogenea" should be annotated as "[esplenomegalia [homogenea](TE)](FI)" and "normal" in "normal size and echotexture" should be annotated as a *measure* and as a *texture*.

Figures 3.1, 3.2 and 3.3 show examples of annotations of some sentences. All images were taken from brat annotation tool[16] [241].

[14] RadLex should be used as a source to detect which is the largest possible concept corresponding to an AE (i.e. if right iliac fossa exists as AE in RadLex then it should be annotated as an AE)

[15] One entity is embedded into another if the text of one is a proper subset of the text of the other. Entities are said to be multilabeled if there exist more than one different label for the same entity.

[16] brat annotation tool http://brat.nlplab.org/ (accessed Nov. 2017).

entity name	UMLS STY
anatomical entity	Body Part, Organ, or Organ Component; Body Space or Junction; Body System; Tissue
finding	Anatomical Abnormality; Congenital Abnormality; Acquired Abnormality; Finding; Sign or Symptom; Pathologic Function; Disease or Syndrome; Mental or Behavioral Dysfunction; Neoplastic Process; Injury or Poisoning

Table 3.3: Mapping of UMLS Semantic Types (UMLS STY) to our annotation schema (anatomical entities and clinical findings) to perform the automatic annotation.

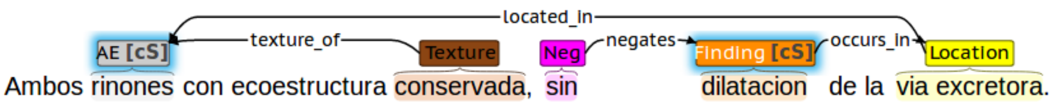

located_in
texture_of
negates
occurs_in
AE [cS]
Texture
Neg
Finding [cS]
Location
Ambos rinones con ecoestructura conservada, sin dilatacion de la via excretora.

Figure 3.1: Example of an annotation. "Both kidneys of conserved echotexture, without enlargement of the excretory pathway."

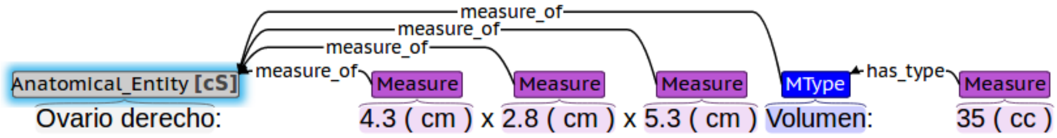


Figure 3.2: Example of an annotation. Measures of the right ovary and its volume are given.

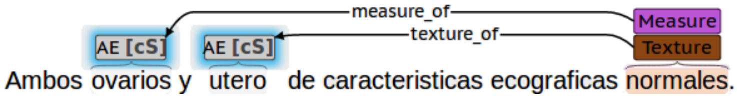

measure_of
texture_of
AE [cS]
AE [cS]
Measure
Texture
Ambos ovarios y utero de caracteristicas ecograficas normales.

Figure 3.3: Example of an annotation. "Both ovaries and uterus of normal echographic signs."

3.3.3 Automatic pre-annotation

In order to decrease the annotation time, entities were pre-annotated automatically. For this purpose, regular expressions, UMLS and a manually-created dictionary were used.

Regular expressions were used to detect the concept *measure*. *Anatomical entities* and *clinical findings* were detected with the use of some semantic types (STY) of UMLS (see mapping among our concepts and UMLS STYs in Table 3.3). First, concepts of the Spanish UMLS were mapped to the radiology reports. If one of the corresponding semantic types corresponded to an anatomical entity or to a clinical finding, then the concept was pre-annotated. Finally, a manually-created dictionary, that contains terms and their corresponding entity type -such as negation and uncertainty terms, locations or textures- was used. For example, *no puede descartarse* (*cannot be discarded*) is mapped to *Uncertainty*. Many of the concepts in this dictionary have been included within an iterative process followed from the annotations performed by the annotators.

A dictionary lookup algorithm was used to look for terms of UMLS and of the terms of the dictionary in the texts. Therefore, terms were stemmed with the Spanish version of Snowball implemented in NLTK.

After applying automatic pre-annotation, data was processed by human annotators. Annotations wrongly made by the tool were removed or corrected and missing concepts were included.

3.3.4 Manual annotation

The manual annotation was carried out by two Spanish native speakers: one of them with medical background (Annotator 1) and the other with a technical background (Annotator 2), that were not trained in medical document annotation. **brat** annotation tool[17] [241] was used for this purpose.

Many meetings were held in order to solve doubts. After having annotated a first pilot dataset (*Annotation iteration 1* in Table 3.7) doubts and differences in criteria were reviewed and the annotation guidelines (described in Section 3.3.2) were written with more detail. After two annotation-revision iterations, the annotations stabilized, the final guidelines were defined, and annotations were performed (in what we call *iteration 3* or *dataset 3*). The annotation guidelines development process is similar to the MAMA portion of the MATTER cycle proposed by Pustejovsky and Stubbs [210] and can be seen in Figure 3.4 (taken from the book).

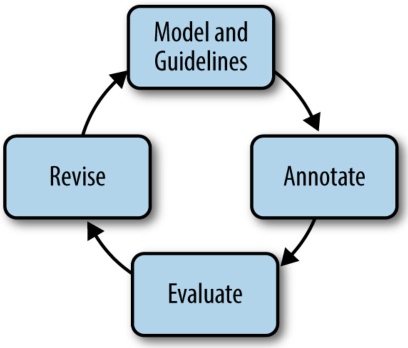


Figure 3.4: Annotation guidelines development process followed. Figure taken from Pustejovsky and Stubbs [210].

Disagreements were solved by a computational linguist and a computer scientist with expertise in the biomedical domain together with a physician.

3.4 Annotated dataset analysis

Once the annotation was performed, the annotated dataset was analyzed in order to know how many entities and relations of each type were found. We present the results next. Also, details about the size of the annotated dataset are provided. The analysis of the annotations is calculated for the whole set of 513 annotated reports. For those reports annotated by both annotators the annotation done by the medical student was chosen.

Table 3.4 describes the composition of the annotated dataset.

[17] brat annotation tool http://brat.nlplab.org/ (accessed Jan. 2018).

concept	number
number of radiology reports	513
total amount of words	36,211
total amount of sentences	4,175
avg. words per document	71
avg. sentences per report	8
avg. words per sentence	9

Table 3.4: Composition of Spanish annotated dataset. avg. means average

Table 3.5 shows the number of entities, modifiers of entities and other characteristics found in the annotated reports (abbreviations and acronyms, temporal expressions and multi-name terms). In all cases the total number of concepts and the number of different concepts is shown. It can be seen that there is a total of 880 abbreviations and acronyms. 470 of them correspond to anatomical entities and 7 to findings. The rest correspond to type of measures (266), locations (20) and 117 have no associated entity type. Table 3.6 shows for each type of relation, the entities related by them, the total number of relations and the number of different relations appearing in the annotated texts.

type	**total**	**different**
anatomical entities	4,398	405
finding	2,637	745
location	722	201
measure	3,210	975
texture	1,890	74
type of measure	1,127	72
negation	1,489	51
uncertainty	109	26
abbreviations	880	105
temporal expressions	35	15
multi-name terms	788	210

Table 3.5: Type and amount of entities, modifiers and other characteristics in the annotated reports.

There are 867 relations across sentences and a total of 10,987 relations in the 513 reports. So, 7.89% of the relations are across sentence relations.

The most frequent multi-name terms are *via biliar extrahepática* (*extrahepatic bile duct*) (232) and *via biliar intra hepática* (*intrahepatic bile duct*) (219).

3.4.1 Inter-annotator agreement

To evaluate the consistency among the annotations performed by both annotators, the inter-annotator agreement (IAA) was calculated using the Cohen's Kappa coefficient (κ) [51]. An explanation of why κ is an appropriate measure can be seen in [17].

The kappa coefficient (κ) measures agreement among a pair of annotators taking into account the fact that the they could have assigned the annotation tags totally by chance and is defined as follows:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{3.1}$$

relation	**entities**	**total**	**different**
occurs in	FI-AE	2,161	750
	FI-LO	233	218
located in	LO-AE	538	165
area of	AE-LO	65	53
measure of	TM-AE	1,007	154
	TM-LO	46	36
	TM-FI	59	56
	ME-AE	1,651	578
	ME-LO	74	48
	ME-FI	407	346
has meas. type	ME-TM	1,123	831
texture	TE-AE	1,495	192
	TE-LO	387	54
	TE-FI	90	37
negates	NG-FI	1,478	164
speculates	UT-FI	96	86
not present	CT-FI	33	33

Table 3.6: Relations with more than five occurrences annotated among entities in annotation iteration number 3. Entities are abbreviated in the following way, AE: anatomical entities, CT: conditional temporal, FI: findings, ME: measure, LO: location, NT: negation terms, TE: texture, TM: type of measure and UT: uncertainty terms.

, where P(A) is the proportion of times that the annotators agree and P(E) is the proportion of times that we expect them to agree by chance.

κ was calculated with the scikit-learn toolkit[18] on a token level. Therefore, each input file was tokenized (using the NLTK tokenizer).[19] Multiple annotations per token are possible (and are frequently used) due to various meanings and to the existence of overlapping concepts,[20] e.g. *normal* can be labeled as measure, as texture or as both of them (consider, for example, the phrase *normal size and echotexture*). For the calculation of the IAA we decided to consider that a token is labeled in the same way by both annotators if and only if both annotators assigned the exact same set of labels to it (or no label at all).

Table 3.7 shows the IAA for each of the annotation datasets. It can be appreciated that it improves in each annotation iteration step. Annotation dataset 3 was the final one and was annotated once the annotation schema and criteria (see Sections 3.3.1 and 3.3.2) were stabilized.

ann. iter.	**# reports**	**# annotated by both annotators**	κ
1	16	16	0.5883
2	20	20	0.8577
3	513	61	0.8893

Table 3.7: Inter-annotator agreement (κ) and number of annotated reports in different annotation iterations. Column *# reports* describes the number of reports contained in each dataset.

[18] scikit-learn http://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html (accessed Jan. 2018).

[19] nltk http://www.nltk.org/ (accessed Nov. 2017).

[20] Two annotations are overlapped if they share some text.

The subset of dataset 3 annotated by both annotators contains 427 tokens with more than one annotation of a total of 5894 tokens. That is 7.24% tokens belonged to more than one entity according to one of the annotators.

3.5 Previous work

The definition of annotation guidelines is a time consuming and difficult task. There exist some previous definitions for more generic entity types (e.g. persons, organizations and geographical locations). For example, MUC-7 and ACE (Automatic Content Extraction) competitions defined guidelines for the named entity recognition tasks organized by them in the past.[21,22] Both define annotation guidelines for general domain entity types (e.g. persons, organizations and locations). The annotation criteria is not easy to establish. For example, both guidelines differ in the way that the name of a Saint has to be annotated.

ISO space[23] and ISO TimeML standards[24] establish guidelines of space-related features and of temporal relations.[25]

Wilbur et al. [281] defined annotation guidelines to categorize segments of scientific sentences in research articles of the biomedical domain (see also[226]).

There is usually a scarcity of available data for the biomedical domain. The department of Radiology Informatics of Stanford University owns a large dataset of radiology reports, that is not annotated, nor publicly available, as far as we know.[26] There are some annotated datasets available for languages different to Spanish in the clinical domain, e.g. for English [201, 208, 209], for Swedish [232], for French [191], for Polish [188] and for German [213]. Oronoz et al. [198] presented IxaMed-GS, an annotated dataset in Spanish for adverse drug reactions analysis. Although the dataset is in Spanish and addresses the biomedical domain, it concerns a different use case and covers different information (different entity types, not medical reports). Furthermore, it is not publicly available, to the best of our knowledge. Recently, in 2017, Marimon et al. [165] and Cruz et al. [61] annotated negations in Spanish clinical reports.

Bellow we continue with the analysis of previous work organized by topic.

Complexity. An overview about the complexities of annotations projects can be seen in *The Handbook of Linguistic Annotation* [127] and in *Natural Language Annotation for Machine Learning* [210]. The formatting of the files to be annotated, the way to deal with typos and to ensure that a specific number of files is annotated by at least two annotators, how to measure the inter-annotator agreement (IAA), processes and tools for annotation creation and annotation of clinical texts are some of the topics treated in the first book. How to build an annotation dataset and mentions of some annotation standards, among others, are handled in the second. Neves and Leser [194] describe the difficulty of building gold standards in the biomedical domain and the importance of using annotation tools for this task. Stubbs [244] explains the complexity of annotation in the biomedical domain and suggests a methodology

[21] http://itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html (accessed Jun. 2017).

[22] ACE annotation guidelines https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf (accessed Jun. 2017).

[23] ISO space standard https://www.iso.org/standard/60779.html.

[24] ISO TimeML standard https://www.iso.org/standard/37331.html.

[25] They have relation with our “type of measure” entities (longitudinal, transversal, etc) and temporal relations (in the past, etc.). See Section 3.3.1.

[26] http://langlotzlab.stanford.edu/nlp-datasets/ (accessed Jun. 2017).

for creating *light* annotation tasks for biomedical corpora avoiding long annotation periods and time-consuming trainings. Stubbs and Özlem Uzuner [246] perform a light annotation process, consisting of a non-exhaustive annotation in order to have a larger corpus and to avoid a time-consuming and expensive annotation process.

Types of annotations. Leech [154] presents a guide with good practices for doing corpus annotation. Among others, different type of annotations that can be done, such as semantic, pragmatic, syntactic, lexical and others are described.

Crowdsourcing. Most available copora are annotated by experts, but community efforts are also possible, often with the use of crowdsourcing tools, such as Amazon's Mechanical Turk (AMT)[27] or CrowdFlower[28]. If tasks require an important amount of domain knowledge, usually there are experts that carry out the annotation task. Within the context of the *Third i2b2 Workshop on Natural Language Processing Challenges for Clinical Records*, a community annotation experiment was organized. Annotation guidelines and a small set of annotated discharge summaries were released. Results were comparable to those annotations obtained by experts [200], but it is important to notice that the annotators came from the community of interest in the subject.

Additionally, a community annotation experiment, was performed for the i2b2 edition of 2009 and publications about the annotation process for obtaining the datasets of years 2012 and 2014 competitions were developed.

Annotation tools. Manual annotation is a topic that is being developed, surveyed and studied by different groups. Analysis of different tools have also been done. Neves and Leser [194] conducted a survey on existing tools for annotating biomedical texts. They considered 30 freely-available tools. From these, they took into account those that have been reported as successfully used at least once for annotating biomedical facts in biomedical documents. The 13 resulting tools were deeply analyzed with respect to predefined criteria, that include the scope of supported annotations, the possibility to pre-annotate texts and the usability of the interface[29] They do, among others, a review of brat [241] and Callisto [68, 69], the annotation tools used by us.

Annotation projects in the biomedical domain. Some projects have been carried out to annotate biomedical texts. Next, we present some of them. We put focus on the entities and attributes similar to ours and on some decisions taken, that we also had to take. Some details about the annotation schema, annotation decisions and ontologies used to do pre-annotations are presented in order to be able to compare them with our annotation schema and the decisions made by us.

Bozkurt et al. [29] developed an annotation schema to annotate named entities. Therefore, 35 radiology reports corresponding to mammographies were annotated. The annotation schema is composed of entities -anatomical entities and imaging observation, among them-, modifiers and relationships. Only complete noun phrases and adjective phrases were marked as annotation, for example *right breast* was marked as anatomical entity.

[27] Amazon's Mechanical Turk https://www.mturk.com/mturk/welcome (accessed Jun. 2017).

[28] CrowdFlower https://www.crowdflower.com/ (accessed Jun. 2017).

[29] An abstract of the paper can be seen in: https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/resources/annotationtools (accessed Jun. 2017).

The SHARP and THYME projects of the University of Colorado at Boulder developed semantic annotations in the clinical domain for radiology and pathology notes. A number of annotation guidelines are being developed in these projects. In the THYME project[30] biomedical texts were pre-annotated with cTAKES (introduced in Section 2.4.3). Annotations were based on UMLS Metathesaurus. Various entity types are annotated, anatomical entities and clinical findings among them. The annotation of embedded entities is allowed. For instance, the phrase *renal cell carcinoma* should be annotated as *[renal cell]*(AE) and *[renal cell carcinoma]*(FI). Also overlapping annotations are carried out. For instance, the phrase *right lower leg swelling caused by edema*, should be annotated as *[right lower leg swelling]* (SI),[31] *[right lower leg]* (AE), *[leg swelling]* (SI) and *[edema]* (FI). Terms to be annotated include: *conditional* and *history of* indicators (similar to our *temporal terms*) and a *negation* and an *uncertainty indicator*. All relations should be contained within the same sentence. Examples of one of the relations, the *location of* relation, can be seen in Section B.2 of Appendix B. Relations also have attributes. Some of them are *conditional*, *negation* and *uncertainty*.

The 2010 i2b2/VA Workshop on Natural Language Processing Challenge for Clinical Records, introduced in Section 2, presented a concept extraction, an assertion classification and a relation classification task for de-identified discharge summaries and progress notes written in English. A series of annotation guidelines used to annotate the reports employed for performance evaluation of the systems were presented. The assertion annotation guideline[32] instructs to classify each medical problem into one of six assertions categories (present, absent, possible, conditional,[33] hypothetical[34] and not associated with the patient[35]). Assertions are time independent (i.e., a problem experienced in the past can be in the same category as a problem existing in the present). Annotation guidelines for concepts[36] and for relations[37] were also provided. A summary of the provided instructions, that we believe can be useful for the creation of annotation guidelines in the biomedical domain, can be seen in Section B.2 of Appendix B.

Sun et al. [247] describe the process done to annotate temporal expressions and temporal relations in discharge summaries. They also provide the developed annotation guideline.

[30] THYME (Temporal Histories of Your Medical Event) project: http://clear.colorado.edu/compsem/documents/umls_guidelines.pdf, https://clear.colorado.edu/TemporalWiki/index.php/Main_Page (both accessed Jun. 2017).

[31] *SI* refers to sign or symptom.

[32] 2010 i2b2/VA challenge evaluation assertion annotation guidelines https://www.i2b2.org/NLP/Relations/assets/Assertion%20Annotation%20Guideline.pdf (accessed Jul. 2017).

[33] *Conditional* means that the assertion says that the patient experiences the medical problem only under certain conditions, for instance: *patient reports dizziness after standing up.*

[34] *Hypothetical* refers to the mention of problems the patient may develop. For instance: *if he experiences shortness of breath (...)*. It is equivalent to our *conditionals.*

[35] *Not associated with the patient* refers to medical problems associated to someone who is not the patient, for instance: *mother had arrhythmia.*

[36] 2010 i2b2/VA challenge evaluation. Concept annotation guidelines https://www.i2b2.org/NLP/Relations/assets/Concept%20Annotation%20Guideline.pdf (accessed Jun. 2017).

[37] 2010 i2b2/VA challenge evaluation. Relation annotation guidelines https://www.i2b2.org/NLP/Relations/assets/Relation%20Annotation%20Guideline.pdf (accessed Jun. 2017).

3.6 Discussion

As reported before, abbreviations and acronyms are often used in our dataset. Considering that there are 7,035 anatomical entities and findings (see Table 3.5), and that there are 477 abbreviations and acronyms of anatomical entities and findings, we can think that about 6% of the anatomical entities and findings are written in an abbreviated way. Also, there is a total of 105 different abbreviations in 513 reports (see Table 3.5), which we consider a high number.

The other results of Table 3.5 are useful for a posterior use of this annotated dataset. Multi-name terms will probably not be easily recognized by standard entity recognition algorithms. The relation of findings with temporal expressions (*not present* relation shown in Table 3.6) should be taken into account to determine the factuality of a finding. The same occurs with terms that denote negation and uncertainty. Relations between sentences will be also difficult to discover.

As depicted in Table 3.7, the inter-annotator agreement improves in each annotation-revision iteration step. This makes sense, since after each annotation iteration many meetings were held with both annotators to solve doubts and the annotation criteria was changed according to new questions the annotators asked until it stabilized. With this stabilized annotation guidelines, annotators performed the annotation of dataset 3, which had an inter-annotator agreement of 0.89.

We do not have an objective measure related to how the annotation easiness increased after the pre-annotation process. Even though, annotators reported that after some improvements of our manually built dictionary (occurred after some *annotation-automatic improvement of the dictionary* iterations) pre-annotations were much more accurate and, thus, their annotation was much easier. We also noticed an increase of the reports annotated per hour.

Considering Tables 3.5 and 3.6, we can see that 56% of the findings are negated (1,478 out of 2,637). This might lead to future implementation of methods to detect negated findings in reports (see [43], and Chapter 5 of this thesis, that handles with negation detection for Spanish and for German). Only 1.25% of the findings are reported as a past issue or as a conditional issue in the future (33 *not present* relations out of 2,637 *findings*).

The development of the annotation criteria has not been an easy task. New entities (e.g. location) had to be added to the initial annotation schema. The need to add these entities came from the actual annotation process and the questions that the annotators had. The initial set of relations grew also much more than expected due to the complexity of some of the sentences that revealed the existence of relations that were not considered initially. The schema grew more complex after the two initial iterations.

In many cases it was not easy to determine if a concept belonged to an entity type or to another. In particular, we found that a location can be referring to an anatomical entity and an anatomical entity to a location. Many doubts of this kind arose and helped us to define the definitive annotation guidelines (Section 3.3.2).

These facts led to the presence of some errors and inconsistencies in the annotations, that are to be corrected before publishing the dataset. The presence of this issues coincides with the usual presence of inconsistencies in annotations, especially in the biomedical domain that was introduced in the first section and that is referred in [255, 238]. Some of the issues are: “retroperitoneo vascular” and “via biliar intra hepática” are annotated as abbreviations, while they are not; “via biliar” has been wrongly omitted as an AE in at least one report; “vesicula” (gallbladder) has at least once been annotated as finding; the phrase “Ambos riñones” (both kidneys) has been sometimes annotated as AE and sometimes only “riñones” has been annotated as AE and the annotation of "dilatada" as a finding has been missed in more than one occasion.

Our annotation schema and process agree with THYMEs' in having a pre-annotation step, in being based on UMLS concepts and in the annotation of conditional, historic, negation and uncertainty indicators. THYME differs in that it includes attributes to the relations and also accepts embedded and overlapping entities.

Our annotation schema shares almost all the assertion indicators with the 2010 i2b2 challenge annotation guidelines schema. The indicator *not associated with the patient* is added in the last one. With regards to the annotation of entities, in our case for *pain in the chest*, we would annotate *[pain](FI) in the [chest](AE)*, while i2b2 challenge annotation proposal is that a concept to be annotated can include up to one prepositional phrase (PP) following it if the PP indicates a body part or can be rearranged to eliminate the PP. As *pain in the chest* can be rearranged to *chest pain*, *[pain in the chest](FI)* would be annotated by them.

3.7 Conclusions

In this chapter we presented the annotation criterion we developed for a set of radiology reports written in Spanish with the goal to be able to use the annotated corpus as an evaluation resource for name entity recognition, negation detection and relation extraction and as input for the training of supervised learning methods to solve these tasks. We divided the total available set in four subsets in order to be able to extract, in the future, relations among the data that give further information (for instance, the existence of appendicitis), that might be useful for physicians and patients. We anonymized data, created and published annotation guidelines and trained the annotators to do the annotation task.

The shortness of the texts, the abundance of acronyms and abbreviations and the specificity of the medical language made the annotation task difficult. Furthermore, it was not easy to keep up with the goal to achieve a simple annotation criterion.

The analysis of the annotated dataset shows some interesting characteristics, as the abundance of negated findings. That might lead to the development of negation detection algorithms. This annotated dataset is useful for its evaluation.

We noticed the importance of having annotators with expertise in the annotation task and in the medical domain and we consider that in this particular domain it is even more difficult than in others to obtain annotations from experts.

As future work we plan to make our dataset publicly available, after doing some improvements.

3.8 Resumen

La anotación de un texto está definida como el agregado de metadatos que marcan[38] elementos del mismo [210].

Para evaluar algoritmos de extracción de información y para entrenar modelos de aprendizaje supervisados es necesario contar con corpus anotados.

Sin embargo, pese a su importancia, en el dominio biomédico es muy difícil obtenerlos. Especialmente en idiomas distintos al inglés. La falta de corpus anotados por expertos se considera como uno de los principales obstáculos para desarrollar métodos de minería de textos novedosos [194].

[38] Denominamos *marca* al término *tag* del inglés.

El proceso de anotación requiere tener definidos ciertos criterios. Estos deben ser diseñados cuidadosamente y revisados en un proceso iterativo, de forma tal de que sean lo suficientemente claros como para que distintos anotadores puedan anotar los datos con un alto grado de acuerdo.[39] En particular en el dominio biomédico hay un alto grado de diferencias en los criterios de anotación entre distintos anotadores [270] y hasta en los de un mismo anotador. Esto se debe principalmente a la incompletitud de los criterios de anotación, muchas veces originada por la naturaleza compleja de los textos, a la posibilidad de asignarles más de una marca a las entidades y a errores humanos [255, 238]. Algunas diferencias comunes entre anotadores son los límites izquierdos y derechos de las entidades (por ej. un anotador puede asumir que la entidad a anotar es “proteína distrofina”, mientras que el otro asume que sólo “distrofina” es la entidad) y la clasificación de las mismas. Las inconsistencias en la anotación pueden afectar el entrenamiento y la evaluación de técnicas de aprendizaje automático y el resultado de otras técnicas, que la utilizan para su evaluación. Una alternativa a la anotación es la utilización de técnicas de *supervisión a distancia*, que suelen utilizar una base de conocimientos externa. Estas técnicas reducen el esfuerzo en crear los corpus anotados, pero introducen ruido.

Para poder aportar a la medicina mediante la provisión de técnicas de IE que permitan detectar hallazgos explícitos, implícitos, determinar si son fácticos y en qué parte del cuerpo se encuentran, requerimos de corpus anotados. Según nuestro mejor saber y entender no existen corpus anotados disponibles públicamente de informes clínicos escritos en español para detección de entidades, especulaciones, ni relaciones.[40] Por este motivo trabajamos en la creación de un corpus anotado de informes radiológicos escritos en español para detección de entidades, negación, especulación y extracción de relaciones.

En este capítulo describimos los datos con los que trabajamos, su proceso de selección y de anonimización. Luego explicamos el proceso seguido para la anotación, el esquema de anotación, las dificultades encontradas y las decisiones tomadas, de forma tal de que puedan reutilizarse por otros investigadores que trabajan en temas similares. Presentamos también un análisis del corpus anotado.

El proceso de anotación resultó ser muy complejo, debido al uso de términos altamente especializados en los informes médicos y a la gran cantidad de abreviaturas y acrónimos existentes en los textos. En el análisis realizado al corpus anotado se detectó, entre otros, una gran cantidad de términos negados.

[39] El grado de acuerdo entre anotadores es medido con el *inter-annotator agreement* (IAA).

[40] Recientemente, en 2017, se publicó un corpus para detección de negaciones [165].

CHAPTER 4

Named entity recognition

In this chapter we present with further detail the named entity recognition problem, its importance and its challenges in the biomedical domain. We explain the evaluation methods used for named entity recognition and we introduce the metrics that will be used to evaluate our NER implementations. We then present a review of previous work in the area, including challenges in the biomedical domain. We present our proposal of two methods to automatically detect anatomical entities and clinical findings in radiology reports written in Spanish. The first, called SiMREDA, is thought for cases where there is scarcity of linguistic resources and of annotated corpora. The second, a CRF based method, can be used in cases where there exists annotated corpora. We evaluate both methods and draw conclusions. We also propose a classification method among reports containing affirmed findings and reports not containing them.

4.1 Introduction

As introduced in Chapter 2, named entity recognition is an information extraction task, whose goal is to identify instances of specific kind of information units in text and assign them a class. Additionally, a score designating the confidence that an expression is of a given class can be given. It has been applied to different textual genres and domains and to different entity types.

The term named entity recognition (NER) was introduced in the Sixth Message Understanding Conference (MUC-6) [107], in a task that involved recognizing names of people, organizations, and geographic locations, time, currency and percentage expressions in well-written texts,[1] such as newswire. Afterwards, it began being applied to other domains, such as the biomedical, for identifying genes, proteins, drug names and diseases, among others.

The approaches to solve the NER problem include: dictionary-based, rules-based, statistically based and combined approaches [50, 229].

As we previously mentioned, the biomedical domain has specialized terminology and a lot of abbreviations and non-standardized naming conventions (so, unseen words usually appear) and there is no standard, even among specialists, regarding

[1] We say that a text is *well-written* if it is carefully composed and does not have abundance of orthographic, grammatical or syntactic errors.

to which is the boundary of an entity. All these situations make the named entity recognition problem more difficult in the biomedical domain than in the general domain. In addition, following situations, that highlight the challenges of NER task in the biomedical domain are described by Cohen and Hersh [50] and Leaman and Gonzalez [150]:

- the absence of complete dictionaries for some biological or medical named entities and the fact that new entities are added frequently,
- the same word or abbreviation might refer to different concepts depending on the context (ambiguity and polysemy),
- there might be different ways of referring to the same entity, and
- the fact that biological and medical entities may have multi-word names, so there is a need to determining name boundaries and resolving overlap of candidate names. As mentioned in [150], it is easier for a system and for a human to determine if an entity is present or not in a text than to determine its boundaries.

Additionally, there is no standard criteria in the evaluation of biomedical named entity recognition systems. Not only the boundary of named entities, but also their class is often ambiguous, due to criteria differences among specialists. Therefore, different matching criteria have been used for Bio-NER system[2] evaluations. Furthermore, datasets are usually not published due to confidentiality issues. Accordingly, usually gold standards have to be generated. The lack of standard metrics, of publicly available datasets and of standard annotation criteria makes the comparison of different implementations difficult.

Much of the work in biomedical NER has focused in the recognition of gene and protein names in formal texts and for English [50]. Less work has been done for the medical domain and for languages other than English.

The processing of medical reports in languages other than English, such as Spanish adds a further difficulty, since there are less resources available. Our research questions are:

- how well do dictionary-based techniques for NER detection work in the biomedical domain in Spanish?
- does the use of rules improve dictionary-based techniques?
- if there are no terminology resources available in the language of the reports, is the quality of the translations important or are there ways to overcome translation problems?
- does SNOMED CT work better than RadLex for our purpose?[3]
- does the study of word morphology improve results?
- is the exact match an appropriate way of measuring results or are results based on partial matching more desirable?
- is it possible to work with supervised machine learning techniques with a small annotated dataset?

We describe in this chapter different approaches we implemented in order to detect anatomical entities and clinical findings in a set of Spanish radiology reports.

Our goal is to identify all named mentions of a specific type of object. We are not addressing the identification of entities referred by pronouns or nominally.

[2] Bio-NER refers to biomedical named entity recognition systems.

[3] SNOMED CT and RadLex were introduced in Chapter 2. RadLex is a terminology specific for the radiology domain and does not exist in Spanish. SNOMED CT has general clinical terms and exists in Spanish.

We work in the recognition of two types of entities: anatomical entities (AEs) and clinical findings (FIs). The recognition of these entities is useful because: a) it enables the possibility to structure the information, b) it offers the opportunity to detect relations among findings and anatomical sites where they occurred [201], c) if negation is taken into account, identifying which reports contain clinical findings could allow the indexing of only relevant documents and discard those which are not relevant (do not contain clinical findings). This is as a classification task and can serve for the purposes of identifying later on, which are the specific occurrence of clinical findings in the relevant reports and d) it could serve to notify physicians about the findings, some of which could require immediate action. The obtention of timely information is critical in case of urgent or important findings [24, 37, 23]. The automatic detection of critical issues, such as appendicitis and pyloric stenosis, is of interest and is being studied (see e.g. [78, 183]) and could allow their communication by pager or alternatives methods, as [149] describe.

To detect entities, we propose and evaluate two different approaches: 1) SiMREDA, a Simple Entity Detection Algorithm for Medium Resource languages, that is based on a lookup of terms from a specialized vocabulary, on morphological knowledge and on knowledge of PoS tag patterns of anatomical entities and clinical findings, and that was conceived by us and 2) a machine learning approach. At the beginning of our research we had a small annotated dataset (of approximately 200 reports), that did not allow us to apply machine learning techniques, due to its reduced size. We worked in the development of SiMREDA. Meanwhile, we achieved to annotate a set of 513 reports (see Chapter 3). So, later we were able to develop the second approach, based on conditional random fields (CRF).

As explained in Chapter 2 and summarized in Table 2.2, there exist different ontologies, terminologies and coding systems in the medical domain like MeSH, ICD-10, LOINC, UMLS, SNOMED CT and RadLex.

As mentioned, RadLex is a lexicon of radiology terms written in English. It has specifically been developed to satisfy standardized indexing and retrieval of radiology information. It satisfies the needs in this domain by adopting features of existing terminology systems as well as producing new terms to fill critical gaps. However, there is no radiology ontology or machine-readable dictionary data that can be used to identify terms that denote anatomical entities and clinical findings of the radiology domain in Spanish. A direct automatic translation from English ontologies present a number of difficulties:

- some terms are frequently used in Spanish with synonyms, that are less frequently used in English. For example, *arteria mamaria interna* for *internal mammary artery* is commonly used in Spanish, while in English it would be referred as *internal thoracic artery*
- sometimes terms in Spanish are preferred in an adjectival way rather than as a noun. For example, *folículo ovárico* for *ovarian follicle* is commonly used, while in English *follicle of ovary* is the preferred term, and
- terms of interest can be composed of more than one word, which often leads to problems in the order of the translated words.

For SiMREDA algorithm we use a translation of RadLex as the lexicon for the detection of anatomical entities and clinical findings. We evaluate some variants taking into account improvements in RadLex translation and since there are studies that detect that SNOMED CT, that also exists in Spanish, covers some radiology concepts [5], we also test our algorithm using SNOMED CT instead of RadLex as an information source. Finally, we improve our results adding modules of syntactic analysis, based on PoS tag patterns detected in a subset of annotated reports and of morphological knowledge, through the recognition of Graeco-Latin morphemes, that compose a great number of biomedical terms (such as *itis* in *hepatitis*). Morphological analysis helps us detect and understand specific terms that do not appear in terminological resources.

For the CRF implementation we test different sets of features, one proposed by us and others already existing.

We use the same dataset to test our algorithms. Therefore, we use a portion of the dataset, whose annotation was described in Chapter 3. The other part of the dataset is used for training the CRF implementation.

The rest of the chapter is organized as follows. Section 4.2 presents the criteria used in the evaluation of NER systems, Section 4.3 presents related work, specific NER resources for the biomedical domain, NER challenges in general domain in languages other than English and NER challenges in the biomedical domain. Then, in Section 4.4 we present our proposed methods: SiMREDA and its different variants and modules and CRF with different features. We also specify the pre-processing done to the data, how we retrieved anatomical entities and clinical findings from RadLex and from SNOMED CT and we briefly describe the technical details of our implementations. We also present a classification task performed by us in Section 4.4.5. After that, Section 4.5 presents results of both NER methods, analysis of each of them and a comparison among them. Also, a comparison with previous works is done. Finally, Section 4.6 describes our conclusions and future work that could be carried out and Section 4.7 provides an abstract of the present chapter in Spanish.

The chapter includes part of the content of following co-authored publication: [55].

4.2 Evaluation of NER systems

First, we will present two ways of representing named entities. Then, we will present different criteria used in evaluating NER systems.

Representation of named entities

We make reference to named entities in two ways:

- **delimited by opening and closing tags**, where entities are preceded by an opening tag and followed by a closing tag (both labeled with the type of the entity -*FI* for findings and *AE* for anatomical entities-) and where multi-word entities are written between brackets. For example, "the <FI>cyst</FI> is in the <AE>[upper part of the liver]</AE>." indicates that *cyst* is an entity of type clinical finding and that *upper part of the liver* is an anatomical entity, and

- **IOB or IOB2 format**, a way of representing the boundaries of an entity and of showing how different evaluation metrics are calculated.

The IOB2 format contains two columns, each separated by a single space. The first column is composed by each word of the sentence, and the second by a syntactic chunk tag.[4] The chunk tag has the format I-TYPE which means that the word is inside a phrase of the type TYPE. The first word of each phrase has the content

[4] There are alternative representations of IOB2 format, for example having three columns, where the first corresponds to the words, the second to their PoS tags and the third to its syntactic chunk tag.

B-TYPE. A word with tag O is not part of any phrase.[5] We can adequate the IOB2 format, replacing syntactic chunks by named entity categories.

For example, "the <FI>cyst</FI> is in the <AE>[upper part of the liver]</AE>." would be written in IOB2 format as follows:
the O
cyst B-FI
is O
in O
the O
upper B-AE
part I-AE
of I-AE
the I-AE
liver I-AE
. O

IOB2 and IOB format differ that in the second, not always chunks begin with a B-type. We will use IOB2 format but will call it indistinctly IOB or IOB2.

Criteria in evaluating NER systems

In named entity recognition systems, often the evaluation is performed per entity and not per token. That means that TP, FP, FN and TN metrics (see Table 2.3 in Section 2.5) are calculated on an entity level and not on a token level.[6]

As we previously mentioned, in the biomedical domain there is little agreement, also among the annotators, about which are the boundaries of an entity.[7] E.g. if a human annotator tags *dystrophin protein* as a protein, and a system tags *dystrophin* as a protein, then if an exact match is considered for evaluation purposes, a false negative (for not having discovered *dystrophin protein*) and a false positive (for having discovered *dystrophin*) would be generated, although the term dystrophin has been detected. So, results would be worse than if the term would have not been tagged.

Therefore, we can say that exact match is a too strict criterion in some cases. It could be sufficient to know that a specific protein, gene or clinical finding is mentioned in a sentence [87, 255].[6] Christopher Manning explains in an informal publication why standard measures (precision, recall and F1, all based on exact match) are not necessarily the best way to evaluate named entity recognition systems and describes three type of errors (besides FP and FN).[8] He also explains that some entities would be rather missed in order to avoid counting two errors,[9] but that users would prefer systems retrieving those entities than those systems, that would miss them. An exact match evaluation would encourage systems not to tag entities.

Hence, there are approaches, other than the exact match, that assign partial (or total) credit for matches with boundary errors (sometimes called *lenient or approx-*

[5] The description of the IOB2 has been taken almost textually from the Software and Data section of the conll2003 challenge https://www.clips.uantwerpen.be/conll2003/ner/ (accesssed Jun. 2017).

[6] Stanford University course of NLP (Coursera). Unit 9 - 2 - Evaluation of Named Entity Recognition- Dan Jurafsky & Chris Manning https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html (accessed Jan. 2018).

[7] This problem is not exclusive of the biomedical domain.

[8] Doing Named Entity Recognition? Don't optimize for F1, http://nlpers.blogspot.com.ar/2006/08/doing-named-entity-recognition-dont.html, Chris Manning. (Accessed Jan. 2017).

[9] The two errors refer to the issue explained in previous paragraph.

imate matching). Some of them are MUC and ACE challenges evaluation metrics, *right-match* and *left-match*. How much credit to give to partial matches in these cases is also a studied question. Esuli and Sebastiani [87] describe different criteria proposed by various studies. In some cases, gold standards are annotated with different boundaries. In Section 4.3 a brief abstract of surveys performed by other authors about approximate matching techniques is presented.

Consider following example for a better understanding of the different error types, inspired on an explanation of Manning[6] and on a different example provided by Nadeau and Sekine [189]. Assume following gold standard: <FI>[enlarged esophagus]</FI>. A <FI>cyst</FI> has been observed in the <AE>[upper part of the liver]</AE>. <FI>Edema</FI> in the <AE>lungs</AE>, and the output obtained by a NER system: <FI>enlarged</FI><AE>esophagus </AE>. A <FI>cyst</FI> has been <FI>observed</FI> in the upper part of the <FI> liver</FI>. Edema in the <FI>lungs</FI>. The system produces five errors and an exact match. All are explained in Table 4.1. The description includes the name given by Manning and by MUC evaluation standards to the different type of errors.

id	**correct solution**	**system output**	**description**
1	observed	<FI>observed</FI>	The algorithm discovers a non-existent entity (FP)
2	<FI>Edema</FI>	Edema	The algorithm misses an existent entity (FN)
3	<AE>lungs</AE>	<FI>lungs</FI>	The algorithm detects an entity, but assigns it the wrong label (*text* according to MUC, *labeling error* according to Manning)
4	<FI>[enlarged esophagus]</FI>	<FI>[enlarged]</FI>	The algorithm detects an entity, with the right label but with wrong boundaries (*type* according MUC, *boundary error* according to Manning)
5	<AE>[upper part of the liver]</AE>	<AE>liver</AE>	The algorithm detects an entity with wrong labels and wrong boundaries (label-boundary error according to Manning)
6	<FI>cyst</FI>	<FI>cyst</FI>	The algorithm detects an entity, with right label and right boundaries (TP)

Table 4.1: Type of errors and matches in named entity recognition.

4.3 Related work

In this section we present some surveys carried out for the NER task in the general and the biological domain and different solutions implemented for the NER task. We also describe work carried out for recognizing neoclassical morphemes and some NER resources for the biomedical domain. Finally, we mention NER challenges in the general domain mainly for languages different than English and NER challenges in the biomedical domain.

Previous surveys

A number of surveys have been carried out on the NER task. Various address the biological domain. Nadeau and Sekine [189] present a survey of research in the NER field from 1991 to 2006. They describe supported languages, textual genres and entity types, algorithmic techniques proposed, and the evaluation criteria used in different challenges. The Journal of Biomedical Informatics published two special issues: *Named entity recognition in biomedicine* [12] and *Current issues in biomedical text mining and NLP* [41]. Both address the NER task in the biological domain. Sondhi [238] makes a survey of NER in the biological domain addressing the challenges in this domain with respect to the general domain, the available resources and features used for machine learning methods involved in the resolution of the task. Finally, Tasneem and B [253] review NER in the biological domain.

NER development

Most NER research has been developed for English. Nevertheless, also some work has been carried out for other languages, such as German, Dutch, Japanese, Chinese, French and many others [189]. Spanish has been introduced in CoNLL-2002 and MET-1 events.

Usually efforts in NER are dedicated to a specific genre and domain. Not much work has been devoted for NER in diverse genres and domains. In their study, Nadeau and Sekine [189] give an overview of works dedicated to different genres and domains and also reference previous studies, such as [206], which demonstrated that it constitutes a major challenge to port a system to a new domain or textual genre since there is a drop in performance when there is a change in genre and domain.

An overview about the evolution of different entity types to be detected can also be read in [189].

The initial approaches for NER were dictionary and rule-based. Dictionary based techniques look for the appearance of terms belonging to terminologies in the texts. Sometimes inexact string matching [256] is carried out in order to improve performance. Rule-based techniques use domain knowledge or information obtained through analysis of a subset of the data. Syntactic parsing and the composition patterns of the named entities (PoS pattern, orthographic patterns) are examples of information that can be used to build rules. Rule-based methods usually have good results [261], but its construction is time consuming and often not reusable in other datasets.

Statistical methods are also used for NER. They are sometimes combined with dictionary or rule-based techniques [20]. Machine learning (ML) methods can be supervised, for which a considerable amount of training data is needed, semi-supervised, as bootstrapping [257], or unsupervised.

Among the supervised methods, there are classification-based and sequence-based approaches. Examples of the first are Naive Bayes (NB) and Support Vector Machines (SVM) [139, 252]. Sequence-based approaches consider sequences of words instead of individual words or phrases considered in the classification-based approaches. Some examples include Hidden Markov Models (HMM) [227] and Conditional Random fields (CRF).[10] CRFs were the best performing systems in various challenges and have been highly ranked in others (2010 i2b2, BioCreAtIve gene mention and JNLPBA bio-entity recognition) [229]. Some implementations can be seen in [224, 22, 21]. Different features used for these methods are described in [238]. Many HMM and CRF approaches used are reported in [238, 189, 229].

Many semi-supervised methods for NER in the general domain are reviewed in [189].

Unsupervised learning methods are typically based on clustering. Methods are usually based on lexical resources and on large corpus of statistics taken from unan-

[10] Conditional random fields, CRF, are defined in Section 4.4.4.

notated texts. Some unsupervised learning methods for the general domain are reviewed in [189]. A research to recognize NER in tweets (noisy data) in an unsupervised way can be seen in [86].

Weegar et al. [277] examine the impact of feature engineering in order to improve the baseline of different models, such as CRF, SVM or neural networks in the clinical NER task for Spanish, English and Swedish. They try different window sizes,[11] with and without the transformation to lower case, the addition of prefixes and suffixes, the addition of features regarding words only formed by capital letters and number types (e.g. only digits and digits with hyphen), lemmas, POS tags and features based on SNOMED tags. They recommend trying different window sizes, prefixes and suffixes of lengths three and four, and conclude that there are important differences on the impact of features with respect to each language. They work, among others with Spanish with a set of 121 manually annotated texts, that has 3,362 diseases and 1,406 drugs and test features with a perceptron.

Recently, Roller et al. [214] implemented a CRF algorithm for the detection of entities in medical reports written in German. They use a feature set proposed in CLEF 2015 by Jiang et al. [130] for texts in French, that will be presented later in Section 4.3.

Regarding NER in Spanish biomedical texts, Castro et al. [40] implement a tool similar to UMLS MetaMap Transfer (MMTx)[12] for the identification of Spanish SNOMED CT terms corresponding to SNOMED CT *procedures* and *disruptions* hierarchies in Spanish clinical notes. The tool is tested with 100 clinical notes. An inverted index is used and a score is assigned to the retrieved terms, depending on the length of the query with respect to the retrieved terms. It is integrated with MOSTAS [128], a tool that normalizes abbreviations and acronyms, anonymizes reports and corrects spelling errors.

Santiso et al. [220] present a NER for Spanish EHRs with the goal to access their factuality with a NegEx implementation. They try different techniques. Their best result consists in a CRF implementation, tested with 75 electronic health reports annotated with an IAA of 90.53%. As features they use four characters prefixes and suffixes and transform terms to lower case. They consider entities that overlap as partial match.

Oronoz et al. [197] use Freeling-Med, presented in Section 2.4.3, as a way to automatically tag named entities. They test it with 20 clinical reports looking for diseases, drugs and substances. They achieve high F1s, but use following extremely loose matching criteria: “two elements are considered to be equivalent if an element given by the system is entirely contained within an extension of a manually tagged element by six positions both to the left and to the right”.

Table 4.2 shows the results for Spanish NER in the medical domain.

Knowledge sources

Aleksovski [5][13] does a test annotating 381,000 radiology reports with RadLex and SNOMED CT terminologies. His goal is to determine if RadLex could be extended including some SNOMED CT terms. He discovers that there are many medical terms relevant to radiology that are missing in RadLex.

[11] Window sizes refers to the number of words that are taken into account.

[12] MetaMap Transfer has the same functionalities as MetaMap, but some technical differences. Some of them can be seen in `https://ii.nlm.nih.gov/Publications/Papers/09.08.20.MetaMap-MMTx.updated.ppt` (accessed Jan. 2018).

[13] Not available online. Personal communication.

paper	**# reports**	IAA	**P**	**R**	**F1**	**doc. types**	**ent. types**
Castro et al. [40]	100	66%	0.43 (0.72*)	0.06 (0.09*)	0.11 (0.16*)	CN	DRP (SN)
			0.35 (0.70*)	0.07 (0.55*)	0.06 (0.10*)		PR (SN)
Santiso et al. [220]	75	90.53%	0.36 (0.70*)	0.45 (0.83*)	0.40 (0.76*)	EHR	DS
Oronoz et al. [197]	20	-	- (0.97**)	- (0.80**)	- (0.88**)	ClR	DS
			- (1.00**)	- (0.96**)	- (0.98**)		DR
			- (0.84**)	- (0.92**)	- (0.88**)		SB

Table 4.2: NER results for Spanish in the medical domain. References on the type of documents: ClR: clinical reports, CN: clinical notes, EHR: electronic health reports. References on the entity types: DS: diseases, DRP: disruptions, DR: drugs, PR: procedures, SB: substances, and SN: SNOMED CT. Other references: doc.: documents, ent: entities. First results correspond to exact matches. Results marked with (*) correspond to lenient matches and (**) to extremely loose lenient matches.

Neoclassical Morphemes

In many languages, including Spanish [136], Graeco-Latin morphemes are used in medicine, biology and other health-related science disciplines. Furthermore, a small number of Graeco-Latin morphemes can generate a large amount of terms [136, 233, 271]. In particular, Vivaldi and Cabré [271] explain that among 30% and 40% of the monolexical medical terms[14] in biomedical texts written in Spanish include at least a Graeco-Latin morpheme. This number is much larger than in technical domains such as computer science.

Zweigenbaum and Grabar [292], Ananiadou [9, 10] and Vivaldi [269] use the existence of Graeco-Latin morphemes for different stages of NLP in biomedical texts. Ananiadou [10] introduces a morphological grammar and a lexicon for handling neoclassical compounds of terms. Zweigenbaum and Grabar [292] present a method for automatically acquiring morphological knowledge applicable to different languages. Vivaldi and Cabré [271] describe a module based in Graeco-Latin morphemes that uses a finite automaton and a dictionary of morphemes and their related information to segment words and recognize the morphemes included in them. Vivaldi [269] uses neoclassical morphemes knowledge and term segmentation to improve a term extraction system in the medical domain.

NER resources in the biomedical domain

In the next paragraphs we present a brief overview of the NER component of biomedical text mining support tools. For those systems introduced in Sections 2.4.3 and 2.3.2, we only give details of the NER component.

LingPipes named entity recognition module is based on a supervised machine learning algorithm, on an exact and a partial dictionary matching algorithm, and on a regular expression-based approach (the last two techniques are useful for entities that are completely listed in dictionaries or for whom a regular expression can be written).

[14] *Monolexical terms* are those terms composed by only one word.

ABNER NER system for biological entities (genes, proteins, cell lines, cell types, RNA and DNA) is based on conditional random fields and uses orthographic and contextual features [224, 225]. It was trained on BioCreative and NLPBA 2004[15] shared tasks annotated data.

BANNER NER system, primarily thought for biomedical text, is based on CRF and uses features based on those published in the biomedical NER literature [150]. It does not employ semantic features nor rule-based processing steps for the sake of doing it domain independent.[16]

DNorm is an automated NER and normalization tool for diseases detection in biomedical texts [151]. DNorm achieved the best performance in the 2013 ShARe/-CLEF shared task on disease normalization in clinical notes.

Apache cTAKES$^{\text{TM}}$ (Clinical Text Analysis and Knowledge Extraction System) NER module uses a dictionary look-up algorithm within a noun-phrase look-up window, that takes into account non-lexical variations [222]. A subset of UMLS, including disorders/diseases, signs/symptoms and anatomy, UMLS synonyms and a Mayo clinic-maintained list of terms, are used as dictionary. 160 EHR from the Mayo Clinic were annotated for evaluating the NER task. The results for the UMLS semantic group disorder are 80.1% (88.9%) precision, 64.5% (76.7%) recall and 71.5% (82.4%) F1. Results among parenthesis correspond to lenient metrics.

Table 4.3 shows the comparison of BANNER, ABNER and LingPipe in the NER task on two different corpora: the training corpus of the BioCreative II GM task [234] and the diseases of the BioText disease-treatment corpus,[17] introduced in Chapter 2. The table was extracted from Leaman and Gonzalez [150].

corpus	BioCreative II gene mention task (Training set)			BioText disease/treatment (diseases only)		
system	**P (%)**	**R(%)**	**F1**	**P (%)**	**R (%)**	**F1**
BANNER	85.09	79.06	81.96	68.89	45.55	54.84
ABNER	83.21	73.94	78.30	66.08	44.86	53.44
LingPipe	60.34	70.32	64.95	55.41	47.50	51.15

Table 4.3: Results of BANNER, ABNER and LingPipe on the BioCreative 2 gene training set and on the Biotext diseases, as shown by Leaman and Gonzalez [150].

Events dedicated to the NER task in the general domain mainly in languages other than English

As previously mentioned, MUC-6 introduced the task of NER for names of people, organizations, geographic locations, time, currency and percentage expressions in 1995 for English newspaper articles from the Wall Street Journal [107]. In 1996 the Multi-lingual Entity Task (MET-1) was organized. Its goal was to identify named entity expressions of name of persons, locations, organizations date and time

[15] NLPBA: Natural Language Processing in Biomedical applications. The NLPBA corpus is a modified version of the GENIA corpus [140].

[16] BANNER: http://banner.sourceforge.net/ (accessed Jan. 2018).

[17] The BioText disease-treatment corpus, previously introduced, is taken from abstracts and titles of MEDLINE (correctly written text).

name	**entity type**	**language**	P (%)	R (%)	F1 (%)
MUC-6[a]	PER, LOC, ORG, DATE, TIME, NUM	EN	97	96	96.49
MET-1 [168]	PER, LOC, ORG, DATE, TIME	SP, CH, JP			SP 93.04, CH 84.51, JP 92.12
MUC-7[b][46]	PER, LOC, ORG, DATE, TIME, NUM	EN			94
MET-2[c][46]	PER, LOC, ORG, DATE, TIME	CH, JP			CH 91, JP 87
CONLL 2002[d] news articles [218, 219]	PER, LOC, ORG, MISC	SP	81.38	81.40	81.39
CONLL 2002. news articles	PER, LOC, ORG, MISC	DU	77.83	76.29	77.05
CONLL 2003. news articles	PER, LOC, ORG, MISC	EN	88.99	88.54	88.76+-0.7
CONLL 2003. news articles	PER, LOC, ORG, MISC	GR	83.87	63.71	72.41+-1.3
TAC EDL 2015	PER, GPE, ORG, LOC, FAC	SP, EN, CH	SP 83.4 EN 83.4 CH 85.1	SP 77.2 EN 74.0 CH 77.7	SP 79.9 EN 76.1 CH 79.9

[a] MUC-6 http://www.aclweb.org/anthology/C96-1079 (accessed Mar. 2018).

[b] MUC-7 http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/overview.html (accessed Mar. 2018).

[c] MET-2 http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/overview.html (accessed Mar. 2018).

[d] CONLL-2 http://delivery.acm.org/10.1145/1120000/1118877/p24-tjong_kim_sang.pdf?ip=181.90.59.99&id=1118877&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=948168262&CFTOKEN=46269260&__acm__=1497905948_cb2678b74e558f66d3b6438f0b9946c2, http://www.aclweb.org/anthology/W03-0419.pdf (both accessed Mar. 2018).

Table 4.4: NLP challenges summary in NER in general domain. References: entity type DATE: dates, FAC: facility, GPE: geopolitical entity, LOC: locations, NUM: monetary amounts and numbers associated to percentage, ORG: organizations, PER: persons, MISC: entities that are not PER, LOC or ORG, and TIME: time. Languages: CH: Chinese, DU: Dutch, EN: English, GR: German, JP: Japanese, and SP: Spanish. Results correspond to exact matches.

in Spanish, Chinese and Japanese news articles [14, 168]. MET-2 was organized in 1997, with the same goal as MET-1 for Chinese and Japanese. MUC-7 repeated MUC-6 task [46].[18],[19]

The CoNLL-2002 and CoNLL-2003 shared tasks handled with the extraction of named entities in Spanish and Dutch (2002) and English and German (2003). Named entities considered were persons, locations, organizations and entities not belonging to any of the before mentioned entity types in newspaper articles [218, 219].

ACE-2, ACE 2003 and ACE 2004 studied, among others, NER for Chinese, Arabic and English [80].

Table 4.4 shows the results of events dedicated to the NER task in the general domain in different languages.

[18] MUC-7 NER task: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html. MUC-7 overview of results: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/marsh_slides.pdf (both accessed Oct. 2017).

[19] An overview of tasks descriptions and results from MUC-3 through MUC-7, MET-1 and MET-2 can be senn in http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/overview.html (accessed Oct. 2017).

Previous challenges results in NER in the medical domain

Competitions are a way to improve the state of the art, specially in the biomedical domain, due to the lack of publicly available gold standards. In the next paragraphs we present the main challenges held in the NER task in the medical domain. As mentioned, a visualization of biological vs. clinical problems, challenges organization between 2002 and 2014 and challenges subtasks according to the different NLP areas can be seen in [124]. Table 4.5 presents a summary of the best scores obtained in each of the challenges of the medical domain shown in Table 2.4.

2010 i2b2/VA challenge on named entities, assertions and relations in clinical text. The 2010 i2b2/VA challenge had three subtasks. For the first, named entities of *problems*, *tests* and *treatment* had to be extracted from discharge summaries and progress notes.

Almost all named entity recognition systems evaluated in the 2010 i2b2/VA challenge used conditional random fields. Some were fed with the output of a rule-based NER system. Others used CRF with an ensemble with different algorithms. The best result implemented a modification of a HMM, called semi-Markov model using lexical features, including context and length of sentences and sections. More details can be seen in [70].

Evaluations were done considering an exact match and a partial match (with correct type but incorrect boundaries).

SEMEVAL 2014 Task 7. Analysis of clinical text. The goal of the task is to use supervised methods for named entity, abbreviation and acronym recognition and normalization -mapping to UMLS CUIs- in clinical notes. The task is composed of two subtasks: subtask a, that proposes NER of concepts that belong to the semantic group *disorders* of UMLS and task b, that deals with the mapping of the *disorder* mentions to a unique UMLS CUI, that belongs to SNOMED CD.

The ShARe corpus, introduced in Section 2.4, was used. It contains clinical reports from MIMIC II database, manually annotated for disorder mentions and normalized to UMLS CUIs, when possible.

CLEF

CLEF-ER - Entity Recognition @ CLEF 2013. As mentioned in Chapter 2, the ER task consisted in the automatic annotation of named entities and normalization to CUIs in English, French, German, Spanish, and Dutch corpora. The corpus was composed by the documents that later were part of the QUAERO corpus (introduced in Section 2.4). Texts in Spanish stemmed from EMEA corpus.

ShARe/CLEF eHealth evaluation lab pilot 2013 had two subtasks: 1a) NER of diseases and disorders in clinical reports and 1b) their normalization against a concept in SNOMEDC CT, that belongs to the *Disorder* semantic group (composed of one of the following UMLS semantic types: *Congenital Abnormality*, *Acquired Abnormality*, *Injury or Poisoning*, *Pathologic Function*, *Disease or Syndrome*, *Mental or Behavioral Dysfunction*, *Cell or Molecular Dysfunction*, *Experimental Model of Disease*, *Anatomical Abnormality*, *Neoplastic Process* or *Signs and Symptoms*. The ShARe corpus was used [208]. Exact and approximate matches were evaluated.

Clef eHealth Evaluation Lab 2015 Task 1. Consisted in IE from texts written in French with subtasks 1) NER and 2) entity normalization.[20] The input

[20] CLEF 2015 Task 1-bhttps://sites.google.com/site/clefehealth2015/task-1/task-1b (accessed Mar. 2018).

texts were MEDLINE titles and EMEA documents from the QUAERO corpus [196]. MEDLINE documents are short and comprise two sentences at most. Each EMEA document has several hundred sentences. Both kind of documents have a formal, well-written language. Entities had to be extracted according to following UMLS semantic groups: *Anatomy*, *Chemical and Drugs*, *Devices*, *Disorders*, *Geographic Areas*, *Living Beings*, *Objects*, *Phenomena*, *Physiology* and *Procedures*. Normalization was also done with regards to these semantic groups. For subtask 1, only exact matches were taken into account. Subtask 2 also took partial matching into account.

Methods used by participants included the use of ML techniques (CRF with the use of lexical resources as features) and algorithms that did not use the training corpus and that relied on lexical resources (medical terminologies and ontologies) and on translation software. The authors of the algorithm with the best results expanded the coverage of the French UMLS with the automatic translation of English UMLS terms into French with Google Translate and Microsoft Bing. The corpus was then indexed with Peregrine, a dictionary-based concept recognition system, developed by the authors. Finally, many post-processing steps were applied [2].

Jiang et al. [130] presented a CRF implementation with good results called WiENRE. They propose lexical, morphological and orthographic features. They achieve 75.9 and 54.6 F1 in percentage for AE and FI respectively. More information about CLEF 2015 Task 1b can be seen in [190].

Clef eHealth Evaluation Lab 2016 Task 2. Consists in Multilingual IE, that, as the 2016 edition, deals with NER in French biomedical texts (scientific articles and drug inserts, among others). For the 2016 edition a new corpus of death reports (CépiDC Causes of Death Corpus) was added to the QUAERO corpus used in 2015. The new corpus consists of free-text descriptions of causes of death as reported by physicians. The task, for this corpus is to extract the causes of death and is considered as a text classification task.

Entities of the QUAERO corpus have to be recognized based on the same UMLS semantic groups as in 2015. The CépiDC corpus has to be mapped to ICD-10 codes. Two subtasks were organized: 1) NER, 2) named entity normalization. For subtask 1 only exact matches were taken into account. Subtask 2 also took partial matching into account. The CépiDC task consists of extracting ICD10 codes from the death certificates.

Methods used vary mainly from machine learning techniques (CRF, LDA,[21] SVM) to the use of lexical resources (medical terminologies and ontologies, including the training data as additional knowledge source) combined with indexing methods. Statistical machine translation was used by some participants to address the limitation of French lexical resources. The authors of the best algorithm of subtask 1, the same that had the best results in 2015, expanded the terminology used with the 2016 training data [263]. For more information on Task 2 of the 2016 CLEF eHealth evaluation lab see [192].

Evaluation of biomedical named entity recognition

The evaluation of term extraction systems is not standardized. There are many studies that refer to ways to evaluate NER systems. Vivaldi and Rodríguez [270] describe the difficulty of comparing different implementations due to the lack of evaluation standards and introduce many possible evaluation methods. Nadeau and Sekine [189] explain the criteria used in the main challenges for general domain NER

[21] LDA: Latent Dirichlet allocation.

name	corpus	lang.	P	R	F1	doc. types	entity types
ShARe/CLEF 2013 Task 1a[a]	ShARe	EN	0.80 (0.93*)	0.71 (0.83*)	0.75 (0.87*)	ClR	DISE, DISO (SN)
CLEF 2015[b]	QUAERO	FR	0.71	0.62	0.66	PT, DO	DISE and ANAT (UMLS)**
CLEF 2016 EMEA[c]	QUAERO	FR	0.63	0.78	0.70	PT, DO	DISE and ANAT (UMLS)**
CLEF 2016 MEDLINE[d]	QUAERO	FR	0.61	0.69	0.65	PT, DO	DISE and ANAT (UMLS)**
2010 i2b2/VA[e]		EN	0.87	0.84	0.85 (0.92*)	DS, PN	PROB, TEST, TREAT
SEMEVAL 2014 Task 7 run 1[f]		EN	0.84 (0.94*)	0.79 (0.87*)	0.81 (0.90*)	CN	DISO (UMLS)

[a] ShARe/CLEF 2013 Task 1a. https://sites.google.com/site/shareclefehealth/ (accessed Mar. 2018).
[b] CLEF 2015 http://www.clef-initiative.eu/ (accessed Mar. 2018).
[c] CLEF 2016 EMEA http://www.clef-initiative.eu/ (accessed Mar. 2018).
[d] CLEF 2016 MEDLINE http://www.clef-initiative.eu/ (accessed Mar. 2018).
[e] 2010 i2b2/VA https://i2b2.cchmc.org/faq#data1 (accessed Mar. 2018).
[f] SEMEVAL 2014 Task 7 results: https://docs.google.com/spreadsheets/d/1yE8cQSOK3LhRRlZblcqwUT7MFh8nWepAlW5BM6W0pCo/edit#gid=2094125 (accessed Mar. 2018).

Table 4.5: BioNLP challenges results in the NER task in the medical domain. References on the languages: EN: English, FR: French. References on the type of documents: ClR: clinical reports, CN: clinical notes, DS: discharge summaries, PN: progress notes, PT: paper titles, DO: other types of documents. References on the entity types: ANAT: anatomy, DISE: diseases, DISO: disorders, PROB: problems, TEST: tests, TREAT: treatments and SN: SNOMED CT. Lang. refers to language and doc. to documents. Results correspond to exact matches. ** entity types include DISE and ANAT, among other UMLS categories. Results marked with * correspond to lenient matches.

systems (MUC[22], CoNLL[23] and ACE[24]). In particular, in MUC a system is scored taking into account its ability to find the correct type (called *type*) and its ability to find exact text (called *text*) [47] (see Table 4.1).

Tsai et al. [255] describe and evaluate various criteria in the evaluation of biomedical NER, such as left match, right match, left-right match, partial match, and approximate match, among others. Jiang et al. [131] propose two methodologies to evaluate the most well known named entity recognition systems at that time: exact match and a partial match, that is counted when there are boundary errors, but the

[22] MUC-7 Named Entity Task Definition http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html (accessed Nov. 2017), Nancy Chinchor.

[23] CoNLL, Conference on Computational Natural Language Learning http://www.conll.org/2017 (accessed Nov. 2017).

[24] ACE, Automatic Content Extraction Program https://www.ldc.upenn.edu/collaborations/past-projects/ace (accessed Nov. 2017).

detected type is correct (*boundary errors* or *MUC type errors* in Table 4.1). Esuli and Sebastiani [87] state that there is a lack of agreement on an evaluation measure for IE systems and propose a sophisticated metric to evaluate it. Dlugolinsky et al. [77] also evaluate many NER tools, with two evaluation methods: strict and lenient matching. Dingare et al. [76] discuss the effect of variability in annotation criteria in system performance. In some cases, entities have alternative annotations (with different boundaries). In these cases, they are considered a true positive if there is an exact match with one of the alternative annotations.[25]

4.4 Methods

In this section we will explain which radiology reports we used for training (when it applied) and for testing purposes, the preprocessing applied to reports, and the lexicons used. The SiMREDA algorithm and its variants and the conditional random field algorithm and its feature selection are presented next. Then, we explain the exact match and a lenient matching evaluation metrics used and how they work. Finally, we present the technical details of our implementation.

In the next section we will present the data used and the pre-processing steps done.

4.4.1 Data

We worked with the 513 annotated reports, whose characteristics, selection and anonymization process were described in Section 3.2. We will explain the preprocessing done to the reports and the datasets generated to test results and to train the CRF method.

Pre-processing

In Table 3.2 we showed that the percentage of accentuated vowels in our 513 annotated radiology is more than ten times smaller than the number of accentuated vowels in a corpus of abstracts of scientific articles of the medical domain written in Spanish with the same amount of words. Besides, the terms obtained by RadLex translation to Spanish also lack many accents.

To uniform the text, a decision was taken to remove accents and to change the ñ by n. It is very usual not to take into account accents and special characters. A drawback of this decision is that in some cases the PoS tagging results, that we will use later in our methods, are incorrect. For example, *líquido* (*liquid*, a noun) is written as *liquido* (*to finish*, *to complete*) and interpreted as a verb by the Freeling PoS tagger trained for Spanish texts with accents.

We also normalized our reports transforming every word to lowercase.

Datasets

The reports, whose annotation was described in Chapter 3, are going to be used to train and to test an entity recognition algorithm (CRF) and to test SiMREDA, the NER algorithm developed by us. Both NER algorithms are going to be evaluated with the same dataset.

For machine learning methods a portion of the data has to be used for training and adjustment of the model (we will call it the *development dataset*) and another

[25] Alternative annotations can be seen in `http://biocreative.sourceforge.net/biocreative_2_gm.html` (accessed Nov. 2017).

portion of the data for the testing of the developed model (the *testing dataset*). Therefore, we will partition our annotated dataset into two sets. The development dataset with 80% of the annotated reports and the testing dataset with the remaining 20% of the reports.

We will evaluate different features for the CRF. Therefore, for each feature set, reports belonging to the development dataset are going to be used with a 5-fold cross-validation (that is 80% of the development set will be used for training and the remaining 20% will be used for validation). This will be performed five times and an average of the results will be considered as the result for each set of features. Finally, the whole development dataset will be used as training set with the best features, selected by the previously described method. See Figure 4.1 for a visual explanation. More details about the dataset used for CRF can be seen in Section 4.4.4.

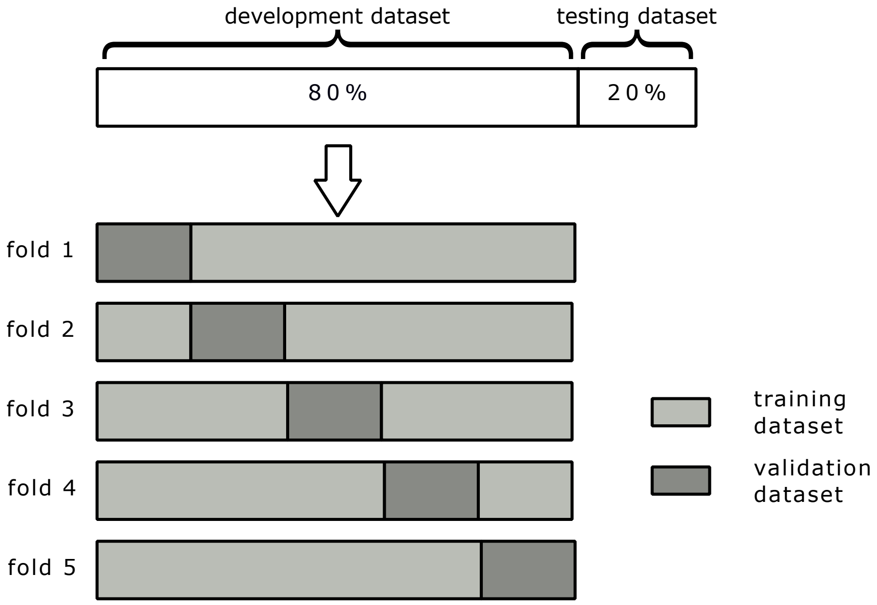


Figure 4.1: Datasets preparation.

For the analysis of PoS tag patterns in SiMREDA algorithm 20% of the development dataset will be used.

Finally, the testing set of 20% of the 513 reports will be used to test the CRF and SiMREDA algorithms. This dataset has not been used to train nor to infer patterns of the data.

4.4.2 Lexicons

In this section we present the details of the use of RadLex and SNOMED CT, introduced in Chapter 2, that will be the two sources of information used for the SiMREDA algorithm. Section 4.4.3 describes how the information sources are used in SiMREDA algorithm.

RadLex

We use RadLex, the RSNA ontology specific to the radiology domain, as the lexicon for the detection of anatomical entities (AE) and clinical findings (FI). RadLex, has different versions. After analyzing them, we decided to use version 3.6, which has more than 30,000 terms. We processed data in order to obtain only terms referring to clinical findings and anatomical entities (see Section 4.4.7 to view the details of how they were obtained).

As we previously mentioned, to the best of our knowledge, there is no complete RadLex translation to Spanish (the translation mentioned in Castilla et al. [39] is partial and not every term is precise and was refused to use by a specialist of the radiology domain).

In order to be able to use RadLex with texts written in Spanish, we had to obtain a translated version. All RadLex terms were translated to Spanish with Google Translate.[26] Later, this translation was improved by a physician of the radiology domain. We tested our algorithm with both: the Google Translate translations and its improvement. We considered also:

- a mapping of RadLex to UMLS terms, and through UMLS we obtained the translation to Spanish of the mapped terms, and

- a mapping of English-Spanish Wikipedia terms, and through an exact search of RadLex single word terms in the English Wikipedia terms, we obtained their Spanish translation.

Table 4.6 shows the number of terms translated using different translation sources. Of the more than 30,000 RadLex terms, 10,357 correspond to anatomical entities and to clinical findings and were translated with Google Translate. 972 anatomical entities and clinical findings translations (almost 10%) were corrected by the physician. 628 of them correspond to anatomical entities and 344 to clinical findings. 857 anatomical entities and findings were obtained by the UMLS translation and 896 through Wikipedia.

In Section 4.4.3 we will explain how we used different translations in our tests.

source of translation	**# of anatomical entities and clinical findings**
RadLex - Google Translate	10,357
RadLex -Google Translate improved	972
UMLS	857
Wikipedia	896

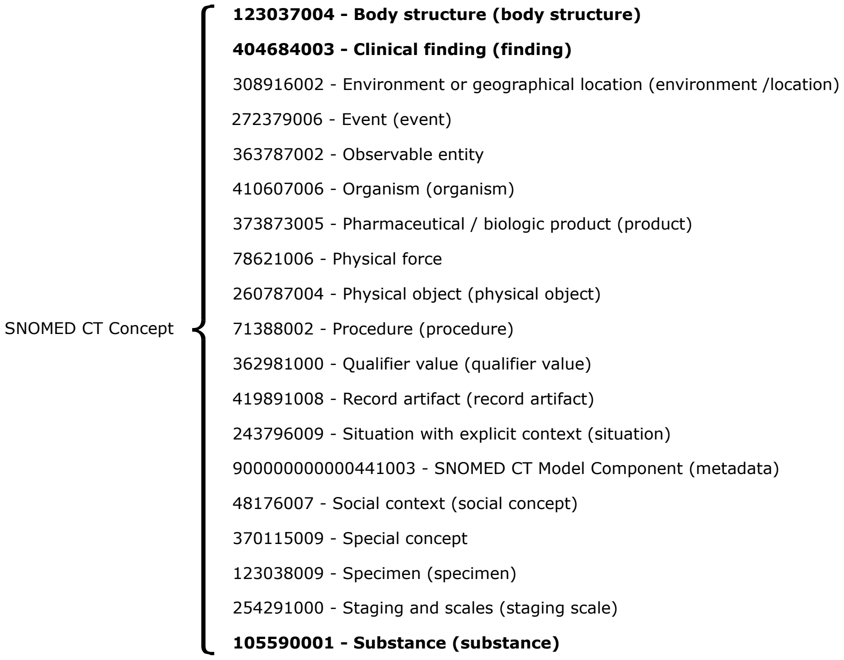
Table 4.6: Number of English-Spanish RadLex translated terms. The rows refer to the source of the translation and the second column refers to the number of translated terms, that correspond to clinical findings or to anatomical entities. For instance, there were 972 anatomical entities and findings, whose translation has been reviewed by the specialized physician and 857 anatomical entities and findings that were translated through UMLS.

[26] Google Translate. https://translate.google.com/ (accessed Mar. 2018).

SNOMED CT

As mentioned in Chapter 2, SNOMED CT[27] has 19 hierarchies. In order to extract anatomical entities and clinical findings of SNOMED CT (SN) Spanish edition we proceeded in two different ways.

For the first, we considered SNOMED CT hierarchies *Clinical Findings*, *Body Structure* and *Substance*, and retrieved part of them in order to obtain SNOMED CT findings and anatomical entities that meet our definition of those entities (presented in Section 2.2). Figure 4.2 shows SNOMED CT hierarchies retrieved.



SNOMED CT Concept

123037004 - Body structure (body structure)
404684003 - Clinical finding (finding)
308916002 - Environment or geographical location (environment /location)
272379006 - Event (event)
363787002 - Observable entity
410607006 - Organism (organism)
373873005 - Pharmaceutical / biologic product (product)
78621006 - Physical force
260787004 - Physical object (physical object)
71388002 - Procedure (procedure)
362981000 - Qualifier value (qualifier value)
419891008 - Record artifact (record artifact)
243796009 - Situation with explicit context (situation)
900000000000441003 - SNOMED CT Model Component (metadata)
48176007 - Social context (social concept)
370115009 - Special concept
123038009 - Specimen (specimen)
254291000 - Staging and scales (staging scale)
105590001 - Substance (substance)

Figure 4.2: SNOMED CT hierarchies used for selecting clinical findings and anatomical entities. The 19 SNOMED CT hierarchies are shown. Only the bold-faced hierarchies are considered for selecting FIs and AEs.

For the second, we built a different set of findings by using a subset of SN terms, called *CORE Problem List Subset*. Thus, we performed two tests: one considering as clinical findings most terms under the *Clinical Finding* category and some terms of *Substance* and *Body Structure* categories (we will call it *Our findings*) and the other considering as clinical findings, those terms that appear in the CORE subset. In both cases we considered the same set of anatomical entities.

In the next paragraphs we describe how we retrieved AEs and *our clinical findings* and we explain what the CORE subset, used as an alternative set of findings, is.

Our selection of anatomical entities and of *our clinical findings* was performed as follows. Among the categories *Clinical Finding* and *Body Structure* we found some terms that we believed corresponded to a category different to the one it was depending from (taking into account our definition of AEs and FIs described in Section 2.2). In those cases, we changed the category according to our opinion. For instance, *Body structure, altered from its original anatomical structure*, that is listed

[27] SNOMED CT Spanish edition browser can be found in: http://browser.ihtsdotools.org/?perspective=full&conceptId1=404684003&edition=es-edition&release=v20171031&server=https://prod-browser-exten.ihtsdotools.org/api/snomed&langRefset=450828004 (accessed Mar. 2018).

in SNOMED CT as *Body structure*, was taken by us as a clinical finding. There are also some subcategories, that we consider do not correspond to the parent categories. These categories and the decisions taken are described below.

Administrative statuses are not under our definition of finding, thus we did not consider this category. There are some sub categories that in our criteria in part correspond to the father category and in part not, for instance: under *fetal finding* (descendant of *finding*), *fetus normal* and *fetus present* are not considered findings by us. *Fetal state* contains some terms considered finding by us, and other terms not considered findings by us, and *fetal problem* contains findings -according to our criteria-. Anyway, we considered all *fetal category* descendants as findings. *General clinical finding* has also some descendants that we believe that do not correspond to finding (e.g. *gender finding*). Nevertheless, we took *General clinical state finding* and all its descendants as findings into account. Regarding the *substance* category. One of its descendants was considered as clinical findings by us (*cancer-related substance*) and some others (*biological substance, body substance and material)*) as anatomical entities. Tables 4.7 and 4.8 show SNOMED CT identifiers (SCTID), SN concepts, and the description of those concepts that we considered as *our findings* and those that we considered as anatomical entities. All descendants of the categories are also taken into account with the same category as their parents. Terms marked with (*) contain descendants that we consider that do not correspond to the main category, but that were anyway taken as if they would correspond to it. Table 4.9 shows the number of anatomical entities retrieved: 40,241 and the number of clinical findings retrieved with core the dataset (6,357) and with our dataset (144,060).

The *Clinical Observations Recordings and Encoding (CORE) Problem List Subset* contains a subset of UMLS terms useful for encoding clinical information in a summarized way. It was generated from the analysis of datasets collected from seven health care institutions, that use controlled vocabularies for data entry. This list of terms is mapped to SN concepts and codes and constitutes the CORE Problem List Subset of SNOMED CT.[28] Its main purpose is to facilitate the use of SN as the primary coding terminology for problem lists and thus maximize data interoperability among different institutions. The CORE subset contains terms from *Clinical finding*, *Procedure*, *Situation with explicit context* and *Events* SN hierarchies. In the cases of *Clinical finding* hierarchy concepts that are very similar, only one of them (the disorder concept) is selected.[29]

4.4.3 SiMREDA algorithm

We implemented SiMREDA, a Simple Entity Detection Algorithm for Medium Resource languages. As its name suggests, the algorithm is useful for the detection of named entities for languages with limited lexical resources.

SiMREDA has three modules and some variants. The basic module consists in a lookup of terms that come from a specialized vocabulary through the use of an inverted index. As specialized vocabulary we try two alternatives: RadLex, specific of the radiology domain, but that had to be translated into Spanish and SNOMED CT, that is not specific of the radiology domain, but exists in Spanish.

[28] Those terms of SNOMED CT that are considered useful for the problem list and are not mapped to its terms are sent to IHTSDO (International Health Terminology Standards Development Organisation) consideration and eventually added to the CORE subset.

[29] Information taken from The CORE Problem List Subset of SNOMED CT https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html (accessed Mar. 2018).

SCTID	**SNOMED CT concepts**	**description**
61321005	Substance	Cancer-related substance
118956008	Body structure	Body structure, altered from its original anatomical structure (morphologic abnormality)
405533003	Finding	Adverse incident outcome categories
131148009	Finding	Bleeding
313413008	Finding	Calculus finding
250171008	Finding	Clinical history and observation findings
80631005	Finding	Clinical stage finding
3415004	Finding	Cyanosis
417893002	Finding	Deformity
64572001	Finding	Disease (disorder)
79899007	Finding	Drug interaction
267038008	Finding	Edema
419026008	Finding	Effect of exposure to physical force
424017009	Finding	Enzyme activity finding
247441003	Finding	Erythema
441742003	Finding	Evaluation finding
106112009	Finding	Fetal finding (*)[a]
118234003	Finding	Finding by site
384740007	Finding	Finding of grade
300475002	Finding	Finding of measures of urine output
127357005	Finding	Finding related to physiologic substance
418799008	Finding	Finding reported by subject or history provider
365860008	Finding	General clinical state finding (*)
18165001	Finding	Jaundice
102957003	Finding	Neurological finding
443871003	Finding	Papule
365858006	Finding	Prognosis/outlook finding
271587009	Finding	Stiffness
65124004	Finding	Swelling
225552003	Finding	Wound finding

[a] Not all the descendants are findings. For example: Fetal state contains some terms that constitute findings according to our definition and others that do not (e.g. fetus normal and fetus present).

Table 4.7: SNOMED CT concepts considered as clinical findings. Column SCTID shows SNOMED CT Identifiers. Concept descendants were also considered as findings. Terms marked with (*) contain descendants that we consider that do not correspond to the main category, but that were anyway taken as if they would correspond to it.

As previously explained, we tried different ways of translating RadLex into Spanish and different ways of obtaining SNOMED findings. Since among 30% and 40% of the monolexical terms written in Spanish can be formed by a small number of Graeco-Latin morphemes [271], their lookup can help discovering clinical findings that do not appear in the lexicons, that are not correctly translated to Spanish or that are not well written in reports. Thus, the second module considers the appearance of those

SCTID	**SNOMED CT concept**	**description**
442083009	Body structure	Anatomical or acquired body structure
91832008	Body structure	Anatomical organizational pattern
258331007	Body structure	Anatomical site notations for tumor staging
115668003	Substance	Biological substance
91720002	Substance	Body substance
260769002	Substance	Material

Table 4.8: SNOMED CT concepts considered as anatomical entities. Column SCTID shows SNOMED CT Identifiers. Concept descendants were also considered as anatomical entities.

set	**number of anatomical entities**	**number of clinical findings**	**total number of entities**
core subset	40,241	6,357	46,598
our subset	40,241	103,819	144,060

Table 4.9: Number of anatomical entities and clinical findings retrieved from SNOMED CT (with core data set and with our retrieved findings listed in Table 4.7).

morphemes. Finally, usually anatomical entities and clinical findings satisfy certain PoS tagging patterns. For example, we discovered in a subset of our annotated reports that 56.25% of the anatomical terms beginning with a noun continue with an adjective, that is also considered part of the anatomical term (e.g. *testículo izquierdo* -*left testicle*- and *pared abdominal* -*abdominal wall*-). We analyze the PoS tag patterns present in the previously mentioned subset of our development dataset and look for these patterns in the radiology reports in order to improve SiMREDA results. This constitutes module 3. A graphical image of SiMREDA algorithm, its modules and variants can be seen in Figure 4.3.

In the rest of this section we will explain each module and variants.

SiMREDA module 1: inverted index

The algorithm uses the terms of RadLex referring to anatomical entities and to clinical findings as a base to determine if a term appearing in a radiology report refers to an AE or to a FI.

We translate to Spanish all RadLex anatomical entities and clinical findings (different translations are tested: a translation obtained through Google Translate and its improvement done by a physician of the radiology domain, and translations obtained through UMLS and Wikipedia, see Section 4.4.2). Each word appearing in the translated terms is added to an inverted index (defined in Section 2.2). Stopwords are not included in the inverted index. Each entry of the inverted index points to the RadLex terms where it appears and gets the class assigned (anatomical entity or clinical finding), that is most frequent in the RadLex terms where it is used. In the case that there is the same quantity of anatomical terms and clinical findings that contain the word, then a *manual decision* is made as whether it corresponds to an AE or to a FI. For example, *pieloureteral* (pyeloureteral) and *subdural* (sub-

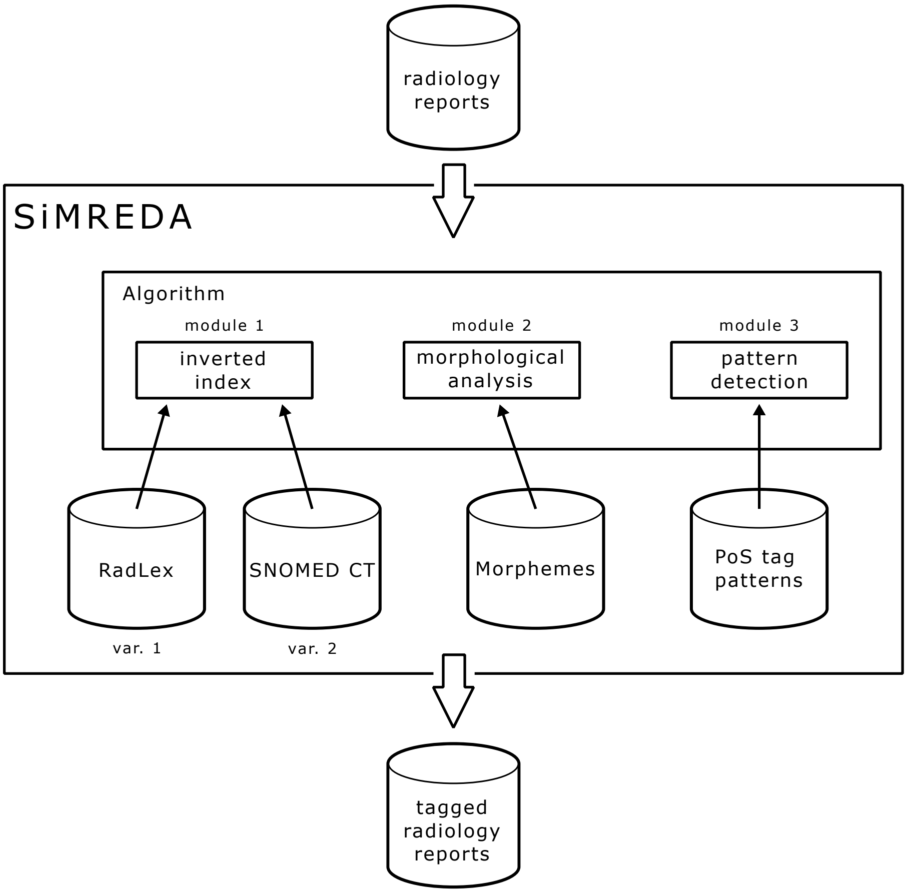


Figure 4.3: Schema of SiMREDA algorithm. Its modules and variants.

dural), that are adjectives referring to locations are taken as anatomical terms and the word *nodulo* (*-nodule-*) as a clinical finding. See Table 4.10 for an example of an inverted index of RadLex terms translated to Spanish.

The use of an inverted index is useful because:

- RadLex terms are usually composed by many words that do not necessarily appear all and in the same order in the reports,
- there are some words that should be detected and that appear as part of a RadLex term, but not as a RadLex term by itself (for example: *vessel* does not appear as RadLex term but is part of more than 100 RadLex terms (as in *blood vessel*), and
- we can avoid problems derived from the wrong order of words in the translations of phrases.

Those words that appear in the reports and that also belong to the inverted index are tagged as anatomical entities or as findings, according to the class assigned in the inverted index. Adjacent sequence of words belonging to the same class are tagged together with their corresponding class. For example, let's assume we have following text: *se visualiza prolapso de la válvula mitral* (*a mitral valve prolapse has been noticed*). After running the algorithm that tags terms according to their presence in RadLex we would get: "se visualiza <FI>prolapso</FI>de la<AE>válvula</AE><AE>mitral </AE>" if we assume that *prolapso* appears

word	**RadLex terms**	**class asigned**
corazón (heart)	"corazon" (AE), "válvula del corazon" -heart valve- (AE), "enfermedad isquémica del corazon" -ischemic heart disease-(FI), "zona basal del corazón" -basal zone of the heart- (AE), ...	AE
válvula (valve)	"válvula del corazon" -heart valve- (AE), "válvula aórtica" -aortic valve- (AE), "válvula mitral" -mitral valve- (AE), "insuficiencia de la valvula mitral" -mitral valve insufficiency- (FI),..."	AE
insuficiencia (insufficiency)	"insuficiencia de fractura" (FI) -insufficiency fracture-, "insuficiencia de la valvula mitral" -mitral valve insufficiency- (FI), "insuficiencia cardiaca" -heart failure- (FI)...	FI

Table 4.10: Example of inverted index for RadLex terms *heart*, *heart valve*, *ischemic heart disease*, *basal zone of the heart*, *aortic valve*, *mitral valve*, *mitral valve insufficiency*, *insufficiency fracture* and *heart failure* translated to Spanish. The first column has the indexed words. The second column has the RadLex terms, where the words occur, and the third column has the class assigned to the word, that depends on the class of the RadLex terms, where the word appears. The table should also have entries for the words *ischemic*, *disease*, *basal*, *zone*, *aortic*, *mitral*, *fracture* and *failure* (we do not add them because of space constraints). AE corresponds to anatomical entity and FI to clinical finding.

in RadLex more times in terms referring to FIs than in terms referring to AEs and if we consider the class assigned to *válvula* in Table 4.10. Then, if there are contiguous words of the same class (in this case we have *válvula* and *mitral*, both tagged as anatomical entities) we tag them together with their corresponding class. In this case we would get: "Se visualiza <FI>prolapso</FI> de la <AE>válvula mitral</AE>".

As a result, as the algorithm output, we have a set of radiology reports with terms referring to AEs and to FIs automatically tagged according to the translation to Spanish of RadLex anatomical and clinical finding terms.

To improve the results, we also considered a list of common terms referring to findings given by a physician specialized in the radiology domain. We did a dictionary lookup of those terms and tagged them in the reports. Moreover, we run the algorithm with the 79,123 anonymized reports and we analyzed the findings most frequently tagged. Some of them did not appear to identify pathologies, so we created another vocabulary with non-pathological terms and these terms were not longer tagged by our algorithm.

Embedded and multi-labeld entities Our method might produce results with entities embedded into others and multilabeled entities. We took the decision not to allow this.

- **embedded entities:** for entities tagged, that are embedded into larger ones, the largest entity is kept. Consider for example:
  - "<FI><FI Ph>formacion</FI Ph> heterogenea</FI>" -*heterogeneous*

formation- was transformed to "<FI>formacion heterogenea</FI>"[30] and
  - "<FI><FI Ph>apendicitis</FI Ph> aguda</FI>" -*acute appendicitis*- was transformed to "<FI>apendicitis aguda</FI>".

- **multi-labeled entities:** entities that had more than one label assigned were revised by experts on the field and a decision was taken about which was the label to be assigned. Some examples are:
  - *apéndice cecal* -*cecal or caecal appendix*- was tagged as anatomical entity and as a finding of the list of findings suggested by the physician. The decision was to tag it as finding, since the physician considers this term as an indicator of a possible finding (appendicitis).[31]
  - *vena coronaria* -*coronary vein*- was also annotated as anatomical entity and as as finding of the list of findings suggested by the physician. In this case the decision was taken to tag it as an AE.

Module 1 variant 1: Google Translation improved. In this variant the translation of RadLex obtained by Google Translate was changed by the translation improved by the physician of the radiology domain. For those terms, whose translation was not checked, the translation of Google Translate, UMLS and of Wikipedia were taken into account. For more details about RadLex translations see Section 4.4.2.

Module 1 variant 2: Use of SNOMED CT. In this variant, SNOMED CT is taken as an information source instead of RadLex. The advantage is that SNOMED CT is in Spanish, so we avoid the translation problems. The disadvantage is that SNOMED CT is not specific of the radiology domain. Details about how we obtained SNOMED CT anatomical entities and clinical findings were described in Section 4.4.2.

SiMREDA module 2: morphologic analysis

As previously mentioned (see Section 4.3), Graeco-Latin morphemes are used in medical terms of many languages, including Spanish. Even a small number of morphemes of Greek and Latin origin can generate a large amount of terms [136, 233, 271]. The knowledge of these morphemes and their meaning is used to understand many specialized terms and to generate new ones.

We implemented a simple module to detect Graeco-Latin morphemes. Therefore, we compiled a dictionary of morphemes, that includes their type -prefix or suffix- and meaning. The dictionary was built based on a reduced subset of *The reference book of Medical Terminology* [8], that provides a detailed description of suffix, roots, and affixes used in different areas of the medical domain.

Those words, that include morphemes corresponding to findings, in the correct position (as suffix or as prefix) are tagged as findings replacing the tag assigned based on RadLex terms (Module 1). For example, *ascitis* -*ascites*- is not tagged as a finding based on RadLex, but our morpheme detection module detects the suffix -*itis*, so it assumes that *ascitis* is a clinical finding and tags it as such.

[30] FI Ph are the terms indicated as clinical findings by the physician.

[31] It is important to note that the term *apéndice cecal* is actually an anatomical term, whose visualization not necessarily implies the presence of appendicitis. In the institution where the physician works the appearance of this term is usually associated with a positive presence of appendicitis, but that might not be interpreted in the same way by physicians of other institutions or that do not belong to the radiology domain.

Table 4.11 presents some morphemes used in medicine and the number of appearances of them in the 79,123 anonymized radiology reports. In our module we only took into account those morphemes of the *finding* category.

morpheme	**example**	**cate-**	**meaning**	**number of**
				appearances
	Spanish (English)	**gory**		**(distinct)**
-itis	hepatitis (hepatitis)	FI	inflammation	1,859 (63)
-algia	cefalalgia (headache)	FI	ache	3 (3)
-megalia	hepatomegalia (hepatomegaly)	FI	enlargement	4,503 (43)
macro-	macrocefalia (macrocephaly)	FI	enlargement	27 (7)
-ragia	hemorragia (hemorrhage)	FI	effusion	73 (4)
-osis	fibrosis (fibrosis)	FI	disease, pathological process	3,086 (96)
-plejía	apoplejía (apoplexy)	FI	paralysis	0 (0)
-patía	cardiopatía (heart disease)	FI	disease	2,240 (17)
-lito	apendicolito (appendicolith)	FI	stone	865 (6)
-grama	electrocardiograma (electrocardiogram)	PR	drawing	386 (12)
-copía	colonoscopía (colonoscopy)	PR	study, examination	13 (5)
-grafía	radiografía (x-ray)	PR	field of study	5,669 (31)
-logo	cardiólogo (cardiologist)	PRF	specialist	5 (1)
peri-	perivesicular (perivesicular)	LO	surrounds	4,366 (234)
retro-	retrohepática (retrohepatic)	LO	behind	2,760 (108)
supra-	supratiroideo (suprathyroid)	LO	above	7,141 (99)
sub-	submandibular (submandibular)	LO	below	4,753 (158)

Table 4.11: Some Graeco-Latin morphemes related with medical terms and number of appearances in the set of anonymized reports. FI corresponds to clinical finding, PR to procedure, PRF to professional and LO to location.

The detection of morphemes related to the medical domain might also help us improving the dictionary-based approach by detecting terms that are misspelled. For example, *epatitis* for *hepatitis*. Nevertheless, not all the words that contain the previously described morphemes are medical terms (consider, for example, *homologo* -homologous- for suffix *logo*). Furthermore, there are words that contain more than one morpheme related with the medical domain (***peri****ton****itis*** -peritonitis-).

SiMREDA module 3: pattern detection

We selected randomly 20% of our development dataset and used it to analyze the PoS tag sequences of the annotated anatomical entities and clinical findings. We discovered, among others, that more than 50% of the anatomical entities that begin with a noun continue with an adjective, and in many cases only the noun is tagged by our inverted index module. Thus, the discovery of PoS tag patterns of the annotated anatomical entities and clinical findings, and the later revision of the entities tagged by SiMREDAs' Module 1, can help us improve the results obtained by our algorithm. Intuitively, these changes should enhance our exact match more than the partial match.

Tables 4.12 and 4.13 show the most frequent PoS tagging sequences of anatomical entities and clinical findings appearing in the selected subset of the development dataset. Examples of AEs and FIs with each PoS tagging sequence are shown. Also, the quantity of terms with each PoS tag sequence, the percentage of annotated anatomical entities and clinical findings that have this PoS tag sequence pattern and the accumulated percentage of anatomical entities or findings that have the present and precedent PoS tagging sequences of the table are shown. For example, in the *accumulated percentage* column of Table 4.12 it can be seen that 94.84% of the anatomical entities are expressed by terms with PoS tag sequences: NC, NP, NC-AQ, NP-AQ and NC-NC.[32,33] Table 4.13 shows that 9.57% of the findings analyzed by us have the pattern NC-AQ (this means that a common noun was tagged as B-Finding and the next word in the report was an adjective and was tagged as I-Finding and finally, the next word was tagged as O). But we also wanted to know how many of the pairs of words with PoS tags NC-AQ, where the noun is tagged as B-Finding, have also the adjective tagged as I-Finding and the next word tagged as O. So, we studied in our evaluation dataset the percentage of phrases tagged with the PoS tag patterns shown (in this example NC-AQ) that are actually of the entity type listed (finding in this case). The result is shown in column 6 of Tables 4.12 and 4.13. For example, 74% of the NC-AQ cases, where the first term (NC) is tagged as B-Anatomical_Entity correspond to an anatomical entity (i.e. are tagged as B-AE I-AE -O in IOB format). As example: some terms manually tagged as anatomical entities that have NC-AQ PoS tag patterns are *pared abdominal*, *vesícula biliar* and *músculo pilórico -pyloric muscle-*. In the case of findings, with 35 patterns 94% of the cases are considered. With 17 patterns, 90% are considered and with 8 patterns, 81%. We worked with the eight most frequent PoS tag patterns (see Table 4.13).

So, in this module, when we discover in a report a sequence of PoS patterns, belonging to the PoS tag patterns listed in Table 4.12 and whose first word was automatically tagged as an Anatomical Entity, then we tag the whole term (the remaining words that correspond to the pattern) as an anatomical entity (independently of how they were previously tagged based on the used lexicon -RadLex or SNOMED CT-). The same is performed with patterns listed in Table 4.13 for findings. Since all the PoS patterns have a probability greater than 50% of actually being of the corresponding type (see sixth column of Tables 4.12 and 4.13), we consider them all. An example of the application of the knowledge obtained from PoS patterns can be seen below:

The phrase "cambio de la ecogenicidad" (-changes in echogenicity-) would be tagged by the inverted index module as "<FI>cambio</FI> de la ecogenicidad", assuming that *cambio -change-* was classified as FI by our inverted index module. Nevertheless, Module 3 would transform it to "<FI>cambio de la ecogenicidad</FI>", because the first word of the phrase, a NC, (*cambio*) is classified by RadLex as a finding and the rest of the PoS tags are SP (*de*) DA (*la*) and NC (*ecogenicidad*). Rephrasing what we explained before, we have discovered that 60% of the cases where in a pattern of the form NC-SP-DA-NC, the word corresponding to the first NC is tagged as a finding, then the whole phrase (corresponding to pattern NC-SP-DA-NC is a finding) (see Table 4.13). So, we tag the whole phrase matching with

[32] The meaning of Freeling PoS tag sequences can be seen in Section B.3.

[33] NP were incorrectly tagged as they should be NC. There are some other tagging errors, mentioned in the tables.

PoS tag sequence	examples	quantity	perc.	acum. percentage	prob. of PoS tag sequence being an AE
NC	bazo (spleen), psoas (psoas)	383	63.73%	63.73%	1.00
NP	estómago (stomach), rinon (kidney)	73	12.15%	75.87%	1.00
NC-AQ	musculo pilórico (pyloric muscle), pared abdominal (abdominal wall), vesícula biliar (gallbladder)	65	10.82%	86.69 %	0.74
NP-AQ	ovario derecho (right ovary)	39	6.49%	93.18%	0.81
NC-NC	ovario der (right ovary), venas porta (portal veins)	10	1.66 %	94.84%	0.63

Table 4.12: Detected anatomical entity (AE) patterns. The first column has the PoS tag patterns ordered according to the number of AEs that have the patterns (column 3). The second column shows some examples of AEs with the corresponding pattern. Column 4 and 5 show the percentage of annotated AEs that have this pattern and the percentage of annotated AEs that have the previously shown patterns (accumulated percentage). The last column shows the probability that a sequence that has the PoS tags analyzed in the row and whose first word is tagged as an AE is an AE. The analysis is based on the 20% of the development dataset selected for doing this study. NP were incorrectly tagged as they should be NC.

this PoS tag pattern as a finding.

4.4.4 Conditional Random Fields (CRF)

Introduction to the algorithm

Conditional random fields are probabilistic models used to predict sequences of labels, based on sequences of input samples.

A text can be seen as a sequence of tokens. We can say that each token has an associated vector of features, such as the words' part of speech tag, the words' suffix of a given length and an indication as to whether the word is capitalized or not. The input of CRF is the sequence of tokens of the text. The features of a token and the pattern of labels assigned to previous words are used to determine the most likely label for the current token. In linear chain CRF only the label of the previous token is used.

As mentioned in a previous section, CRF have been successfully used for named entity recognition and also for some other natural language processing tasks, such as PoS tagging. Introductions to conditional random fields can be seen in [146, 275,

PoS tag sequence	**examples**	**quantity**	**perc.**	**acum. percentage**	**prob. of PoS tag sequence being a FI**
NC	ovariocele (ovariocele), alteracion (alteration), cambios (changes)	154	33.48 %	33.48 %	1.00
VMP	dilatadas (dilated), engrosadas (thickened)	67	14.57 %	48.04 %	1.00
VMI-AQ	liquido libre (free fluid)	58	12.61 %	60.65 %	1.00
NC-AQ	dilatación pielocalicial (pyelocalicial dilation), hipertrofia pilórica (pyloric stenosis), varices perivesiculares (perivesical varices)	44	9.57 %	70.22 %	0.81
VMP-SP-NC	aumentada/disminuida de tamaño (increased/decreased in size)	21	4.57 %	74.78 %	0.92
AQ	heterogeneo (heterogeneous), bífida (bifid), adenomegalia (adenomegaly)	11	2.39 %	77.17 %	1.00
NP	cavernoma (cavernoma), esplenomegalia (splenomegaly)	10	2.17 %	79.35 %	1.00
NC-SP-DA-NC	incremento de la vascularizacion (increase in vascularization), cambio de la ecogenicidad (change of echogenicity)	9	1.96 %	81.30 %	0.60

Table 4.13: Detected finding (FI) patterns. The first column has the PoS tagging patterns ordered according to the number of FIs that have the patterns (shown in column 3). The second column shows some examples of FIs with the corresponding pattern. Columns 4 y 5 show the percentage of FIs that have this pattern and the percentage of annotated FIs that have the previously shown patterns (accumulated percentage). The last column shows the probability that a sequence that has the PoS tags analyzed in the row and whose first word is tagged as an FI is actually a FI. All is based on the 20% of the development data set, used for doing this analysis. NP were incorrectly tagged as they should be NC. *liquido* is tagged as a verb, while it should be a noun (this happens because the accent is missing, *líquido* is the correct word), *adenomegalia* is tagged as adjective, while it should be a noun, this is probably due to not being included in the Freeling reference dictionary.

249, 52].[34]

For feature engineering an exhaustive search can be done, or attributes can be selected in a greedy-forward or greedy backward elimination fashion, among others.

We are going to try different features proven to work well in different entity recognition tasks and propose our own features.

Dataset

We use the development dataset presented in Section 4.4.1 and Figure 4.1 in order to decide the best set of features. Once we decided it, we used the whole development dataset as training set, and we tested the results with our testing dataset. For more information about the training process see Section 4.5.2.

Performance of CRF depends on the features selection. Our focus is not set in doing feature engineering. Instead, we re-used some previously used features, that will be described next and propose a set of features based on the knowledge of our data. A summary of the features used is given next. For precise details of the features tested see Section B.5 in Appendix B.

Feature set 1. Baseline.

As Feature set 1 we use a case of study presented in CRF ++ for noun phrase chunking as a baseline.[35]

Features include the current word, the current PoS tag, and the context of the word and PoS tag (previous and later tokens and PoS tags). Bigrams are also considered.

Feature set 2. Our proposal.

This feature set is proposed by us.

Features used are: lexical features (lemmas) and reduced PoS tags of current, previous and posterior tokens, and context (lemma of previous and current token, and current and next token; the same for PoS tags); morphological features (4 letters prefix of current, previous and posterior tokens and 4 and 7 letters suffix of the current token), and orthographic features (whether all the characters in the current token are capital letters, the length of the word and whether the current token is formed only by letters, only by numbers by both of them or by none of the above). Morphological and orthographic features are included because they are related with language characteristics of the medical domain. Bigrams are also considered.

Feature set 3. Wi-ENRE

As feature set 3 we test features proposed by Jiang et al. [130] for the solution of CLEF eHealth 2015 task 1b (named entity recognition for French). It was also used by other authors for named entity recognition in German [214].

Features used include: lexical (lower case), morphological (four characters prefix and four characters suffix), reduced PoS tags, orthographic features and shape-related features (length of the token, whether the token begins with a capital letter, whether all its characters are capital letters, whether it contains only digits, only

[34] An informal explanation about CRF can be seen in http://blog.echen.me/2012/01/03/introduction-to-conditional-random-fields/ (accessed Mar. 2018).

[35] CRF++ is an open-source implementation of Conditional Random Fields https://taku910.github.io/crfpp/ (accessed Mar. 2018).

letters or letters and digits). It also takes into account context for morphology features and for PoS tags and uses bigrams.

4.4.5 Classification

In [55] we presented a method for classifying Spanish radiology reports into two sets: the ones that contain affirmed clinical findings (i.e. not negated nor speculated) and the ones that do not. In addition, the entities corresponding to clinical findings were identified in the reports.[36]

Related Work

MoSearch [185], RADTF [79] and Render [67] allow searching for terms in radiology reports taking into account negation and modality information and using NLP techniques. In the last two, results are linked with images from a picture archiving and communication system (PACS). In RADTF, if the user searches for a RadLex term, it returns its RadLex id. Bretschneider et al. [30] use a grammar-based sentence classifier to distinguish among reports with clinical findings and without them. They report 0.54 precision and 0.74 recall. Both are implemented for German and use a German available version of RadLex as linguistic resource. RadMiner adds new terms taken from the annotation performed by a specialist. LEXIMER [82] uses information theory to classify English radiology reports based on the presence or absence of positive findings. They report 0.98 precision and 0.99 recall.

Method and results

For the identification of clinical findings, we used SiMREDAs inverted index module using RadLex Google Translate translation as knowledge base. To detect negation and speculation terms we compiled a list of negations and speculation terms (based on a translation to Spanish of RADTF [79] negations and hedges). These two sets of words were used by a dictionary lookup algorithm to tag these words in the reports. If one term is contained in another we get the largest of the two terms, for example, if *no -not-* and *no se encontró -was not found-* belong to the negation dictionary and *no se encontró* appears in the report we will tag this phrase, rather than the phrase *no*. We classified a report as containing a clinical finding if at least a clinical finding is automatically tagged in the report and no negation and no hedges exists in the sentence where the finding has been detected.

We had a gold standard composed by 248 radiology reports, that were annotated with the Callisto annotation tool [68, 69] by a physician of the radiology domain indicating medical findings. An example of a report annotated in Spanish and its translation to English can be seen below. It can be noticed that the annotation differs from the one presented in Chapter 3 in that the concept of clinical findings is different: it includes the anatomical entities where findings occur and in case of being negated they are not annotated. With our classification method we obtained a precision of 0.63, a recall of 0.83 and F1 of 0.72.

Annotation of radiology reports for classification

33289—16a 4m—20070807—A27611 HIGADO:<FI>lobulo caudado aumentado

[36] A journalistic article of scientific dissemination of the results can be seen in http://nexciencia.exactas.uba.ar/hospital-garraham-diagnostico-imagenes-radiologia-computacion-jose-castano-viviana-cotik-dario-fillipo (accessed Mar. 2018).

de tamano</FI>, resto de higado de ecoestructura conservada. VIA BILIAR intra y extrahepatica: no dilatada. VESICULA BILIAR: alitiasica. Paredes y contenido normal. PANCREAS: tamano y ecoestructura normal. <FI>BAZO: minimamente aumentado de tamano</FI>. Diametro longitudinal:13.5 (cm) RETROPERITONEO VASCULAR: sin alteraciones. No se detectaron adenomegalias. No se observo liquido libre en cavidad. Ambos rinones de caracteristicas normales.

33289 —16y 4m —20070807—A27611 LIVER: <FI> caudate lobe with increased size </FI>, the other lobes of the liver appear normal. Intra and extrahepatic BILIARY TREE: not dilated. GALLBLADDER: no gallstones were seen. Wall and content appear normal. PANCREAS: normal size and echotexture. <FI> SPLEEN: minimally increased in size </FI>. Longitudinal diameter: 13.5 (cm) VASCULAR RETROPERITONEAL COMPARTMENT: unremarkable. No lymphadenopathy was detected. No free fluid in the peritoneal cavity was observed. Both kidneys unremarkable.

Analysis and Conclusions

The results show that there is room for improvement, in particular regarding precision results. Nevertheless, they are promising, considering that we are working with very noisy data, given that terms used to identify clinical findings were obtained through automatic machine translation. We can assume that as a first step to identify reports containing clinical findings, the results are good.

LEXIMER [82] has better results for English and our work has better results than that of Bretschneider et al. [30] for German, but in both cases the results are incomparable, since they have been obtained with different data and for different languages.

As future work, it would be interesting to use the version of RadLex translation that was corrected by the physician and to use NegEx for negation and speculation detection.

4.4.6 Evaluation of biomedical NER systems

We will measure our algorithms with the classical exact match metric and with a lenient (or approximate) match metric, based on the MUC challenge evaluation metric and that scores partial matches (matches with wrong boundary and same entity type) as half of an exact match. In this section we explain how both kind of metrics work.

For the purpose of evaluating our system we transformed our gold standard and the output of the algorithms to the inside-outside-beginning (IOB2) format (see Section 4.2 for an IOB2 format explanation).

To evaluate the performance of the algorithms with the exact match measures of precision, recall and F1 we used the Perl script conlleval.[37] Conlleval tasks use as input a file, that has in each line a word, its PoS tag -that in our case does not matter so it will always have the same label (TAG)-, the tag assigned by the gold standard and the tag assigned by the prediction algorithm, both in IOB2 format. Table 4.14 shows an example of a fragment of report with the human-annotated entities (gold standard) and the algorithm predicted entities in conlleval input format for the phrase *Ambos rinones de ecoestructura normal. Derrame pleural derecho.* (*Both kidneys of normal echotexture. Right pleural effusion.*). It can be seen that there

[37] conlleval script, examples and explanations can be seen in https://www.clips.uantwerpen.be/conll2000/chunking/output.html. (accessed Jan. 2018)

are nine tokens. The gold standard found two entities (*rinones* and *derrame pleural derecho*). The algorithm to be evaluated found three entities (*rinones*, *normal* and *derrame pleural*).

token	word	TAG	annotation tag according to GS	predicted tag according to algorithm
1	Ambos	TAG	O	O
2	rinones	TAG	B−AE	B-AE
3	de	TAG	O	O
4	ecoestructura	TAG	O	O
5	normal	TAG	O	B-AE
6	.	TAG	O	O
7	Derrame	TAG	B-FI	B-FI
8	pleural	TAG	I-FI	I-FI
9	derecho	TAG	I-FI	O

Table 4.14: Example of a fragment of a report in conlleval input format. Conlleval input has a white space instead of tabs. The table is shown with tabs instead of white spaces with the purpose of improving the readability.

The exact match evaluates the correct detection and classification of the complete entity, while the approximate match will consider also as partial correct those matches with boundary errors, only if the detected type is correct (see Table 4.1 in Section 4.2).

Metrics can be calculated per entity type and an overall measure (considering all entities together) can be given. In this example we will consider overall measures instead of considering independent metrics for anatomical entities and for findings.

If **exact match** is considered, there is one true positive (*rinones*), two false positives (*normal* and *derrame pleural*) and one false negative (*derrame pleural derecho*). In an **approximate match** there is one true positive (*rinones*), one partial match (*derrame pleural* has a partial match with *derrame pleural derecho*), and one false positive (*normal*).

Once it is defined if exact match or lenient match are going to be considered, metrics have to be defined. For exact match we will use precision, recall and F1 taken into account entities as defined in Section 2.5 (see also Section 4.2). For approximate match we use the metrics of precision and recall presented in MUC challenge, that are explained in detail by Chinchor et al. [47] and in the *Scoring Software User's Manual*[38], and that are presented below. F1 (defined in Section 2.5) is also used. Table 4.15 explains the scoring criteria.

POS: positives

$$POS = COR + PAR + INC + MIS$$

ACT: predicted as positives

$$ACT = COR + PAR + INC + SPU$$

$$precision = \frac{COR + 0.5 * PAR}{ACT}$$

[38] The Message Understanding Conference Scoring Software User's Manual.https://www-nlpir.nist.gov/related_projects/muc/muc_sw/muc_sw_manual.html, accessed June 2017.

abbrev.	**category**	**criteria**	**meaning in confusion matrix**
COR	correct	algorithm guess is positive and gold standard value is positive	TP
PAR	partial	algorithm and gold standard values have a partial coincidence (both with positive values)	accounted as a fraction of TP
INC	incorrect	corresponds to two incorrect results in the confusion matrix, one counted as FP for a spurious answer and one counted as FN for not getting the correct positive answer	
SPU	spurious	algorithm guess is positive and gold standard value is negative	FP
MIS	missing	algorithm guess is negative and gold standard value is positive	FN
NON	noncommittal	algorithm and GS values are negative	TN

Table 4.15: Scoring criteria for evaluating MUC-3 results.

$$recall = \frac{COR + 0.5 * PAR}{POS}$$

4.4.7 Technical Details

In this section we explain the technical details of our implementation.

SiMREDA implementation. At the very beginning *NLPTools-ES*, a Spanish plugin for GATE was used for doing tokenization, PoS tagging and lookup of RadLex terms in specially built Gazetteers. After comparing the results with other methods and taking some decisions on further steps, we decided to implement the different modules in Python, which was better suited for our problem, because it was easier to modify algorithms and to show the results in the expected format.

Given a set of reports we created a new file that includes the text of the original reports, where each meaningful entity referring to an anatomical entity or to a clinical finding is tagged with its type. We used HTML tags in order to have a visual tool easily checkable by the physicians and by us.

Dataset preparation. The development and testing datasets were partitioned randomly using scikit-learn toolkit.

NLP techniques. For SiMREDA we used a tokenizer previously implemented by us. PoS tagging was obtained with the use of Freeling [35].

Retrieval of RadLex terms. RadLex was downloaded in Protégé format. The selection of terms of *anatomical entity* and *clinical finding* RadLex categories has been done in Java with the help of the tutorial done by MantasCode.[39]

[39] JAVA: How to programmatically manipulate a Protégé-Frames lexicon/ontology/dictionary using Protege API and Java. http://mantascode.com/java-how-to-programmatically-

Retrieval of SNOMED CT terms. We used the package pymedTermino[40] for retrieving all the descendants of a SNOMED CT node and then an SQL query to retrieve the Spanish descriptions corresponding to the SNOMED CT ids (SCTID). More details can be seen in Section B.4 of Appendix B.

Implementation of exact match evaluation. As previously mentioned, to evaluate the performance of the algorithms with exact matching and measures of precision, recall and F1 we used the Perl script conlleval.[41] The output of the algorithm and the gold standard had to be transformed into IOB format and into the conlleval input format.

Therefore, a script was elaborated to convert the SiMREDA tags and their enclosed texts into a text in IOB format. Another script was created to transform the gold standard (in BRAT standoff format) into IOB format. Finally, a third script was implemented to create from two IOB format files one IOB format file in conlleval input format and finally conlleval.pl script was used to obtain the exact match results.

Partial match evaluation. Although there are MUC scorer implementations publicly available,[42] they are complex, because they take many other scores into account, so we finally developed our own MUC-based partial match implementation in Python. The detail of the implemented metrics can be seen in Section 4.4.6.

CRF implementation. There are many CRF implementations. We used CRF++[43], an open source implementation of linear chain CRF.

There is no implementation of cross validation in CRF++, so we partitioned randomly our development dataset and implemented it by ourselves in Python.

4.5 Results

In this section we present results of SiMREDA and CRF algorithms. For each of them we begin explaining experimental settings. Then, we show results and we carry out an analysis of the results. Finally, we compare SiMREDA and CRF results.

Precision (P), recall (R) and F1 measure (F1) were calculated against every entity type (AE and FI) and a final overall score, that considers both entity types is also given for all the measurements. Similarly, precision, recall and F1 for partial boundary matching is calculated for every entity type (AEPM and FIPM) and a final overall score (totalPM) is calculated.

In both cases we used the testing dataset composed by 20% of the annotated 513 reports (103 reports) (see Figure 4.1).

We are interested in a solution that retrieves a high rate of relevant entities and that the entities retrieved by the solution are actually positive (high recall and high precision). Hence, we will choose F1 metric, that balances precision and recall, as the metric in order to compare results.

manipulate-a-protege-frames-lexicon-ontology-dictionary-using-protege-api-and-eclipse/.

[40] pymedTermino https://pypi.python.org/pypi/PyMedTermino, http://pythonhosted.org/PyMedTermino/tuto_en.html (both accessed Nov. 2016)

[41] Perl script conlleval https://www.clips.uantwerpen.be/conll2000/chunking/output.html (accessed Jan 2018).

[42] MUC scorer implementations https://catalog.ldc.upenn.edu/docs/LDC2001T02/MUC_scorer3.3/ (accessed Nov. 2017)

[43] CRF++ https://taku910.github.io/crfpp/ (accessed Mar. 2018).

4.5.1 SiMREDA algorithm

Experimental Setting

First, for SiMREDA Module 1 we evaluated Variants 1 (RadLex improved translation) and 2 (use of SNOMED CT instead of RadLex). For Variant 2 we evaluated which collection of findings of SNOMED CT (CORE problem subset or *our findings*-)[44] had better results. Based on these results we decided which variant to use: Module 1 with RadLex translation obtained through Google Translate or Module 1 with Variant 1 or Variant 2. Then, we tested the results of the best Module 1 configuration with the incorporation of Module 3, and, finally with Module 2.

Results

Results of the original SiMREDA Module 1 (Google Translate -GT- RadLex translation without improvement) and SiMREDA Module 1 Variant 1 (GT RadLex translation with corrections in the translation) can be seen in Table 4.16. It can be noticed that SiMREDA Module 1 has better results with Variant 1.

	SiMREDA Module 1 and Module 1 Variant 1 (with RadLex)							
	Module 1				**Module 1 Variant 1**			
NE	**#**	**P (%)**	**R (%)**	**F1 (%)**	**#**	**P (%)**	**R (%)**	**F1 (%)**
AE	1,044	53.93	67.83	60.09	1,054	57.40	72.89	64.23
FI	456	34.21	31.64	32.88	422	37.20	31.85	34.32
total		47.93	54.35	50.94		51.63	57.60	**54.45**
AEPM	1,044	63.57	74.85	68.75	1,054	64.89	77.58	70.67
FIPM	456	50.69	44.93	47.63	422	50.11	45.12	47.49
totalPM		59.57	63.64	61.54		60.3	65.44	**62.76**

Table 4.16: **SiMREDA Module 1 and Module 1 Variant 1 (with RadLex)**. Evaluation results of SiMREDA algorithm using RadLex as information source. The translation to Spanish was carried out with Google Translate (GT) (Module 1) and was improved with the physicians' translation in Module 1 Variant 1. Exact and approximate matches are shown for anatomical entities (AE, AEPM) and findings (FI, FIPM). Columns # show the number of entities of each type detected by the algorithm. For example, the algorithm detected 1054 anatomical entities in Module 1 Variant 1. For both types of match, metrics are calculated by entity and also the overall value (total) is calculated.

Table 4.17 shows results of SiMREDA Module 1 Variant 2 (use of SNOMED CT instead of RadLex as information source). As explained in Section 4.4.2 two tests were carried out. The first uses as findings, those retrieved from the CORE problem subset. The second uses *our findings*. The anatomical entities are the same for both tests. It can be noticed, that SiMREDA Module 1 Variant 2 has better results with the CORE problem subset than with *our findings*.

From Tables 4.16 and 4.17 it can be seen that SiMREDA Module 1 has better results with Variant 1 than with Variant 2 (higher F1 for exact and for partial match). Therefore, we decide that SiMREDA will be configured with Variant 1 of Module 1. Next, we test results with the addition of Module 3 to Variant 1 of Module 1, and then with the addition of Module 2 to Modules 1 and 3. Results are shown in Table 4.18.

[44] *Our findings* were selected from *substance*, *body structure* and *finding* hierarchies.

	SiMREDA Module 1 Variant 2 (with SNOMED CT)							
	CORE problems subset				**Our findings**			
NE	**#**	**P (%)**	**R (%)**	**F1 (%)**	**#**	**P (%)**	**R (%)**	**F1 (%)**
AE	1,696	35.50	72.53	47.66	310	30.00	11.20	16.32
FI	540	29.81	32.66	31.17	2,154	9.01	39.35	14.66
total		34.12	57.67	42.88		11.65	21.69	15.16
AEPM	1,696	45.04	82.75	58.33	310	12.41	14.53	13.38
FIPM	540	36.93	46.15	41.03	2,154	21.9	54.39	31.23
totalPM		42.69	69.03	52.76		18.46	32.57	23.56

Table 4.17: **SiMREDA Module 1 Variant 2 (with SNOMED CT)** Two tests are carried out, considering two different sets of findings. 1) the CORE problem subset, 2) *our findings*. Anatomical entities are the same in both tests. Exact and approximate matches are shown for anatomical entities (AE, AEPM) and for findings (FI, FIPM). Columns # show the number of entities of each type detected by the algorithm. For both types of match, metrics are calculated by entity and also the overall value (total) is calculated.

	SiMREDA Modules 1 (Variant 1), 2 and 3							
	SiMREDA Modules 1 and 3				**SiMREDA Modules 1, 2 and 3**			
NE	**#**	**P (%)**	**R (%)**	**F1 (%)**	**#**	**P (%)**	**R (%)**	**F1 (%)**
AE	1,036	58.49	73.01	64.95	1,035	58.74	73.25	65.20
FI	425	55.29	47.67	51.20	427	56.21	48.68	52.17
total		57.56	63.57	60.42		58.00	64.10	**60.90**
AEPM	1,036	65.35	77.11	70.75	1,035	65.46	77.23	70.86
FIPM	425	58.97	53.55	56.13	427	59.86	54.49	57.05
totalPM		63.39	68.48	65.83		63.73	68.9	**66.21**

Table 4.18: **SiMREDA Modules 1, 2 and 3.** In both cases Variant 1 of Module 1 is used. The first columns show results of Modules 1 and 3. The last columns show results of Modules 1, 2 and 3. Exact and approximate matches are shown for anatomical entities (AE, AEPM) and for findings (FI, FIPM). Columns # show the number of entities of each type detected by the algorithm. For both types of match, metrics are calculated by entity and also the overall value (total) is calculated.

Table 4.19 shows the terms that match with suffix *megalia* and that are related to *hepatomegalia -enlargement of the liver-*. Table 4.20 presents those Graeco-Latin morphemes listed in Table 4.11, that were discovered in the testing dataset. In our algorithm we only look for the *finding* category. The morphemes found were:[45]

- **megalia:** *esplenimegalia*[46] (1), *esplenomegalia -splenomegaly-* (14), *hepatomegalia -hepatomegaly-* (3),
- **itis:** *apendicitis -appendicitis-* (1), *ascitis -ascites-* (4), and
- **osis:** *esteatosis -steatosis-* (3), *estenosis -stenosis-* (3), *etenosis*[47] (1), *poliquistosis -polycystosis-*(1).

[45] The number between brackets indicates the number of appearances of the term in the testing dataset.

[46] It refers to *splenomegaly* but it is incorrectly written.

[47] It refers to *stenosis* but it is incorrectly written.

number of occurrences	**term**
1	hapatomegalia
2	hapetomegalia
1	heaptomegalia
1	heatomegalia
1	hepatmegalia
1	hepatoamegalia
1,350	hepatomegalia
1	hepatoomegalia
1	hepattomegalia
1	heptaomegalia
1	heptoesplenomegalia
6	heptomegalia
1	hetomegalia

Table 4.19: Appearances of terms referring to *hepatomegalia* (enlargement of the liver) in the 79,125 anonymized reports.

morpheme	**category**	**number of appearances (distinct)**
-itis	finding	5 (2)
-megalia	finding	18 (3)
-osis	finding	8 (4)
-grafía	procedure	4 (1)
peri-	location	58 (7)
retro-	location	3 (2)
supra-	location	10 (3)
sub-	location	4 (2)

Table 4.20: Morphemes related with medical terms appearing in the test set. Only those referring to the finding category were taken into account.

Analysis of results

The best SiMREDA results for named entity recognition of anatomical entities and clinical findings is obtained using modules 1, 2 and 3. For this configuration F1s are 65.20 (AE), 52.17 (FI) and 60.90 (overall). Results of adding Module 2 (morphology) to Modules 1 and 3, reported in Table 4.18, are not so noticeable (they increase overall F1 in less than 1%). We believe, that this fact is related with the reduced number of reports in the testing dataset, where a small number of words appear. Table 4.20 shows that only 31 findings with Graeco-Latin suffixes appear in our testing dataset. However, the detection of morphemes related to the medical domain helped us to detect terms that are misspelled. For example, *etenosis* for *estenosis* -*stenosis*- and *hesplenomegalia* for *esplenomegalia* -*splenomegaly*- were found in reports and detected as findings by module 2. Notice also results shown in Table 4.19, that presents the terms that match with suffix *megalia* and that are related to *hepatomegalia* -enlargement of the liver, *hepatomegaly*-. 18 of 1,368 references to *hepatomegalia* were misspelled in the set of 79,125 anonymized reports and could be found by Module 2.

It is interesting to notice, that the improvement of only 10% of RadLex translations derived in a relative F1 increase of anatomical entities and findings of ∼ 7% and ∼ 4% respectively. Therefore, we conclude that it makes sense to invest effort in improving the translations.

At a first glance, it is surprising what happens with both experiments carried out with SNOMED CT. As previously mentioned, they use the same anatomical entities but different clinical findings. Table 4.9 shows the quantity of terms retrieved for each entity type. From the analysis of Table 4.17 results, it can be seen that when working with the CORE problem subset much more anatomical entities and much less clinical findings are detected, than when working with *our findings* (1,696 vs. 310 AEs and 540 vs 2,154 FIs). F1 measure is also three times larger in AE and two times larger in findings with the CORE problem subset than with our findings dataset. In order to understand a possible reason for this issue, we analyzed the terms *hígado -liver-* and *vejiga -bladder-*, that should be ideally considered as AEs. As expected, they appear the same amount of times to terms categorized as anatomical entities, but they appear much more times embedded in terms categorized as findings in our finding subset than in the CORE problem subset (see Table 4.21). To further analyze the situation, we observed that some of the findings of our dataset that include the world *vejiga* are: ***lesion traumatica*** *de la vejiga -traumatic lesion of the bladder-*, ***diverticulo congenito*** *de vejiga -congenital diverticulum of bladder-*, ***aborto espontaneo*** *con perforacion de vejiga -spontaneous abortion with bladder perforation-* and *vejiga* ***hipotónica*** *-hypotonic bladder-*.[48] As can be seen, *vejiga*, an actual anatomical entity, appears many times embedded in terms corresponding to clinical findings.

set	**# appearances of *hígado* as FI**	**# appearances of *vejiga* as FI**
core subset	16	29
our subset	150	244

Table 4.21: Number of appearances of terms *hígado* (liver) and *vejiga* (bladder) as part of terms tagged as findings in the CORE problem subset and in *our finding* dataset.

Finally, as expected, partial match results are always higher than exact match results. Table 4.22 shows the percentage in which F1 values of partial match increase with respect to exact match with the different SiMREDA configurations. For example, as reported in Table 4.18, the overall F1 for SiMREDAs final configuration (Modules 1, 2 and 3) is 60.90 with exact match. Partial match achieves a relative increase in F1 of 8.72%.

In all cases findings show a greater increase in partial match F1s than anatomical entities. We believe this is motivated, because it is much more complex to determine the boundaries of a clinical finding than those of an anatomical entity. This issue was noted during annotation. Furthermore, in our dataset findings have longer terms (in amount of words composing them) than anatomical entities, which makes its boundary detection a harder problem.

It can be also noticed, that the increase of F1s partial match is much higher in Module 1 than it is when Module 3 is incorporated. We expected this result. We understand that this is due to the fact that when we extend the detected findings and anatomical entities in Module 3 based on the analysis of patterns of a subset

[48] Terms in bold are those that we understand that refer to findings.

	Mod 1 (%)	**Mod 1 Var 1** (%)	**Mod 1 Var 2 CORE subset** (%)	**Mod 1,3** (%)	**Mod 1,2,3** (%)
AE	14.41	10.03	22.39	8.93	8.68
FI	44.86	38.37	31.63	9.63	9.35
total	20.81	15.26	23.04	8.95	8.72

Table 4.22: Relative improvements with partial match with regards to exact match in F1 results for different SiMREDA configurations.

of our development dataset we increase the number of exact matches and we reduce the number of partial matches.

Analyzing results shown in Tables 4.16 and 4.18 it can be seen that the incorporation of Module 3 achieves a relative increase of the performance of the exact match recognition of clinical findings in almost 50% (from 34.32% to 51.20%), while the performance of anatomical entities recognition remains almost the same. We can conclude that the incorporation of rules to the dictionary-based algorithm improves its performance.

Error analysis

Some errors are due to following causes:

- tokenization problems: the text *(...)ascitis-* appeared in one of the reports. The tokenizer did not separate the word *ascitis* from the symbol -, so ascitis was not recognized by our algorithm as a finding
- annotation criteria
  - a decision was taken to annotate implants, such as *kidney implant* as an AE. The algorithm does not annotate implant as part of an anatomical entity.
  - for example, *ovarian cyst* should be annotated as [ovarian cyst](FI), while the algorithm detects [ovarian](AE) [cyst](FI).
- annotation inconsistencies: As mentioned in Section 3.6, there is a number of errors and inconsistencies in the annotations. Some of them, like the omission of annotation of entities (such as *bile duct* and *dilated*), the incorrect classification of entities (such as *gallbladder* as FI) erroneously worsens the results. The annotation of entities with wrong boundaries explains, in part, the difference of performance among the exact match and the partial match.

4.5.2 CRF algorithm

Experimental Setting

As previously mentioned, we will use 5-fold cross-validation in order to select the best set of features. Therefore, reports belonging to the development dataset were randomly ordered. Then, 5 disjoint folds of the same size were selected. Next, for each feature set, five training instances were performed always leaving one of the folds out and using it for validation purposes. See Figure 4.1 to review the dataset construction method and the dataset selection for cross-validation. For each feature set, the average of precision, recall and F1 measures and their standard deviations over all the training instances were calculated.

Based on F1 and on its standard deviation, we decided which feature set to use and with this feature set we retrained the CRF with the whole development set and tested it with the testing dataset shown in Figure 4.1.

Results

Table 4.23 shows exact and partial match results of CRF intermediate results tested with 5-fold cross validation carried out on the development dataset for different features.

Baseline feature set (CRF++)			
	P-mean±std (%)	**R-mean±std (%)**	**F1-mean±std(%)**
AE	93.34 ± 0.95	86.91 ± 1.59	90.0 ± 1.11
FI	84.36 ± 1.85	72.88 ± 3.34	78.18 ± 2.58
total	89.87 ± 1.03	81.23 ± 1.36	85.33 ± 1.18
AEPM	93.69 ± 0.53	88.08 ± 1.19	90.79 ± 0.78
FIPM	88.72 ± 1.56	77.15 ± 2.6	82.52 ± 2.04
totalPM	91.77 ± 0.77	83.64 ± 1.03	87.51 ± 0.84

Wi - ENRE feature set			
	P-mean±std (%)	**R-mean±std (%)**	**F1-mean±std(%)**
AE	93.96 ± 0.73	90.46 ± 0.96	92.17 ± 0.72
FI	85.06 ± 1.91	76.01 ± 2.56	80.26 ± 1.93
total	90.52 ± 1.02	84.6 ± 1.27	87.45 ± 1.05
AEPM	94.96 ± 0.59	91.72 ± 0.86	93.31 ± 0.51
FIPM	88.82 ± 1.7	80.17 ± 1.92	84.26 ± 1.56
totalPM	92.56 ± 0.91	87.0 ± 0.97	89.69 ± 0.78

Our feature set			
	P-mean±std (%)	**R-mean±std (%)**	**F1-mean±std(%)**
AE	94.29 ± 0.91	89.57 ± 1.65	91.86 ± 0.91
FI	83.88 ± 2.6	75.6 ± 3.02	79.52 ± 2.75
total	90.22 ± 1.26	83.9 ± 1.32	86.95 ± 1.25
AEPM	94.83 ± 0.48	90.67 ± 1.57	92.7 ± 0.9
FIPM	88.57 ± 2.14	80.55 ± 2.29	84.37 ± 2.15
totalPM	92.38 ± 0.92	86.56 ± 0.85	89.37 ± 0.8

Table 4.23: CRF intermediate results with different feature sets for exact match and partial match. Columns are precision, recall and F1 and include standard deviations. AE, FI and total refer to exact match of anatomical entities, findings and overall. The suffix PM corresponds to partial match results.

Figure 4.4 shows F1 values of anatomical entities, clinical findings and the overall F1 (total) for each feature set tested and their standard deviations for exact match. As can be seen, the feature set that has the highest average overall F1 (Wi-ENRE) has also the highest average F1 for findings and for anatomical entities. Furthermore, there is no much variation in standard deviation among the different feature sets. That is why we chose as the best results, those obtained by feature set WI-ENRE. Nevertheless, it can be noticed that the relative difference among the F1s is very low (0.57 % in regard to our feature set and 2.42% with respect to the baseline).

Table 4.24 shows the results of testing (with the testing dataset) the CRF trained with the whole development dataset and with the features that had the best results on the development dataset (Wi-ENRE features).

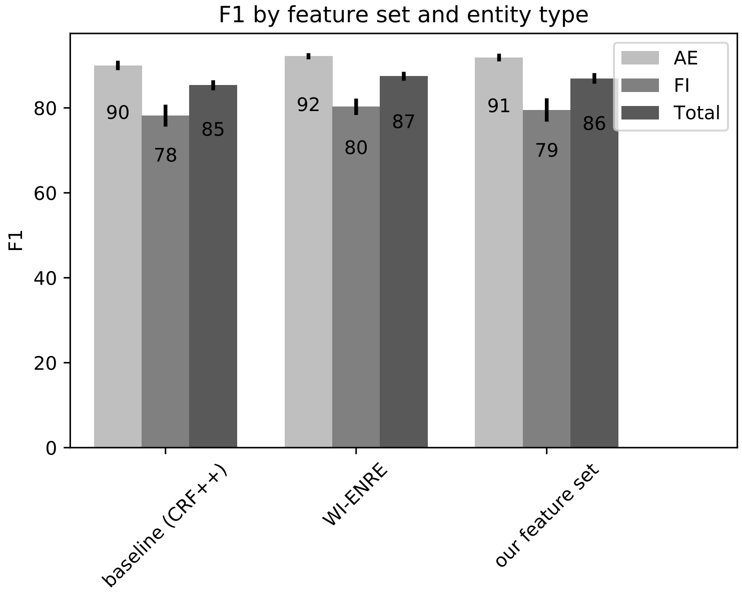


Figure 4.4: Average F1 values for anatomical entities, clinical findings and overall F1 of different feature sets for exact match (baseline, Wi-ENRE, our feature set). Decimals are not shown in the bars.

	CRF		
NE	**P (%)**	**R (%)**	**F1 (%)**
AE	92.09	91.56	91.82
FI	85.78	74.95	80.00
total	89.68	84.70	87.12
AEPM	95.00	92.61	93.79
FIPM	88.46	80.06	84.05
totalPM	92.42	87.45	89.87

Table 4.24: Results of CRF run with development dataset, Wi-ENRE features [130] and tested with the testing dataset. Exact and partial match results are shown.

Analysis of results

Table 4.25 shows the result of CRF implementation with the same feature set for French [130], German [214] and Spanish (our results) datasets. The features used are those proposed by Jiang et al. [130] and referred as Wi-ENRE. For building the table, what we consider AE and FIs in the French dataset are *anatomy* and *disorders* hierarchies of UMLS. In the case of for German, what we consider AEs for building the table corresponds to *organs* and what we consider FI corresponds to *symptoms*, *diagnoses* and *observations*. Since all results are tested with different genre of data and in different languages it is not easy to draw a conclusion about the differences in the results. In Spanish and in French anatomical entities have a higher F1 than findings. That is what usually happens. It can be also noticed that results with our Spanish dataset are better in both entity types than in the original French implementation. This might have to do with the fact that our corpus is of a restricted domain -only radiology reports, while the French implementation has EMEA and MEDLINE articles-, that in our case we had two entity types, while the other case had to select among 10 entity types, and that we trained with 410 reports and tested with 103, while in the French case, 836 MEDLINE titles and 4 EMEA documents were used for training and 832 MEDLINE titles and 12 EMEA documents were used for testing. Besides, the definition of AE and FI among both systems does not necessarily coincide.

implementation	**entity type**	**P (%)**	**R (%)**	**F1 (%)**
Spanish	AE	92.09	91.56	91.82
	FI	85.78	74.95	80.00
German	AE	96.96	65.23	77.9
	FI	95.17	75.16	83.98
French	AE	72.8	79.4	75.9
	FI	61.5	49.1	54.6

Table 4.25: Results of different CRF implementations with Wi-ENRE features for Spanish, German and French.

4.5.3 General analysis

Table 4.26 shows results of SiMREDA and CRF algorithms that were already shown in Tables 4.18 and 4.24. For all metrics, CRF outperforms SiMREDA results.

As can be seen in Table 4.27, where the relative increase in F1 values of CRF with respect to SiMREDA is shown, the improvement occurs for exact as well as for partial match.

	SiMREDA compared to CRF					
	SiMREDA			**CRF**		
NE	**P (%)**	**R (%)**	**F1 (%)**	**P (%)**	**R (%)**	**F1 (%)**
AE	58.74	73.25	65.20	92.09	91.56	91.82
FI	56.21	48.68	52.17	85.78	74.95	80.00
total	58.00	64.10	60.90	89.68	84.70	87.12
AEPM	65.46	77.23	70.86	95.00	92.61	93.79
FIPM	59.86	54.49	57.05	88.46	80.06	84.05
totalPM	63.73	68.9	66.21	92.42	87.45	89.87

Table 4.26: **SiMREDA implementation including Modules 1, 2 and 3 compared to CRF implementation with Wi-ENRE features.** Exact and partial match results are shown for each entity type and the overall measure (total) is also shown.

Considering Tables 4.4 and 4.26, we can see that Spanish NER in the general domain has better results that the Spanish NER we developed in the medical domain. This is usually the case. Though, our CRF solution has better results than CONLL 2002 NER challenge for Spanish. Table 4.4 also shows that in the general domain results are better in English than in other languages (compare MUC-6 and MET 1, MUC-7 and MET-2 and CONLL 2002 in different languages).

Based on Tables 4.5 and 4.26, it can be seen that SiMREDAs results (considering exact match F1) for anatomical entities are similar to best CLEF 2015 results and CLEF 2016 results for MEDLINE articles (both in French). They are worse than

entity	exact match F1 improvement	partial match F1 improvement
	(%)	(%)
AE	40.83	32.36
FI	53.35	47.33
total	43.05	35.73

Table 4.27: **Improvement of CRF over SiMREDA.** Results for exact and partial match F1 are shown for each entity type and the overall measure (total) is also shown.

CLEF 2016 results for EMEA articles. Finding detection results and overall results for Spanish are worst, though. As expected, results with English datasets outperform those obtained with Spanish and French datasets.

Concluding, the development of a dictionary-based algorithm enhanced with rules is more laborious than a machine learning approach such as CRF. In cases as ours, where we did not have specific resources for the radiology domain in Spanish it is even more difficult. Nevertheless, this method has the advantage of needing few annotated data. At the beginning we did not have the perspective of having more than 200 annotated reports, so we put our efforts in developing the SiMREDA algorithm. We then could work on the annotation project described in Chapter 3, where we achieved to get 513 annotated reports and discovered that with a few hundred reports we could train a CRF with competitive results. Based on the good perspective of the results, feature engineering can be carried out in order to improve results.

4.6 Conclusion

In this chapter we presented SiMREDA, a dictionary-based entity recognition algorithm, enhanced with morphology analysis and with a post-processing based on the analysis of PoS patterns of the entities of interest, and an algorithm based on CRF. We also presented a method for classifying reports.

SiMREDA approach can be used when there are no datasets annotated for implementing machine learning techniques and when there are no dictionaries in the original language.

From the results obtained and the analysis carried out we can draw following conclusions.

Despite the conclusion about the coverage of SNOMED CT terms in the radiology domain obtained in [5],[49] we obtained better results with SiMREDA using a translated version of RadLex -although it is not a high-quality translation- than with SNOMED CT terms that are already in Spanish.

Looking at Table 4.16 we can conclude that our algorithm is sensitive to a poor translation. We could experience that the improvement of only 10% of RadLex translations improves our results. Therefore, we conclude that it makes sense to invest effort in improving the translation.

From the results depicted in Table 4.18 we can see that adding rules to our algorithm based on the analysis of its PoS tagging patterns improved the results.

[49] The paper is not available online. Results were discussed in a personal communication.

In the same table, we can see that the morphological processing (Module 2) improvement is almost imperceptible, but we can appreciate that it recognizes more anatomical entities and clinical findings and that the limited increase in performance is probably due to the reduced size of the test set. We could also see that the morphological module helped in recognizing misspelled entities.

Lenient match draws better results than the exact matching for every entity type across all settings of both algorithms tested. Besides in this use case it is more important to determine if an entity is present than to correctly determine its boundaries. Therefore, we conclude that it is important to report a precisely defined partial metric accompanying the exact match results.

Despite having a small annotated dataset (513 reports), we could successfully apply a machine learning technique.

Based on the previous analysis we can say that we answered our research questions.

There are many studies than can be carried out as future work. Some of them are currently being performed by us.

There are some phrases, we call *prefix terms*,[50] such as "could suggest", "is visualized" that usually determine that the following noun phrase corresponds to a clinical finding. Detecting those phrases and the noun phrases that come after them, could help improve the recall of retrieved findings.

The abbreviations databases created in previous works [286, 172] and referred in Chapter 2 are not useful in our cases, given that they are for English abbreviations. The construction of similar databases for Spanish radiology reports, would probably be less useful, since, as we previously mentioned, many of the abbreviations used in these kinds of reports do not follow naming conventions and would, therefore, be difficult to generalize to other texts. However, efforts could be done to study the subject and to create an abbreviation database. Therefore, previous efforts could be studied [147].[51]

It would be interesting to detect of all the morphemes composing a word, as [271] carried out. This can help to a better understanding of the words. For instance, words that have more than one morpheme related with the medical domain (e.g. **peri**ton**itis** -peritonitis-) can be found, and their semantics can be better comprehended. Consider also *cardiopatía* -cardiopathy- and *linfoadenopatía* -lymphadenopathy-, whose decomposition into morphemes (cardio-patía and linf-o-adeno-patía) explains in which anatomical entity the findings have occurred.

There are some patterns that would also probably help to improve finding retrievals. Consider:

- AE FI, as in [ovarian](AE) [cyst](FI),
- FI AE,
- and FI (en (el |la(s?) |los|λ) |de (la(s?) |los |λ) |del) AE,[52,53] as in "[luxación](FI) de la [cadera](AE)" (hip dislocation).

With the current version of SiMREDA, these patterns are not considered as findings, but they were annotated as findings. An additional SiMREDA module that detects those patterns as entities could be constructed. It is also important to notice that detecting [ovarian](AE) [cyst](FI) as a first step, has as advantage, that it can be

[50] In Spanish they usually occur before the terms of interest.

[51] Acronyms and abbreviations provided by the National Academy of Medicine of Colombia http://dic.idiomamedico.net/Siglas_y_abreviaturas and by the Spanish Ministry of Health http://www.redsamid.net/archivos/201612/diccionario-de-siglas-medicas.pdf?0 (both accessed Mar. 2018).

[52] in |in the |from

[53] Written as a regular expression.

determined where the finding is located. If [ovarian cyst](FI) would have been detected, then this understanding would be lost.

There are ways of improving SiMREDAs Module 1. As commented in Section 4.5.1 and particularly in Table 4.21, there are some terms as *liver* and *bladder* that are recognized as FI instead as AEs. We are working in some heuristics in order to avoid these errors. Finally, we are considering an improvement of RadLex translation through the use of SNOMED CT and DBPedia. In this variant we would take the translation of RadLex obtained by Google Translate and partially improved by the physician. For those terms that were not checked by the specialist, the translation of SNOMED CT will be considered. For those cases that do not have a SNOMED CT translation, the translations of UMLS, would be taken into account. And finally, for the rest of the terms Google Translate, DBpedia and Wikipedia would be taken into account.

4.7 Resumen

El reconocimiento de entidades nombradas (NER) es una tarea de extracción de la información, cuyo objetivo es localizar en un texto instancias de determinado tipo de unidad de información y asignarles una categoría predefinida. Esta tarea se ha aplicado a distintos géneros textuales, dominios y tipos de entidades. Entre las técnicas para detectar entidades se encuentran los métodos basados en búsqueda en diccionarios u otro tipo de terminologías, los basados en la elaboración de reglas y los basados en métodos estadísticos. También existen soluciones híbridas, que combinan algunos de los métodos precedentes. Las características de los informes médicos mencionadas anteriormente, la ausencia de terminologías completas, acentuado por la falta de las mismas en determinados idiomas, la ambigüedad y la polisemia, y el hecho de que las entidades biomédicas pueden contener nombres compuestos por muchas palabras (con el consecuente problema de definición de límites de las entidades y de superposición de las mismas) hacen que el problema de reconocimiento de entidades sea más difícil en el dominio biomédico que en el dominio general [50, 150].

Gran parte del trabajo en el NER biomédicas ha estado enfocado en la detección de genes y de nombres de proteínas en artículos científicos escritos en inglés. Mucho menos se ha realizado para el dominio médico y para idiomas distintos al inglés. El procesamiento de textos informales y en idiomas distintos al inglés, agrega dificultades adicionales, debido a que hay menos recursos disponibles.

En este capítulo describimos distintos métodos implementados para la detección de entidades anatómicas (AE) y hallazgos clínicos (FI) en un conjunto de RR escritos en español. Para esto proponemos, implementamos y evaluamos dos métodos distintos para la detección de entidades. La elección de los métodos está relacionada con los recursos disponibles. Inicialmente sólo contábamos con un conjunto de aproximadamente 200 informes anotados. Su tamaño no hacía factible la aplicación de técnicas de aprendizaje automático supervisado. Posteriormente al anotar un conjunto de 513 informes (ver capítulo 3), pudimos desarrollar el segundo método.

El primer método, SiMREDA, está basado en la búsqueda de términos pertenecientes a RadLex.[54] Dado que no hay una versión para español de RadLex, se la tradujo y se utilizó una técnica de índice invertido para encontrar sus términos en los informes. Se evaluaron distintas formas de traducción y los efectos de estas sobre los resultados. La traducción introduce, entre otros, los siguientes problemas: determinados términos son frecuentemente utilizados en español con sinónimos, que

[54] Ontología del dominio radioógico escrita en inglés.

son menos frecuentemente utilizados en inglés,[55] a veces los términos en español son preferidos de una manera adjetival en lugar de como sustantivo,[56] y las entidades pueden estar compuestas por más de una palabra, lo cual en muchos casos trae problemas en el orden de las palabras traducidas. Dado que una gran cantidad de términos médicos está compuesta por morfemas grecolatinos (por ej. los sufijos *itis* y *osis* indican patología) SiMREDA tiene en cuenta la aparición de los mismos para la detección de hallazgos clínicos. Finalmente, se analizaron las categorías gramaticales de un subconjunto de AEs y FIs anotados. A partir de este análisis se elaboraron reglas, que se utilizan para mejorar la NER.

El segundo método utiliza campos aleatorios condicionales (CRF),[57] una técnica muy utilizada para NER. No hicimos foco en realizar una optimización de las características,[58] pero de todas formas obtuvimos muy buenos resultados testeando diversas características existentes y propuestas por nosotros.

Para testear nuestros algoritmos utilizamos un mismo conjunto de datos. Para esto particionamos el corpus anotado en dos conjuntos: de desarrollo y de test. El conjunto de test lo dejamos apartado y sólo lo utilizamos para testear ambos algoritmos. Con el conjunto de desarrollo elegimos el mejor conjunto de características. Para esto hicimos validación cruzada de 5 iteraciones y luego utilizamos todo el conjunto de desarrollo como conjunto de entrenamiento para testear el CRF con el conjunto de test separado previamente. Parte del conjunto de desarrollo fue utilizado también para el estudio de los patrones de las categorías gramaticales, utilizado en SiMREDA.

A lo largo del capítulo explicamos en detalle la problemática del criterio de coincidencia exacta para la evaluación de resultados. Para evaluar los resultados usamos dicho criterio y uno de coincidencia parcial. También presentamos un método sencillo para la clasificación de informes en función de la existencia de hallazgos clínicos afirmados o de la inexistencia de estos.

Los resultados obtenidos por SiMREDA son prometedores dadas las limitaciones de recursos con las que contábamos. La implementación de CRF obtiene mejores resultados, con lo cual se propone trabajar a futuro en una mejor elección de características. De todas formas, consideramos que SiMREDA es una alternativa recomendable para idiomas que carecen de altos recursos lingüísticos, terminológicos y de un volumen alto de corpus anotados.

[55] Por ej., “arteria mamaria interna.^{es} utilizada normalmente en español en vez de “arteria torácica interna”, la traducción de *internal mammary artery*, que es la forma en la que normalmente se la nombra en inglés.

[56] Por ejemplo, se utiliza “folículo ovárico”mientras que en inglés se utiliza más frecuentemente *follicle of ovary*, que sería traducido como “folículo de ovario”

[57] *Conditional random fields* en inglés.

[58] *Features* en inglés.

CHAPTER 5

Negation Detection

In this chapter we introduce the negation and speculation detection problem, its importance in the biomedical domain and especially in clinical reports, and the particularities of the problem in Spanish and German clinical reports. We then present a summary of related work in the negation and speculation detection area. Next, we present the algorithms developed and techniques adapted for the detection of negation in Spanish radiology reports and the detection of negation and speculation in German discharge summaries and clinical notes. Finally, we close the chapter presenting results, discussions and concluding remarks.

5.1 Introduction

A clinical condition mentioned in a medical report does not necessarily mean that a factual condition is reported, since the term referring to the condition could be under the scope of negation or epistemic modality markers (hedges).

Consider, for example, following report with tagged findings. Negation terms are bold. *"Pancreas: tamano y ecoestructura normal. Retroperitoneo vascular: **sin** <FI> alteraciones</FI>. **No se detectaron** <FI>adenomegalias </FI>. (...)" "Pancreas: normal size and echotexture. Vascular retroperitoneum: **without** <FI> changes </ FI>. **No** <FI> lymphadenopathies </FI> were detected.(...) "*. The tagged findings *changes* and *lymphadenopathies* are negated.

Negation and speculation are linguistic phenomena that modify the meaning of terms under their scope. The aim of detecting them is to distinguish facts from impressions and hypothesis and to determine which conditions are present and which are absent. In order to deal with this problem, not only the presence of negation or speculation terms has to be detected, but also its scope has to be determined. Besides terms, relations can also be negated and speculated.

We refer to language constructions that denote negations and hedges as *negations* and *speculations* respectively.

There are different ways to express negations. Syntactic negations are expressed using negation particles (such as *no* and *has not been detected*). Semantic negations are denoted by the use of expressions that mean negation (such as *disappeared*) and morphological negation is expressed by the use of words with a negation affix (such as ***a**morphous*).[1]

There are many ways to detect negations and speculations in biomedical texts. The two more relevant techniques are based on rules (some include syntactic methods) and on machine learning techniques.

According to Chapman et al. [42], many of the medical conditions described in medical reports are negated. In our annotated corpus (see Chapter 3), 56% of the findings are negated, 27.89% of the sentences contain negations and 2.04% contain hedges. 22.80% sentences in Spanish radiology reports are reported to have negations in [61]. In the BioScope corpus[2] 13.55% of the sentences belonging to the radiology reports subset of the corpus contain negations and 13.30% contain hedges [267].

This suggests that the detection of negations in texts of the biomedical domain is an important task in the field of BioNLP and in a clinical IE pipeline. The detection of negation and speculation has also received attention in general domains [207] and in other tasks, such as sentiment analysis [280] and automatic translation.

In the last years the interest in the subject has increased and several workshops and challenges, -many of them specific of the biomedical domain-, have been organized. Some of them are BioNLP'09 Shared Task 3 [141], *Workshop on Negation and Speculation in Natural Language Processing in 2010* [179],[3] *CoNLL 2010 Shared Task* [89], the *2010 i2b2 NLP challenge* [201], *SEM 2012 Shared Task* [174], *Negation and Speculation detection in Biomedical Texts* tutorial in RANLP 2017[4] and *Workshop of Negation in Spanish, SEPLN 2017*[5]. Also books and reviews about negation and speculation detection in biomedical texts have been published [75, 169].

Most of the research in negation and speculation detection has been performed for English written texts. Its application to other languages, such as Spanish and German is more difficult due to the lack of corpora, the need to do some translations and some characteristics of the languages, such as the existence of circumfixes in German. The absence of publicly available corpora is more remarkable in the clinical reports genre, because of the data privacy issues discussed in previous chapters.

Our goal is to detect negation and speculation in clinical reports in order to be able to discover which tagged findings are factual and distinguish them from those that are not. For reaching this goal, we implement different techniques.

In particular, we were interested in applying negation detection methods to our Spanish radiology reports. Within the context of a collaboration with German research labs, we were able to obtain other kind of medical reports (discharge summaries and clinical notes of the nephrology domain) and decided to test the best of our implemented methods in German clinical reports. Thus, our contribution is the elaboration of negation and speculation detection techniques for Spanish and German medical reports. Working with medical reports of different characteristics (different length and level of formality) and adapting the same technique to both

[1] Affixes can also convert words not implying medical conditions into clinical findings (e.g. ***un**comfortable*).

[2] BioScope is an English corpus of biomedical texts annotated for uncertainty, negation and their scopes. It consists of medical reports (radiology reports), full papers and abstracts, both of the biological domain. More details about BioScope are described in Section 5.2.

[3] `http://www.clips.ua.ac.be/NeSpNLP2010/program.html` (accessed Dec. 2017).

[4] Negation And Speculation detection In Biomedical Texts Tutorial, RANLP (Recent Advances in Natural Language Processing) 2017 `http://lml.bas.bg/ranlp2017/tutorials.php#cruz` (accessed Mar. 2018).

[5] SEPLN (Spanish society of natural language processing), Taller de NEGación en Español `http://sepln2017.um.es/neges.html` (accessed Mar. 2018).

languages, allowed as to extract conclusions about the advantages, disadvantages and possible optimizations of the technique used. These conclusions could be employed for further implementations of negation and speculation detection in other Indo-European languages and other domains.

In this chapter we introduce different algorithms developed to determine if a clinical finding is under the scope of negation in radiology reports written in Spanish and under the scope of negation or speculation in discharge summaries and clinical notes written in German. In both cases we focus on negation determined by the appearance of syntactic negation terms and terms that semantically determine negation.

The methods implemented for Spanish include a syntactic technique based on rules derived from the detection of negation patterns inferred from the analysis of paths in dependency parse trees and an adaptation to Spanish of NegEx, a well known rule-based negation detection algorithm [43]. The input to NegEx are i) sentences with tagged clinical findings and ii) a list of negation and speculation terms called *triggers*. Using this information, NegEx determines if the finding is within the scope of negation or speculation. We also adapt NegEx for German.

In both cases, we compare our implementations with a simple dictionary lookup algorithm that we developed as baseline for each language.

NegEx was chosen not only because it has good results for other languages, but also because it is straightforward to implement, and the results can be easily understood. Furthermore, it has been successfully implemented for languages other than English.

As far as we know, of our methods only NegEx has been implemented for Spanish. As the adaptation was for other type of texts (EHRs extracted from SciELO, some of them more formally written than usual EHRs) and it was not available for public use we decided to develop our own adaptation. Working with Spanish presents some challenges: we had to build a corpus and annotate it, since at the time we developed our algorithm there was -to the best of our knowledge- no publicly available annotated corpora for negation detection in Spanish medical reports.[6] Furthermore, syntactic parsing tools are less developed for languages other than English, and translations needed for the development of the work incorporates errors.

The detection of negation and speculation in German text is not developed either (see descriptions of previous work in Section 5.2). There were also no publicly available annotated corpora for negation detection in German medical reports, so we had to build our own corpus.

In comparison to English, Spanish and German clinical data differ in various characteristics which have to be taken into account for the successful application of an algorithm detecting non-factuality. First of all, Spanish and German are richly inflected languages (e.g. *no* can be translated as *ningún*, *ninguna*, etc. in Spanish and *kein*, *keiner*, *keine* etc. in German). Furthermore, German includes *discontinuous triggers*, such as *kann … ausgeschlossen werden …*[7] (*can be ruled out*). Triggers may precede, but may also follow the negated expression, as presented in Table 5.1. Regarding this situation, Wiegand et al. [280] state, that the detection of negation scope in German language is more difficult than in other languages, such as English.

[6] At that time, the annotation performed by us and presented in Chapter 3 had not been carried out either.

[7] Dots indicate potential positions of the finding: (kann ... *finding*... ausgeschlossen werden, ... *finding*... kann ausgeschlossen werden).

language	precede	follow
Spanish	***no se detectaron*** *adenomegalias* (*no adenomegalies have been detected*)	*adenomegalias:* ***no*** (*adenomegalies: no*)
German	***frei*** *von Beschwerden* (*free of symptoms*)	*beschwerde**frei*** (*without symptoms*)
German	**nicht** klopfschmerzhaft (*no percussion*) *tenderness*	*Hinweise für eine cerebrale* Metastasierung *gibt es derzeit* ***nicht.*** (*There is no indication of a cerebral metastasis.*)

Table 5.1: Same negation triggers that might precede or follow a finding in Spanish and in German.

Another interesting aspect of German negations are *surrounding triggers*, such as *lehnt ... ab (reject)* and *wies ... zurück (declined).* In many cases it is possible to reduce or shorten the triggers. However, in the case of given examples, a reduction would make the triggers too general, extending them to different meaning: *wies* (without *zurück*) for instance, could mean *to reject*, but also *to verify* in combination with the separated particle *nach.*

Similar to English, in Spanish and in German, negations can be directly bound to a target word as prefix or suffix, such as ***a****morfo (****a****morphous)*, ***des****compuesto (sick)* (in Spanish) and ***un****auffällig (****un****remarkable),* fett**frei** *(*__non__fat*) or motivations**los*** *(**without** motivation*) (in German).

For the Spanish NegEx implementation we translated NegEx triggers and enriched the resulting trigger set. In the case of the German reports our work is based on a previous version of NegEx triggers translated to German [44]. We conducted the following modifications: 1) we corrected and extended the trigger set, 2) we extended the regular expressions to possible expansions, and 3) we classified the triggers according to their position relative to the findings. Our work differs from Chapman et al. [44] in that they provide the German NegEx triggers, but do not evaluate NegEx on German texts. The German triggers developed for our German NegEx implementation are available in `http://macss.dfki.de/german_trigger_set.html`. We plan to release soon our Spanish triggers for public use.

Results mentioned in this chapter have been published in [243, 57, 56]. Part of them were presented as the master thesis of Vanesa Stricker [242].

The rest of the chapter is organized as follows. We first present in Section 5.2 related work in the detection of negation and speculation terms with a focus on the biomedical domain and in Spanish and German languages. Then, in Section 5.3 we present our main contributions, by explaining the algorithms implemented and the NegEx adaptations carried out for negation and speculation detection. We also present the datasets used and provide an analysis of their characteristics and of the types of negation and speculation terms present in the gold standards built by us. Section 5.4 shows the results of evaluating each of the algorithms. We close this chapter presenting Discussions, and Conclusions and Future Work in Sections 5.5 and 5.6 respectively.

5.2 Related work

Negation and speculation detection are problems that are being addressed nowadays. Various workshops and challenges have tackled this topic in the last years. Among them are the *BioNLP'09 Shared Task 3* [141], the *Workshop on Negation and Speculation in Natural Language Processing in 2010*,[8] the *CoNLL 2010 Shared Task* "Learning to Detect Hedges and Their Scope in Natural Language Text" [89], the *2010 i2b2 NLP challenge* [201], the *SEM 2012 Shared Task* "Resolving the Scope and Focus of Negation" [174], the *Negation and Speculation detection in Biomedical Texts* Tutorial in RANLP 2017[9] and the *Workshop of Negation in Spanish in SEPLN 2017.*[10] Diaz [75] published a book about negation and speculation detection in clinical texts. Meystre et al. [169] presented a review of information extraction in biomedical texts, which also addresses negation detection. The detection of negation and speculation has also received attention in other domains [280, 207].

In order to determine if a finding mentioned in a discharge summary is under the scope of negation or speculation, Chapman et al. [43] developed NegEx, a simple and widely used algorithm, which uses regular expressions to detect *triggers* that indicate negation or speculation and a window of words preceding or following each relevant term to determine if the term is under the scope of negation or speculation or not. NegEx has been adapted to Swedish, French, Dutch [231, 71, 1], and Spanish for biomedical texts, for radiology reports and for EHRs [53, 243, 220].[11] NegEx triggers have been extended for Swedish, French and German [44].

Several methods were built upon this simple algorithm. Wu et al. [284] developed a word-based radiology report search engine based in a modification of NegEx. Harkema et al. [114] developed ConText, a NegEx-based tool, that employs a different definition for the scope of triggers. It also expands the detection of negation of findings to three new categories: hypothetical, historical, and experienced and works with different genres of medical reports, including radiology.

Besides NegEx and Context, a wide range of other methods exist. They are mainly based in rules, predominantly using syntactic knowledge, and in machine learning. Huang and Lowe [125] manually construct grammar rules using PoS tagging in order to detect negations in radiology reports. Mehrabi et al. [167] develop DEEPEN, a negation detection algorithm that uses dependency parsing to reduce NegEx false positives and improve results in complex structured sentences. Sohn et al. [237] create rules for negation detection based on the analysis of negation paths of a dependency parser. Mutalik et al. [187] developed Negfinder to identify patterns of negations present in EHRs. Therefore, terms of interest are tagged with UMLS, negation terms are identified, and grammar rules are used to identify their scope and determine if affects terms of interest. Negfinder has been implemented as a web service [96]. The scope of negation with the use of dependency syntactic structures is addressed in [18].

Among the machine learning techniques are following: Uzuner et al. [260] compare a NegEx extension with a machine learning technique that uses lexical and

[8] http://www.clips.ua.ac.be/NeSpNLP2010/program.html (accessed Dec. 2017).

[9] Negation And Speculation detection In Biomedical Texts Tutorial, RANLP (Recent Advances in Natural Language Processing) 2017 http://lml.bas.bg/ranlp2017/tutorials.php#cruz (accessed Dec. 2017).

[10] SEPLN (Spanish society of natural language processing), Taller de NEGación en Español http://sepln2017.um.es/neges.html (accessed Mar. 2018).

[11] One of the Spanish adaptations was performed by us and will be described later.

syntactic information using two corpora of discharge summaries and one of radiology reports. Cruz Díaz et al. [62] compare ML to a regular expression-based method. Machine learning systems to find the scope of negation [181, 175] and of hedge terms [176] in biomedical texts have also been implemented (in order to determine robustness, these systems have been tested with different text types -three subcorpora of BioScope-). Rokach et al. [212] perform automatic negation identification in clinical reports by means of automatically extracting regular expressions and patterns from annotated data and using them to train a decision tree.

Hybrid methods for the detection of scope of speculation were also carried out [171, 290].

The 2010 i2b2/VA assertion classification task does negation and uncertainty detection and extends it to conditional and hypothetical medical problems, indicating also if the problem belonged to a person other than the patient. [201] make a review about the assertion classification task and explains that the most effective systems used support vector machines (SVMs) either with contextual information and dictionaries containing negation and uncertainty terms or with the output of rule-based systems.

Table 5.2 shows the results of some algorithms of negation detection in the biomedical domain.

paper	**lang.**	**genre**	**P**	**R**	**F1**
original NegEx [43]	EN	DS	84.49	77.84	81.02
RadReportMiner [284]	EN	RR	81.00	72.00	76.24
NegEx adap. [231]	SW	EHR	84.5	82.4	83.44
NegEx adap. (2014) [53]	SP	CR, AR	49.47	55.70	52.38
NegEx adap. (2017) [220]	SP	EHR	80.2	68.3	73.8
Negfinder [187]	EN	DS, SN	91.84	95.74	93.75
Huang et al. [125]	EN	RR	98.60	92.60	95.51

Table 5.2: Some results obtained in negation detection tasks in the biomedical domain. Part of the table is also shown in [242]. P refers to precision and R to recall. References for languages: EN: English, SP: Spanish and SW: Swedish. References for genre: AR: scientific articles, CR: case reports, DS: discharge summaries, EHR: electronic health reports, RR: radiology reports and SN: surgical notes.

Wu et al. [285] argue that despite its good results, negation detection in clinical reports is not a solved problem, since its generalization capacity is challenging. They introduce a machine learning method for negation detection in clinical text and compare its performance across domains. Miller et al. [170] also address generalizability issues of negation systems. Therefore, they examine the performance of multiple unsupervised domain adaptation algorithms[12] on clinical negation detection. Szarvas et al. [251] propose a method for cross-genre and cross-domain detection of speculation.

The most well-known corpus for this task is BioScope [267], a publicly available English corpus of biomedical texts annotated for uncertainty, negation and

[12] The authors provide following definition of unsupervised domain adaptation: "Domain adaptation is the task of using labeled data from one domain (the source domain) to train a classifier that will be applied to a new domain (the target domain). (...) When there is no labeled data in the target domain, the task is called unsupervised domain adaptation."

their scopes. It consists of medical reports (radiology reports), biological full papers and biological scientific abstracts. Negation and speculation terms and their scope are annotated in 20,000 sentences. Vincze et al. [268] review biomedical annotated corpora for negation and speculation detection and compare the negation and speculation annotations of the BioScope and Genia Event corpora. Bokharaeian et al. [28] describe the process for annotating the DDI-DrugBank corpus with negation terms and scopes. Other corpus with negation annotation include PropBank.[13]

Morante [173] provides a list of negation terms that occur in English biomedical texts and a description of their scope based on their syntactic context. This description is useful for the annotation of corpora with negation and for automatic negation detection. Its description is based on negation terms that occur in BioScope corpus.

Some guidelines about negation annotation can be seen in [177, 178]. van Son et al. [264] study the morphological or affixal negation for English. Morante and Sporleder [180] provide an overview of how modality and negation have been treated computationally.

With regards to Spanish, previously Costumero et al. [53] performed an adaptation of NegEx to Spanish EHRs extracted from SciELO. It is related to our work, but it is developed for another biomedical genre (texts from EHRs and more formally written than usual EHRs) and as far as we know the triggers are not publicly available. For the implementation, NegEx triggers were manually translated into Spanish and were enriched with synonyms and with previously detected negation terms. The NegEx algorithm was not modified. Recently, in 2017, Santiso et al. [220] presented an adaptation of NegEx to Spanish EHRs, that achieves 73.8% F1 in an evaluation carried out with 75 electronic health reports. Triggers used were based on previous works [53, 231][14] and on negation terms found in discharge summaries.

In 2017 a Workshop focused on negation in Spanish texts was carried out by the Spanish Society of NLP (SEPLN). A summary of the articles submitted is presented below. A review of the past and ongoing research in negation detection in Spanish texts is presented by Altuna et al. [7]. Jiménez-Zafra et al. [134] present a publicly available Spanish corpus of sentiment analysis annotated for negation. They previously presented the main sources of disagreement found during the annotation process [132]. Jiménez-Zafra et al. [133] compile the existing Spanish corpora annotated for negation. They also define some negation related concepts relevant for the annotation of negation. The UAM Spanish Treebank [182], a corpus composed of 1,501 syntactically annotated sentences extracted from Spanish newspapers, that includes the annotation of negations and their scope, was also presented.[15] Guzzi et al. [109] present the syntactic-semantic criteria applied for focus detection of negation in Spanish texts. Llanos et al. [158] analyze the negation terms appearing in clinical notes from a Spanish hospital. Analysis about the scope of negation in Spanish and a syntactic treatment of negation for sentiment analysis is carried out in [266].

Recently, many corpora annotations for negation detection in Spanish were carried out. Marimon et al. [165] presented and put publicly available the *Spanish Clinical Record Corpus* (IULA-SCRC), that contains 3,194 sentences extracted from anonymized clinical reports manually annotated with negation markers and their scope. Therefore, some previous work is taken into account [187, 267, 178]. Detail

[13] PropBank http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf (accessed Dec. 2017).

[14] Triggers were translated, when needed.

[15] More information about the UAM Spanish Treebank can be seen in http://www.lllf.uam.es/ING/Treebank.html (accessed Mar. 2018).

about the negation annotation criteria is provided and syntactic and semantic negations are considered. Finally, a comparative review of different proposals to negation annotation in the biomedical domain is presented. Cruz et al. [61] are building the UHU-HUVR corpus, that contains a collection of 604 radiology and anamnesis reports (8,412 sentences). Negation terms, negated concepts and their relation are being annotated. Affixal (also called morphological) negations are also included. Wonsever et al. [283] presented in 2016 a freely available Spanish corpus annotated for factuality and propose a factuality model based on Saurí [221] proposal. They also test two machine learning methods as factuality classifiers. Other works that included negation annotation in Spanish are Casillas et al. [38] for the extraction of adverse drug reactions and IxaMed-GS [198], the annotated dataset for adverse drug reaction in Spanish EHRs, mentioned in Chapter 3, that also includes speculation annotation and, to the best of our knowledge, is not publicly available.

Table 5.3 provides the comparison of different negation annotations in the biomedical domain for texts written in Spanish. Inter annotator agreement, genre, number of annotators and size of the corpora are described.

paper	**year**	**#sent.**	**#ov. sent.**	**#ann.**	**IAA ann1-ann2**	**IAA ann2-ann3**	**genre**
Costumero et al. [53]	2014	422					FT
Our work**	2016	1,000	200	3	0.97	0.96	RR
IULA-SCRC [165]	2017	3,194	500	3	0.85	0.88	ClR
UHU-HUVR*** [61]	2017	8,412		2	0.94*	-	RR, AR

Table 5.3: Annotated Spanish corpora for negation detection in the medical domain. Only IULA-SCRC is publicly available. References: # ann.: number of annotators, ann1,ann2, ann3: annotator 1,2 and 3 # sent.: number of sentences, #ov. sent.: number of sentences annotated by more than one annotator. AR: anamnesis reports, ClR: clinical reports, FT: formal texts (scientific articles and other type of formal articles) and RR: radiology reports. *IAA was measured with Dice coefficient, ** Our work will be presented later in the chapter, *** work in progress.

Table 5.4 shows the number of negation and speculation terms and sentences in different radiology reports corpora in English and Spanish.

With regards to German clinical reports, Bretschneider et al. [30] classify sentences containing pathological and non-pathological findings in German radiology reports. Their approach uses a syntactic-semantic parsing approach. Gros and Stede [108] present Negtopus, a system that identifies negations and their scope in medical diagnoses written in German and in English. As mentioned before, Chapman et al. [44] translated NegEx triggers to German. The work reports, among others, the frequency of occurrence of German triggers in an annotated corpus of German clinical text [279], that, as far as we know, is not available for public use. Both publications, [108] and [44], are related to the work we are presenting later in this chapter. However, Negtopus focuses currently only on negation terms and it has been evaluated

	BioScope [267]	**Our corpus for neg. det.**[a]	**Our annotated corpus**[b]	**UHU-HUVR** [61]
# documents	1954	-	513	276
# sentences	6,383	1,000	4,175	5,347
% neg. sentences	13.55%	22.5%	27.89 %	22.80%
# neg. terms	877	406	1,489	1,985
% spec. sentences	13.39%	3.3%	2.04%	-
# spec. terms	1,189	0	109	-
language	EN	SP	SP	SP

[a] Corpus presented later in this chapter.

[b] Corpus presented in Chapter 3.

Table 5.4: Negation and hedge statistics in radiology reports. EN refers to English, SP to Spanish. # refers to number, neg.: negation, det: detection, spec: speculation.

on a set of solely 12 cardiology reports for German negation detection. NegEx with the German trigger set [44] has not been evaluated and thus its performance is still unknown to us.

5.3 Methods

In this section we introduce the different methods developed to detect negations in radiology reports written in Spanish: a syntactic technique, namely patterns based on dependency trees, and the adaptation of NegEx. We also introduce the adaptation of NegEx to German. The idea underlying the use of dependency trees is to identify patterns of negations in the paths obtained from the dependency parsing of reports, to manually compile negation rules, and use them to determine if a finding is under the scope of a negation or not.

Rules elaborated based on other syntactic techniques -PoS tag patterns and constituent tree patterns- were also developed and presented in Cotik et al. [57] and Stricker [242].

All our methods only take into account the sentence where the term of interest appears in order to determine whether it is negated or not, i.e. it does not use information of other sentences.

The adaptation of NegEx to any language requires having the set of triggers written in this language. In order to evaluate the new system, a gold standard data set is necessary, consisting of medical reports with tagged *findings* and a classification of those *findings* as negated, speculated or affirmed. Given that at the point of doing this work, we did not have an annotated dataset either for German nor for Spanish we had to elaborate a gold standard for each of the languages in order to be able to evaluate our developed methods. We will explain next how we developed each corpus.

In the next section we explain the different algorithms implemented.

5.3.1 Baseline Algorithm

The baseline algorithm uses a dictionary of negation and speculation terms.

If one of those terms co-occur in the same sentence with a previously tagged *finding*, we assume the *finding* is negated or speculated, depending on the classification of the term as negation term or speculation term. Otherwise, we assume it is affirmed. For Spanish the terms were provided by an expert radiologist. For German the terms proceed form a previous annotation task of a different German dataset.

5.3.2 The NegEx algorithm

NegEx is an algorithm developed by Chapman et al. [43] for negation and speculation detection in medical reports that is used to determine whether a clinical finding is absent, suspected or present in a patient according to the medical record description. As previously mentioned, it takes as input sentences, each of them with a previously tagged *finding*, and a list of triggers (*negation and speculation terms*), and as output it determines whether each finding is negated, speculated or affirmed. Each trigger has a label assigned, which determines the scope of the negation or speculation. PREN and POST labels correspond to negation terms that occur before and after the finding respectively. The same occurs with PREP and POSP, referring to speculation terms. CONJ refers to trigger terms that terminate the scope of a negation or speculation and PSEU to pseudo-negations.[16] Examples can be seen in table 5.5. For more information refer to [43]. In order to provide the output, NegEx looks for triggers in the input sentences and based on their corresponding label, it determines if the tagged finding is under the scope of the negation or speculation or not.

In order to determine the performance of the algorithm, NegEx uses a gold standard that consists of a set of sentences with tagged findings and an annotation telling whether the identified terms are negated, speculated or affirmed.

Algorithm 1 describes the original NegEx implementation in pseudocode for negation. The speculation detection algorithm is similar.

[Example] Consider the following tagged ultrasonography report in Spanish and it's translation to English:
”384 —15y 3m—20090412—A423517 Higado: lobulo caudado <FI>aumentado </FI> de tamano, tamano y ecoestructura normal. Via biliar intra y extrahepatica: no <FI>dilatada</FI>. Paredes y contenido normal. Pancreas: tamano y ecoestructura normal. Retroperitoneo vascular: sin <FI>alteraciones</FI>. No se detectaron <FI>adenomegalias</FI>. Ambos rinones de caracteristicas normales. (...)”
(*”384 —15y 3m—20090412—A423517 Liver: <FI>enlarged</FI> caudate lobe, size and echostructure normal. Intra and extrahepatic bile duct: not <FI>dilated</ FI>. Wall and content appear normal. Pancreas: normal size and echotexture. Vascular retroperitoneum: without <FI> changes </ FI>. No <FI> lymphadenopathy </FI> has been detected. Both kidneys of normal characteristics. (...) ”*)

NegEx has as input sentences. Each of them with a tagged finding and an annotation as to whether the finding is negated, speculated or affirmed. For the previous report, the output of NegEx for the second sentence would be:

”384 ***dilatada*** *Via biliar intra y extrahepatica: no dilatada* ***negated*** *Via biliar intra y extrahepatica:* ***PREN*** *no* ***PREN*** *dilatada* ***negated.”***
(*”384* ***dilated*** *Intra and extra hepatic bile duct: not dilated* ***negated*** *Intra and*

[16] If a *finding* is under the scope of a PSEU trigger, NegEx assumes it is affirmed.

Algorithm 1 NegEx algorithm

```
for each sentence do
    for each negation trigger (NT) do
        if NT in PSEU then
            goto next negation trigger of the sentence
        else if NT in PREN then
            // Define a forward scope of NT
            if a (CONJ or a PSEU OR
                      a POST or a PREP or a POSP trigger OR
                      the end of the sentence) are found then
                //Define a forward scope of NT
                End the scope of NT
            else
                Tag sentence as negated
            end if
        else if NT in POST then
            // Define a backward scope for NT
            if a (CONJ or a PSEU OR
                      a PREN or a PREP or a POSP trigger OR
                      the beginning of the sentence) are found then
                End the scope of NT
            else
                Tag sentence as negated
            end if
        end if
    end for
    if sentence not tagged as negated then
        Tag sentence as affirmed
    end if
end for

if sentence not tagged as negated then
    Tag sentence as affirmed
end if

```

label	**meaning**	**example**	**example with context**
PREN	the negation term precedes the finding	*no se evidencia* (there is no evidence), *no se observa* (not observed)	*no se evidencian adenomegalias* (no adenomegalies are evident)
POST	the negation occurs after the finding	*negado* (denied), *descartado* (ruled out)	*la presencia de adenomegalias es descartada* (the presence of adenomegalies is ruled out)
PREP	the speculation precedes the finding	*habría que descartar* (should be ruled out)	*habría que descartar adenomegalias* (adenomegalies should be ruled out)
POSP	the speculation occurs after the finding	*podría ser descartada* (could be rouled out)	*la existencia de apendicitis podría ser descartada* (the presence of appendicitis could be ruled out)
CONJ	indicates the end of scope of the negation trigger	*pero* (but)	*no se detectaron adenomegalias, pero se puede observar un quiste* (no adenomegalies were detected, but a cyst can be observed)
PSEU	can occur in any order with respect to the finding, contain negation triggers but do not negate the clinical condition	*no hay incremento* (no increase)	*no hubo incremento en el tamaño del tumor* (there is no increase in the size of the tumor)

Table 5.5: NegEx labels, their meaning and examples. In some cases, such us in our PREP example, the precedence of triggers with respect to the finding differs in English and in Spanish.

extra hepatic bile duct: ***PREN*** *not* ***PREN*** *dilated* ***negated***”).

The sentence corresponds to report number 384. *”dilated”* is the previously tagged clinical finding and it was manually tagged as *negated* (the first *negated* indicates that). The output of NegEx tells it is negated (second *negated*) and shows in which position of the sentence the trigger appears. Finally, the *PREN* label indicates that the trigger precedes the finding.

The algorithm takes following decisions: if a finding appears more than once in the sentence, and one of the occurrences is negated, the algorithm assumes that all occurrences are negated. If there are many occurrences of the same trigger in the trigger list (with different labels), the algorithm uses the label according to following precedence list: PREN, POST, PREP and POSP.

5.3.3 NegEx adaptation to Spanish

The set of triggers provided by NegEx[17] was translated using automatic translation[18] (since translation is an expensive task and we are not experts in the domain) and revised by two computer scientists, that speak Spanish as native language. Those triggers that were not correctly translated were eliminated or corrected. Given that English lacks grammatical gender, while Spanish has two (male and female), additional trigger instances were generated due to inflectional properties (for example *"no"* was translated to *"ningún"* and *"ninguna"*).

Our Spanish implementation differs from a preceding Spanish implementation [53] and our previous implementation [243] mainly in that:

- some end of scope triggers were added, and
- coordinated negations, that were not taken into account in the English, nor in the Spanish versions were included as a trigger (*ni -nor-*) and NegEx algorithm was modified to consider this term.

Additionally, tests were performed with two different trigger sets:

- NegEx translated -and improved- triggers (described in previous paragraph). A total of 210 translated triggers were obtained.
- triggers obtained by combining translated triggers, a set of bi and trigrams,[19] and a list of triggers provided by a physician expert in the radiology domain. A total of 350 triggers were obtained.

Creation of a Spanish Negation and Speculation Gold Standard

As we mentioned in previous chapters, working with languages different than English has, among others, the difficulty of the lack of annotated datasets and the existence of less developed linguistic tools. In this case, at the moment of working with negation detection we did not have a gold standard for validating the reliability of the model (the gold standard presented in Chapter 3 was created afterwards) and the annotation of negations presented in Section 5.2, did still not exist.

As we mentioned previously, NegEx needs as input sentences with findings previously tagged. In order to tag the sentences with clinical findings, we used SiMREDA Module 1 (see Section 4.4.3) and [55]. Then, sentence segmentation was performed using NLTK [159]. Only sentences with findings were considered to create the annotated corpus.

A summary of the process to obtain the corpus is described next. A set of sentences with clinical findings tagged were randomly selected in such a way that

[17] NegEx: https://code.google.com/p/negex/ (accessed Dec. 2017).

[18] Google Translate https://translate.google.com/ (accessed Dec. 2017).

[19] bi and trigrams were obtained from the 85,621 report dataset (see Data section). Those, whose first word was *no*, were selected and the resulting triggers were manually analyzed in order to discard those that did not correspond to triggers. 94 triggers were obtained.

approximately half of it had terms that indicate negation and half of it did not,[20] and the following steps were performed: 1) we verified manually that sentences were neither the same (among them) nor very similar, 2) segmentation issues -e.g. sentences that were not separated by the sentence tokenizer- were corrected, 3) sentences with findings tagged by the algorithm and that were not considered actual findings by the annotators were eliminated and replaced by new sentences.

Finally, the resulting set of sentences was annotated, informing whether each sentence has negations or speculations with scope over the clinical finding or not. Annotations were performed by an expert of the radiology domain and two computer scientists. Clinical findings, that are under the scope of negation were annotated as *negated*, those associated with speculation terms were annotated as *speculated* and those, whose findings where not under the scope of negation or speculation were annotated as *affirmed*. A summary of the factuality indicators of findings can be seen in Table 5.6. Some sentences were annotated by more than one annotator, with the objective to calculate the Inter Annotator Agreement (IAA) between annotators to measure their level of agreement. As measure for that goal we calculated Cohen's Kappa coefficient [51], presented in Section 3.4.1. The resulting corpus is composed of 1,000 sentences.

indicator	**meaning**
affirmed	the medical condition exists
negated	the medical condition is absent
speculated	it is uncertain whether the medical condition exists
doubt	the medical condition corresponds to the past or the annotator cannot determine if it is present or not

Table 5.6: Factuality indicators of findings.

For results evaluation, *speculated* annotations were considered as *affirmed*, since physicians are interested in retrieving them, and sentences categorized as *doubt* were replaced by other sentences (that were also annotated). For those sentences annotated by two annotators, if there was no agreement among annotators, usually the radiology-expert criterion was respected. In case of doubt, the annotation criteria were revised by the annotators and the annotation was done according to the results of this process.

The annotation process was performed in two stages, so that we could revise the annotation criteria. We first annotated a dataset of 100 sentences, revised the annotations and the agreement among annotators. Based on that, we revised the annotation criteria and then proceeded to the annotations of the 1,000 sentences, that composed the testing dataset.

The analysis of the data and the development of the trigger set were performed in an independent way (annotated negation and speculation terms were not added as triggers).

Figure 5.1 shows the number of sentences annotated by each annotator individually and by more than one annotator in the testing dataset. Kappa coefficient (κ) was calculated for two sets: 1) 100 sentences annotated by computer scientist annotator 1 and radiology domain expert (annotator 3), and 2) 100 sentences annotated

[20] A list of 15 negation terms was used in order to automatically determine the presence of negation terms in the sentences.

by computer scientist annotator 2 and annotator 3. Table 5.7 shows κ measure for the testing dataset. κ measure for the analysis dataset had similar results.

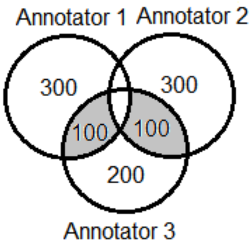


Figure 5.1: Number of sentences annotated by different annotators in the testing dataset.

annotators	κ
A1 and A3	0.97
A2 and A3	0.96

Table 5.7: Inter annotator agreement (IAA) for our Spanish negation dataset. A1 and A2 are computer scientists with a background in BioNLP, A3 is a physician expert in the radiology domain.

Table 5.8 shows the number of clinical findings that are affirmed, negated and speculated in the gold standard. Table 5.9 presents an analysis of the annotated negation terms. The table depicts the most frequent negation triggers and its overall frequency. In all the cases the trigger occurs before the finding.

type of finding	**radiology reports** (%)
affirmed	742 (74.2)
negated	225 (22.5)
speculated	33 (3.3)
findings	1,000

Table 5.8: Number and percentage of affirmed, negated and speculated clinical findings in the Spanish gold standard. Finally, speculated terms were considered as affirmed, so there are 775 affirmed findings.

The annotation of negations in the way we did it is not a complicated annotation task, such as the one described in Chapter 3. No domain-specialists are needed in order to annotate if tagged findings are negated, speculated or affirmed. But, anyway it is costly in time and resources. That is why in a previous work we decided to try the use of a machine learning algorithm in order to create automatic negation annotations, based on a previously human-annotated dataset (i.e. a silver standard). We classified among negated and non-negated terms. Since we were looking for the feasibility of this approach we chose, Naive Bayes, an algorithm that can be considered as a baseline for this method. Alternative models could improve our results. We chose a machine learning toolkit for NLP tasks called MALLET[21] (MAchine Learning for LanguagE Toolkit), that uses the *bag-of-words* model to

[21] MALLET: http://mallet.cs.umass.edu/ (accessed Dec. 2017).

	radiology reports
trigger patterns *(translation,* *frequency)*	no se observ(a |aron |o) (*has/have not been observed*, 32.44 %)
	sin (*without*, 29.78 %)
	no se visualiza(n) (*are not visualized*, 9.33 %)
	no (*no*, 5.78 %)
	ni (*nor*, 5.78 %)
	no se identific(o |a |aron |an) (*are not identified*, 4.89 %)
	no se detect(a |o |an |aron |ron) (*were not detected*, 4 %)

Table 5.9: Most frequent negation terms in the Spanish gold standard.

represent sentences.[22] Results were competitive and could be improved using most sophisticated techniques. They can be seen in [243].

5.3.4 NegEx adaptation to German

For the German version, the translated NegEx triggers provided by Chapman et al. [44] are publicly available and our work is based on them. However, due to various reasons the original translation has been adapted by us.[23] First, in some cases the authors suggest alternative formulations and regular expressions for a trigger. Those alternatives were added to the trigger list and regular expressions were expanded into strings containing all the possible values of the regular expression (the expansions correspond to gender, number and declination variations) (e.g. *kein.*$\{0,2\}$*signifikant.* $\{0,2\}$*(aenderung.* $\{0,2\}$|*Veraenderung.* $\{0,2\}$*)* was expanded to *keine signifikante aendeurng* | *keine signifikanten anderungen*, etc.) (*no significant changes*). Next, a small set of triggers have been changed by an alternative translation. Moreover, new triggers which appeared to be useful were also added to the list. Classification with respect to speculation, proper negation and pseudo-negations and direction of scope was also revised for all triggers (i.e. the appropriate labels were assigned). A set of 506 triggers was obtained.[24] In addition to our trigger set, tests were also performed with the triggers translated by Chapman et al. [44] without modification. In this case the set contains 167 triggers. Alternative translations and regular expressions were not considered.

Creation of a German Negation and Speculation Gold Standard

The data used for the experiments with German language consists of anonymized German discharge summaries and clinical notes of the nephrology domain. Both types of documents (discharge summaries and clinical notes) are written by physicians and have significant differences. Clinical notes are rather short and are written by doctors during or shortly after a visit of a patient. Discharge summaries instead

[22] Bag-of-words defines a dictionary, containing the whole vocabulary included in the training set, in which each word is mapped to a unique position in a vector. The sentences are represented as vectors with the length of the dictionary. Each position, commonly known as feature value, has the amount of occurrences of the word in the sentence.

[23] The team was composed by a computer scientist with expertise in the BioNLP domain, a computational linguists and a student; all German speakers, two of them native speakers.

[24] The link to the trigger data set is available in following url: http://macss.dfki.de/german_trigger_set.html (accessed Mar. 2018).

are written during a stay at the hospital and are prepared for the discharge of the patient. They contain, among others, information about medical history, diagnosis, condition and medication of the patient. Discharge summaries contain much more text compared to clinical notes and often contain longer and more well-formed sentences.

As in other medical reports, both types of documents exhibit non-standard abbreviations, that might include findings and negations among them (e.g. *oB -ohne Befund (without finding)-, opB -ohne pathologischer Befund (without pathological finding)-*).[25] Texts have morphemes representing negation, speculation or findings that are positioned as prefix, suffix or in the middle of a word. Examples are *un**[26], like in *unangenehm -uncomfortable-*, *unklar(e |er |es) -not clear-*, *unverändert(e) -unchanged-* and **los* or **losigkeit*, like in *Appetitslosigkeit -anorexia-* , *Schlaflosigkeit -insomnia-*, and *problemlos(e) -without problems-*. Some of the previous morphemes convert the whole word in findings (such as *Appetitslosigkeit -anorexia-*, *Schlaflosigkeit -insomnia-* and *unangenehm -uncomfortable-*), some in speculated findings (e.g. *unklar -not clear-*) and some are not anymore findings when suffixes are added (e.g. *problemlos -without problems-*).
Table 5.10 provides an overview of the annotated data set used to test our experiment.[27]

	discharge summaries	**clinical notes**
# number of documents	8	175
total amount of words	6,221	6,674
total amount of sentences	1,076	1,158
avg. words per document (std. deviation)	777.63 (322.14)	38.14 (30.49)

Table 5.10: Composition of our German gold standard. Documents correspond to discharge summaries and clinical notes.

Findings were pre-annotated using data of the UMLS Metathesaurus. If a given string can be found in UMLS and its semantic type matches a set of predefined types (Anatomical Abnormality, Congenital Abnormality, Acquired Abnormality, Finding, Sign or Symptom, Pathologic Function, Disease or Syndrome, Mental or Behavioral Dysfunction, Neoplastic Process, Injury or Poisoning), then the string was annotated as a *finding* by the tool. Afterwards, the data was processed by a human annotator. Annotations wrongly made by the tool were removed or corrected and missing concepts were included. Furthermore, the annotator had to decide and annotate whether a given finding occurs in a positive, negative or rather speculative context. Finally, the annotations were corrected by a second -more-experienced-annotator to enhance the quality of the data.

Table 5.11 shows the number and ratio of findings that are affirmed, negated and speculated for the discharge summaries and clinical notes datasets in the German gold standard. It is interesting to note that ratio of affirmed vs. negated is very different in both sets. In clinical notes more than 50% of the findings are negated,

[25] While abbreviations found in German discharge summaries and clinical notes might include findings and negations among them, we did not find abbreviations in Spanish radiology reports with abbreviations containing both negations and an anatomical entity or a finding.

[26] The symbol * implies any combination of letters of the alphabet.

[27] The information was generated by applying a German tokenizer and a sentence splitter. All non alphabetical tokens were removed.

the ratio among the negated and the affirmed findings is much more balanced than in discharge summaries, that contain three times more affirmative than negative findings. Speculations in discharge summaries are very few, but the number is much higher than the number of speculations in clinical notes.

type of finding	**discharge summaries (%)**	**clinical notes (%)**
affirmed	390 (75.29)	255 (42.79)
negated	106 (20.46)	337 (56.54)
speculated	22 (4.25)	4 (0.67)
findings (distinct)	518 (366)	596 (205)

Table 5.11: Number of affirmed, speculated and negated findings in the German gold standard.

Tables 5.12 and 5.13 present an analysis of the annotated negation and speculation terms for each document type. The tables depict the most frequent negation and speculation triggers in combination with trigger order (i.e. the trigger comes before or after the finding) and its overall frequency. Furthermore, the tables present the mean word distance between trigger and finding, including standard deviation (std) and the overall information about how frequently a trigger occurs before (b) or after (a) the clinical finding. Table 5.12, for instance, shows that *kein Nachweis (no evidence)* is used in 14.15% of the cases as negation trigger before the finding. Furthermore, the table shows that the mean word distance between trigger and finding in the discharge summaries is 0.92 with a standard deviation of 1.42. In 97% of the cases the trigger occurs before the finding in the discharge summaries.

	discharge summaries	**clinical notes**
trigger patterns *(translation,* *position, freq.)*	keine *(no, b, 35.85%)* kein *(no, b, 15.09%)* kein Nachweis *(no evidence, b, 14.15%)* ohne *(without, b, 9.43%)* kein Hinweis *(no indication, b, 5.66%)*	keine *(no, b, 64.47%)* kein *(no, b, 27.99%)* keine *(no, a, 3.46%)* kein *(no, a, 0.94%)* ohne *(without, b, 0.63%)*
mean distance (std)	0.92 (1.42)	0.40 (5.62)
position (b/a)	97% / 3%	94% / 6%

Table 5.12: Most frequent negation terms annotated in the German dataset. *a* corresponds to *after*, *b* corresponds to *before*.

The tables show that the variation of triggers in the clinical notes is much smaller compared to the trigger variation in the discharge summaries (with 5 triggers 97.49 % of the clinical notes gold standard negations and 80.18 % of the discharge summaries gold standard negations are covered). This can be explained by the telegraphic style of the clinical notes. In those reports, information is written very quickly, often while the patient is sitting next to the doctor. Due to time pressure and the internal use of the notes, verbose formulations are rare. In this sense clinical notes are very similar to radiology reports.

The analysis of the data and the development of the trigger set were performed in an independent way (annotated negation and speculation terms were not added

	discharge summaries	**clinical notes**
Trigger patterns *(translation, position, freq.)*	Verdacht *(suspicion, b, 30%)* fraglich *(doubtful, b, 10%)* am ehesten *(likely, b, 10%)* wahrscheinlich *(probable, b, 5%)* wahrscheinlich *(probable, a, 5%)*	? *(?, a, 100%)*
mean distance (std)	1.55 (1.64)	0 (0)
position (b/a)	80% / 20%	0% / 100%

Table 5.13: Most frequent speculation terms annotated in the German dataset. *a* corresponds to *after*, *b* corresponds to *before*.

as triggers).

5.3.5 Dependency tree patterns

Dependency parsing, already defined in Section 2.2, allows us to know the syntactic structure of a phrase. The method is based on syntactic context and does not take into account word distance to determine the scope of the negation. Negation patterns were manually created based on syntactic dependency paths in the following way:

1. 30 sentences not belonging to the Spanish gold standard, with clinical findings previously tagged and containing some of the known negation terms *(no, ni, sin) (no, nor, without)* were parsed with a MATE parser trained for Spanish.[28,29] A dependency-based parse tree was obtained for each sentence. For an example of a dependency-based parse tree see Fig. 5.2.,

2. negation terms were located automatically, and an algorithm was developed in order to retrieve the path in the dependency tree between the negation term and the *finding* previously tagged,

3. paths were analyzed and a set of patterns that imply negation of findings was manually developed.

Patterns obtained in the previous step were tested with the gold standard created for our Spanish NegEx adaptation (described in Section 5.3.3).

We detected four patterns, that are described below. *NEG* corresponds to a negation term, *[finding]* and *[anatomical entity]* correspond to previously tagged medical conditions and anatomical entities.[30]

- Pattern 1: sentences of the form "**no** se detectaron adenomegalias" (The Spanish structure of this particular sentence corresponds to *NEG* (no) verb *[finding]*). In these cases the negation has a dependency relation with a word that the finding depends on (see Fig. 5.2).

- Pattern 2: sentences of the form "retroperitoneo vascular: **sin** alteraciones" (vascular retroperitoneum: without alterations) or "tomografía computarizada

[28] The model was obtained as indicated in [15].

[29] MATE and Freeling, were analyzed. Both had its advantages and disadvantages. None of them had optimal results. We chose to use MATE.

[30] Anatomical entities and findings were tagged with SiMREDA Module 1, described in Section 4.4.3 and in Cotik et al. [55].

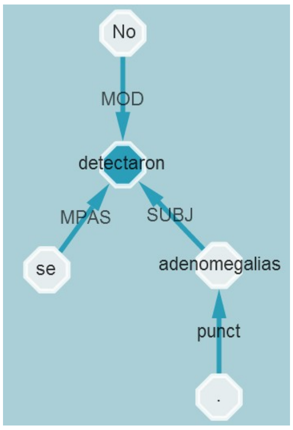


Figure 5.2: Example of a dependency-based parse tree for a sentence of the form of Pattern 1. The parsed sentence is *No se detectaron adenomegalias -no adenomegalies have been detected-*.

encefálica **sin** signos patológicos" (brain computed tomography, without pathological signs) (*[anatomical entity] NEG* (sin) *[finding]*). The finding depends of "sin".

- Pattern 3: sentences like "via biliar **no** dilatada" (*bile duct not dilated*) (*[anatomical entity] NEG [finding]*, where *NEG* is "no").
- Pattern 4: sentences of the form "**No** se detectaron colecciones **ni** liquido libre" (neither collections nor free liquid has been detected) (*NEG*(**no**) verb *[finding] NEG*(**ni**) *[finding]*).

5.4 Results

Next, we will show results of Spanish negation detection and German negation and speculation detection. It is important to notice that for negation detection we take the definition of TP, FP, TN and FN presented in Table 5.14.

	predicted Neg.	**predicted Aff.**
actual Neg	TP	FN
actual Aff	FP	TN

Table 5.14: Confusion matrix for negation detection. *actual* stands for Gold Standard annotation, *predicted* for algorithms output.

5.4.1 Spanish

Table 5.15 shows the performance of our NegEx adaptation to Spanish and our dependency tree pattern method compared to the baseline. We show the best result of NegEx (obtained from the trigger set built from a combination of translated triggers, bi and trigrams and a list of terms suggested by the radiology expert). F1 using NegEx only with translated triggers was similar: 0.91 (220 TP, 36 FP, 5 FN and 739 TN). Results of other syntactic methods based on PoS tagging patterns and constituent tree patterns that we developed previously can be seen in [57].

algorithm	**baseline**	**NegEx (adapted to Spanish)**	**dependency tree patterns**
TP	201	220	194
FP	107	31	61
FN	24	5	31
TN	668	744	714
Accuracy	0.87	0.96	0.91
Precision	0.65	0.88	0.77
Recall	0.89	0.98	0.86
F1	0.75	0.92	0.81

Table 5.15: Performance of different algorithms for negation detection in Spanish radiology reports with testing dataset composed by 1,000 sentences.

trigger type	**trigger**	**number of occurrences**
negation	sin (without)	142
	no se observa(n) (is/are not observed)	80
	no (not)	75
	ni (nor)	36
	no se visualiza(n) (is/are not visualized)	35
	disminuído (diminished)	13
end of scope (conj)	podría corresponder (could correspond)	20

Table 5.16: Triggers used more than five times in Spanish radiology reports.

Table 5.16 shows the negation and speculation triggers that appear more than five times, in the Spanish radiology reports.

Given that only 42 of the 350 triggers are used, and that no more than 7 of them are employed more than five times, we decided to study if the performance of our NegEx adaptation could be maintained with the use of only a subset of our triggers. If the subset is a generic trigger set (i.e. not with specific terms of the radiology domain) it could also be useful to facilitate future adaptations of NegEx to other languages and its application to other domains. *Sin -without-*, *no -not-* and *ni -nor-* are the generic triggers more frequently used. Since the antonym of *sin*, *con -with-*, is probably frequently used to denote the presence of a finding we add it as a PSEU trigger. Some other triggers were included and a set of 16 generic triggers, called *generic trigger set* was built. The list of generic triggers can be seen in Table B.6 of Chapter B.

Table 5.17 shows the results obtained with our generic trigger set of 16 triggers and our compiled trigger set of 350 triggers (the same shown in Table 5.15). There is no much difference among them. This is encouraging, since we suppose that a reduced trigger set can be used for other languages and domains, reducing NegEx adaptation time. Nevertheless, we chose the larger trigger set for our reports, since it has similar precision and higher recall.

measures	**generic trigger set**	**compiled trigger set**
TP	207	220
FP	25	31
FN	18	5
TN	750	744
Accuracy	0.96	0.96
Precision	0.89	0.88
Recall	0.92	0.98
F1	0.91	0.92

Table 5.17: Results of NegEx Spanish implementation with the use of the reduced trigger set (16 triggers) and the use of the full trigger set (350 triggers).

In a previous work we also tested a NegEx implementation carried out by us with a Spanish dataset of formal biomedical texts extracted from the Scientific Electronic Library Online -SciElo-[31], that was used to test a prior NegEx adaptation to Spanish [53]. The dataset was provided by the authors of the article. The results obtained are reported in Table 5.18. The table shows the performance of Costumero et al. [53] implementation as informed in a personal communication,[32] the prior NegEx implementation carried out by us [243] and the NegEx implementation presented in this chapter with the reduced trigger set.

Costumero et al. [53] NegEx adaptation achieved better results that our prior NegEx adaptation. Since the former included as triggers some negation expressions included in their texts, the better results can be motivated by the fact that their triggers are probably better adapted than ours to their texts.

algorithm	NegEx Costumero et al. [53]	NegEx Stricker et al. [243]	NegEx (reduced trigger set)
dataset	SciELO	SciELO	SciELO
F1	0.74	0.67	0.73

Table 5.18: Performance of different Spanish implementations of NegEx with the SciELO dataset used by Costumero et al. [53]. The first column corresponds to the adaptation of NegEx done in [53] (with results obtained from a personal communication), the second column to a prior NegEx adaptation done by us [243] and the third to our adaptation described in this chapter and in [57] with the generic trigger set. All versions are tested with the SciElo dataset selected in [53].

Results of Table 5.18 demonstrate that our NegEx implementation with the reduced trigger set could be used for data different than radiology reports, not only in the medical domain, but probably also in the general domain.

5.4.2 German

In this section we present the negation and speculation detection results of our NegEx adaptation to German (which we call *OTSG* -our trigger set for German-) and the comparison against the original NegEx triggers provided by Chapman

[31] SciElo http://www.scielo.org/php/index.php?lang=en (accessed Dec. 2017).

[32] Results provided by the personal communication outperform those informed in the paper [53].

et al. [44] (which we call *NTSG* -NegEx trigger set for German-) and against our baseline. Results are presented in Table 5.19 and Table 5.20 and evaluated using precision, recall and F1. Furthermore, each table indicates the number of correctly and wrongly predicted instances.

dataset	**discharge summaries**			**clinical notes**		
algorithm	**baseline**	**NegEx**		**baseline**	**NegEx**	
trigger set	–	**NTSG**	**OTSG**	–	**NTSG**	**OTSG**
TP	103	65	99	333	123	328
FP	46	9	13	55	10	19
TN	366	403	399	204	249	240
FN	3	41	7	4	214	9
Accuracy	0.91	0.96	0.96	0.90	0.62	0.95
Precision	0.69	0.88	0.88	0.86	0.92	0.95
Recall	0.97	0.61	0.93	0.99	0.36	0.97
F1	0.81	0.72	0.91	0.92	0.52	0.96

Table 5.19: Performance on the negation detection task for German discharge summaries and clinical notes with the baseline and with NegEx. TP refers to True Positive results, FP to False Positive, TN to True Negatives and FN to False Negatives. NTSG refers to NegEx original German triggers and OTSG to our German trigger set.

dataset	**discharge summaries**			**clinical notes**		
algorithm	**baseline**	**NegEx**		**baseline**	**NegEx**	
trigger set	–	**NTSG**	**OTSG**	–	**NTSG**	**OTSG**
TP	9	0	11	1	0	2
FP	14	0	7	5	5	8
TN	482	496	489	587	587	584
FN	13	22	11	3	4	2
Accuracy	0.95	0.96	0.97	0.99	0.98	0.98
Precision	0.39	0	0.61	0.17	0	0.2
Recall	0.41	0	0.5	0.25	0	0.5
F1	0.4	0	0.55	0.2	0	0.29

Table 5.20: Performance of algorithms for speculation detection in German discharge summaries and clinical notes with the baseline and with NegEx. NTSG refers to NegEx original German triggers and OTSG to our German trigger set.

Table 5.21 shows the negation and speculation triggers that appear more than four times, taking into account discharge summaries and clinical notes.

5.5 Discussion

In this section we discuss results of Spanish negation detection and German speculation and negation detection.

trigger type	trigger	translation	number of occurrences
negation	keine, kein	no	471, 226
	ohne	without	49
	nicht	not	50
	noch	still/yet	40
	aber	but	18
	jedoch	but/however	15
	bis auf	except for	11
	entfernt	removed	7
speculation	verdacht	suspicion	13
	ehesten, eher	rather	13,8
	nicht sicher	not sure	5
	?	?	14

Table 5.21: Negation and speculation triggers used more than four times taking into account all German discharge summaries and clinical notes.

5.5.1 Spanish NegEx and dependency tree patterns

Both, NegEx and the dependency tree pattern algorithms outperform *dictionary lookup*, our baseline algorithm. This makes sense, since the baseline does not take negation scope into account. For example, in "*ectasia* pielica izquierda **sin** cambio de diametro postmiccional" -*left pyelictasis without change in post'void diameter*- the clinical finding (*ectasia*) is not negated, instead *cambio de diametro postmiccional* is under the scope of the negation term *sin*. The baseline algorithm detects the negation (sin) and assumes wrongly that the finding is negated. This scope problem is taken into account in the rest of the algorithms developed.

Dependency tree patters were tested assuming that they would perform better than NegEx in the detection of the negation scope, since it analyzes the structure of the sentence. Nevertheless, NegEx has better results. We understand that two factors influence this situation: 1) the sentences of the reports are usually in our case relatively short (in this dataset they have an average of 14 words and the longest has 74 words). This explains why a simple method like NegEx might be good enough for our data and suggests that we do not need to use more complex methods, that analyze the structure of the sentence. Dependency parsing, that performs an analysis based on the sentence structure, might be left for the most complex sentences. 2) MATE, the tool used to do the dependency parsing was trained on documents from press and several LSP domains (law, economics, computers science, environment and medicine).[33] that include documents of the medical domain, but in this domain, it only includes university handbooks, scientific articles and articles abstracts.

As we previously mentioned, of the 350 triggers of our data only 42 are used and of them and 7 of them more than five times. This fact made us think of a much smaller trigger set with the goal of suggesting easier adaptations of NegEx to other Indo-European languages and the possibility of its use in other domains.

NegEx with the reduced trigger set shows to perform better than a previous implementation for radiology reports in Spanish [243] and similar than an implementation for general medical texts also in Spanish [53] (see Table 5.18).

Sohn et al. [237] results for negation detection in clinical texts in English using

[33] We used a MATE implementation carried out by IULA trained on texts prepared by the Corpus project https://www.upf.edu/web/iula/corpus (accessed Dic. 2017).

dependency parsing are similar to our dependency parser results. They obtain 0.97 precision, 0.74 recall and 0.84 F1, while we obtain 0.77, 0.86 and 0.81 for each of these measures. Nevertheless, it is not easy to compare results with existing papers, since languages and corpora are not the same.

Error analysis

Further analysis of results shows that:

1) the addition of a line of code to NegEx algorithm allows us to handle complex negations. E.g. in "no se detectaron *finding1* ni *finding2*" -"*finding 1* and *finding 2* were not detected"-, when *finding2* is the clinical finding. Those kinds of negations are also handled correctly by the patterns built from our dependency parser, but in some cases, negations are much more complex and are not correctly parsed by the dependency parsing algorithm.

2) Sometimes, negations are not affecting the clinical finding, but a modifier of it and the algorithm tags the clinical finding as negated. For example, in "Riñón derecho sin diferenciación córtico-medular e hiperecogenicidad focal en el polo superior con sombra acústica posterior compatible con *litiasis*." - "Right kidney without corticomedullary differentiation and focal hypoechogenicity in the upper pole with posterior acoustic shadowing compatible with *lithiasis*"-. The trigger *sin* (*without*) is applied to *diferenciación córtico-medular* (*corticomedullary differentiation*), but the clinical finding is *litiasis* (*lithiasis*).

3) The dependency tree patterns method fails when there is lack of punctuation signs. This shows that the characteristics of the noisy text makes the success of syntactic techniques more complicated.

5.5.2 German NegEx

Results show, that the baseline algorithm provides promising results for the negation detection task in the texts written in German. This might have to do with the fact that in German many of the triggers can be used before or after the finding (see Table 5.1). However, the results show, that in all cases the NegEx adaptation achieves better results compared to the baseline algorithm. In particular, the negation and speculation detection applied to the discharge summaries leads to much better results than the use of the baseline algorithm. This can be explained by the fact that the discharge summaries include a larger variety of triggers, which are not covered by the baseline, but are covered by the German trigger set. Moreover, discharge summaries have longer and more complex sentences, that include *CONJ* triggers, which end the scope of negation. However, the results show, that both algorithms achieve better results using the clinical notes. We believe the reason is related to the fact that clinical notes have much shorter and simpler sentences than the ones of discharge summaries. The test with the original German trigger set achieves lower results than our NegEx adaptation and our baseline. The results improve and are similar to ours (F1=0.92 for discharge summaries and 0.94 for clinical notes) if the trigger *keine* is added to NTSG.

Considering the 506 triggers of our data, only 27 occur in the clinical reports (see the ones used more than four times in Table 5.21). This makes us infer that, as we concluded for Spanish, the translation effort could be avoided in further adaptation of NegEx to other languages.

Error analysis

Reviewing the errors, we found that syntactic analysis could improve our results. For instance, in *kein starker Krampf (no strong cramp)*, *Krampf* is under the scope of *kein (no)*, a *PREN* trigger, but *no* is actually addressing to *strong* and not to *cramp.* The use of part of speech tagging or dependency parsing information could help us avoid this error.

Moreover, the original NegEx speculation triggers did not help us to find speculation. In fact with those triggers no speculation terms have been detected (see Table 5.20). Thus, a number of speculation triggers have been added to OTSG. Triggers were taken from general German knowledge and from the transformation of some of the original negation triggers to their corresponding speculation triggers (e.g. *Ohne Verdacht* -without suspicion- originated *Verdacht -suspicion-*). In particular, we added the trigger *?* as a speculation term occurring after the finding, since we knew it is frequently used in the clinical notes to express uncertainty.

Some false negative results were generated by the abundance of acronyms, some of them indicating negation of findings (e.g. in oB -*ohne Befund, without finding*-, B -*Befund, finding*- was annotated as negated, but we don't have *o-ohne, whithout*- as a trigger).

In all cases negation detection achieves better results than speculation detection. This might be due to the fact that there is much greater variety of triggers for indicating speculation than triggers for indicating negation. Additionally, we detected some missing triggers. In some cases, two classifications of the triggers (e.g *nicht*) were possible (see Table 5.1). For those triggers we missed some correct classifications, where the trigger appeared in the less frequent order (for example *Lymphozele nicht mehr sichtbar, Lymphocele not visible anymore)* was classified as positive, since *nicht* was in the trigger list as a PREN trigger. See also trigger preference list in Section 5.3.2.

Parenthesis and commas were not included as CONJ triggers in our trigger set. After evaluating FP and FN results (see Tables 5.19 and 5.20) tests were performed including them. Including parenthesis and commas as triggers reduces the number of false positives. Consider for example those cases that use the trigger *nicht*: *Hat Nitrendipin nicht vertragen (**Flush**) (Did not tolerate Nitrendipin (flush)). Befinden seit Entlassung nicht gebessert, hat weiterhin **Diarrhoe*** (Condition has not been improved since discharge, has still diarrhoea). In the previous examples the findings *Flush* and *Diarrhoe* are out of the scope of negation and therefore misclassified. We also could have avoided false negatives in speculation detection in cases such as *keine Oedeme (...) (serom?) (no edema (...) (serum?))*, because with our trigger set *serum?* is under the scope of *kein.* In a subsequent test, we included parenthesis and commas as CONJ triggers, which increased F1 of clinical notes to 0.98 and F1 of discharge summaries to 0.94 for negations and F1 of clinical notes to 0.62 (with a recall of 1) and F1 of discharge summaries to 0.58 for speculation.

As explained above, clinical notes are much shorter than discharge summaries. The language is less verbose, often just consisting of sequences of noun phrases with some embedded prepositional phrases. Discharge summaries in contrast contain more verbs and full sentences. Thus, it is not surprising when our analysis of triggers shows that the term *kein(e) -no-* as a negative determiner is much more often used in clinical notes (571 vs. 128) whereas the sentence negation *nicht -not-* occurs more often in discharge summaries (32 vs 18).

5.6 Conclusion

This chapter presented the approach taken to solve the negation detection problem in Spanish radiology reports and the detection and speculation problem in two types of German clinical reports. For Spanish two approaches were introduced: an adaptation of NegEx and manually built patterns based on the analysis of dependency tree parsing. For German NegEx was implemented based on a revised version of an existing German NegEx trigger set, that had, as far as we know not been tested.

For both languages, results outperformed a dictionary look-up algorithm, that was implemented and taken as a baseline. Nevertheless, the baseline has good results, especially for the German clinical notes. The Spanish NegEx adaptation worked better than the solution based on dependency tree patterns. We assume that this is influenced by the shortness of reports, that makes the negation detection task easier than in longer reports; the abundance of errors in sentences, such as lack of punctuation signs and lack of verbs, that makes the construction of dependency trees difficult, and, finally the fact that MATE, the tool used to do the dependency parsing was trained based on general domain texts, including documents of the medical domain, but not restricted to them. We believe that for longer sentences, the use of dependency tree patterns would be useful. Therefore, dependency tree patterns should be benefited with a training done with text specific of the medical domain and that contains complex negations.

Our German NegEx adaptation for negations yields very good results. Although not easily comparable (because of being applied to different languages and types of medical reports), results are better than the ones obtained by the original algorithm for English clinical texts and to the adaptations done to Swedish and Spanish (in this last case only for clinical notes, discharge summaries results are similar to results obtained for Spanish). They also outperform results obtained on 12 German cardiology reports carried out by Gros and Stede [108].

The analysis of negations existing in the three types of reports, shows that physicians tend to use a structurally simple and precise language. The degree of lexical variation in the expression of negation and is low. This might explain the good results obtained in medical reports of both languages and suggest that the use of NegEx with other text genres might be more challenging.

In both cases, we believe that the fact of having short sentences with simple syntactic structures helps us to get good results. It should also be considered that our data sets are highly redundant (some negations or negation types occur frequently). In order to improve results, a hybrid method combining syntactic analysis could be used.

Both languages have the additional difficulty of not having publicly available clinical reports, having less annotated data, and having in general less resources.

As Chapman et al. [44] state, the translation of triggers to other languages faces a number of issues. Spanish and German are languages with agglutinative features, where a morpheme representing negation can be added to a word. NegEx does not address this fact. Furthermore, both languages are inflected, so a single term can be translated to many others, because of gender and number agreement. This increased the size of our trigger sets.

In both NegEx adaptations we arrived at the conclusion that only a set of few triggers were used, and this makes us infer that the translation effort could be diminished in upcoming adaptations of NegEx to other languages.

5.6.1 Future work

It would be interesting to detect negation that is represented by bound morphemes (prefix or suffix) of relevant content words. It is not a straightforward task. In German, if a lexeme *lf* stands for a clinical finding, *lf+"los" -without-* should be considered as a negation of the finding, e.g., *schlaflos -without sleeping-*, but also *lf+"los"* or *lf+"losigkeit"* could be included in the thesaurus (e.g. *Appetitslosigkeit (anorexia)* and *Schlaflosigkeit (insomnia)*), and in this case the presence of suffix or infix *los* does not indicate the absence of a finding. van Son et al. [264] study this subject for English.

The implementation of a hybrid methodology, taking the best of NegEx and dependency parsing methods could be carried out. Therefore, dependency parsers should be adapted to these text types.

As future work we would like to evaluate speculation detection in Spanish and use SiMREDA with modules 1, 2 and 3 or CRF and the annotated negations of the dataset presented in Chapter 3 as an input for negation and speculation detection in our Spanish radiology reports.

5.7 Resumen

Una condición clínica mencionada en un informe médico no necesariamente significa que se informa una condición factual, ya que el término que se refiere a la misma podría estar bajo el alcance de la negación o de marcadores de modalidad epistémica. Considérese, por ej., la frase *"Retroperitoneo vascular **sin** <FI>alteraciones </FI>."*. En ella, el hallazgo *alteraciones* está alcanzado por el término *sin*, que denota negación. La detección y determinación del alcance de los términos que expresan negación y especulación permite distinguir hechos de impresiones e hipótesis y determinar qué condiciones están presentes y cuáles están ausentes.

Una gran cantidad de condiciones clínicas descriptas en informes médicos se encuentra negada. En nuestro corpus 56 % de los hallazgos lo están, 27,89 % de las oraciones contiene negaciones y 2,04 % contiene especulaciones. En BioScope, un corpus de textos biomédicos escritos en inglés anotado para negaciones, 13,55 % de las oraciones correspondientes a RR contienen negaciones y 13,30 % contienen especulaciones. Esto sugiere que la detección de negaciones y especulaciones en textos del dominio biomédico es una tarea importante en un proceso de IE médica. También lo es en el dominio general [207] y en otras tareas [280].

En los últimos años creció mucho el interés en el tema y han habido muchos talleres y competencias que lo tratan (algunas son BioNLP'09, NeSpNLP2010, CoNLL 2010, 2010 i2b2, SEM 2012, RANLP 2017, SEPLN 2017). También se publicaron libros y revisiones relacionadas con el tema [75, 169]. Casi todo el trabajo en el área ha sido realizado para el inglés. La aplicación para otros idiomas como el español y el alemán es más difícil por la falta de corpus, la necesidad de obtener traducciones (en algunos casos) y por las características de los lenguajes, como ser los circunfijos en alemán.

En este capítulo mostramos distintas técnicas implementadas para la detección de negaciones en los RR escritos en español. Estas constituyen una adecuación de NegEx [43][34] y una técnica basada en reglas, que detecta patrones de negación inferidos a partir del análisis de caminos en árboles de dependencia. En el contexto de una colaboración con centros de investigación alemanes, tuvimos la posibilidad de obtener otro tipo de informes médicos (resúmenes de alta hospitalaria -también llamados

[34] Referenciado en la introducción.

epicrisis- y notas de evolución clínica) escritos en alemán y decidimos implementar NegEx, la técnica que mejor resultados obtuvo en español para detectar negaciones y especulaciones en dichos informes escritos en alemán. Para ambos idiomas desarrollamos un algoritmo sencillo para utilizarlo de referencia en la comparación con las otras soluciones. El contar con estos tres tipos de informes e implementar la solución para ambos idiomas permitió elaborar conclusiones acerca de las ventajas, desventajas y posibles optimizaciones de la técnica utilizada y la realización de un análisis de su funcionamiento con informes médicos de distintas características en cuanto a longitud y corrección de escritura y a las características de ambas lenguas.

Las adaptaciones de NegEx tuvieron como desafío la no existencia de corpus anotados y la necesidad de traducir y adecuar los triggers.[35] Por otro lado, los analizadores de dependencias para el español no están entrenados específicamente con corpus médicos.

Los triggers de nuestra implementación de NegEx en alemán están disponibles publicamente.[36] En el futuro cercano tenemos previsto publicar nuestros triggers de la implementación de NegEx al español.

[35] Para el alemán hicimos una revisión y adecuación de los triggers provistos por [44].

[36] http://macss.dfki.de/german_trigger_set.html.

Part III

Discussion

CHAPTER 6

Conclusions

In this chapter we summarize the main contributions and conclusions of this thesis, describe some known limitations and possible lines of future work.

6.1 Summary

In this thesis we presented, developed, and evaluated fundamental components of a pipeline of information extraction in the medical domain, specifically for radiology reports written in Spanish. We carried out named entity recognition, negation detection and annotated a corpus, that we plan to put publicly available. The availability of the corpus will enable the comparison of approaches and improvement of information extraction techniques in medical reports written in Spanish. We also carried out negation and speculation detection in German. Each chapter has a detailed section of related work.

In the first part of this document, we introduced and studied the problem of information extraction in clinical reports written in Spanish. We addressed the importance and the difficulty of the subject and we provided an outline of our contributions. We also introduced natural language processing and biomedical text mining, the existing resources in the area, and the previous work and challenges in the domain.

In the second part of this work we presented our contributions. Chapter 3 presents the guidelines created and used for annotating radiology reports, that have the characteristic of being short, with very specific-vocabulary, with abundance of ill-formed sentences and abbreviations and acronyms. These guidelines are useful for future annotation initiatives. Furthermore, we created a corpus annotated for named entity recognition, negation and speculation detection and relation extraction for informal texts of the clinical domain (radiology reports) in Spanish, which we plan to make publicly available. To the best of our knowledge, there exists no publicly available corpus for the extraction of named entities and relations in Spanish medical reports, and the existing corpora for negation detection have been published at the same time we obtained ours. Chapter 4 presents two different techniques for doing named entity recognition in our radiology reports. A technique based on dictionary look-up, improved with morphological analysis and the application of rules based on PoS tagging was developed. The lack of specific vocabulary for radiology in

Spanish made us evaluate the convenience of using less specific resources but providing higher coverage and with the advantage of being available in Spanish (SNOMED CT), instead of specific resources (RadLex). We also implemented a machine learning technique, CRF after having obtained our annotated dataset. We study previous works in the area and different matching criteria for the evaluation of our techniques. We introduce a classification method among reports containing clinical findings and reports not containing them, that we elaborated based on preliminary named entity detection and negation detection results. Chapter 5 deals with negation and speculation detection. We adapted NegEx -a well known negation detection algorithm, initially thought for English clinical reports- to Spanish and compared it with a rule-based method we built based on the detection of negation patterns inferred from the analysis of paths in dependency parse trees and with a baseline implemented by us. Given that NegEx had the best results, we implemented NegEx for German, a language that had -as far as we know- only one limited study for negation detection, and for which we had two kinds of reports available with different characteristics, regarding length and quality of sentence formations. We tested a simplification of NegEx triggers,[1] that can be tried in further implementation of the algorithm in other Indo-European languages. Our German NegEx triggers are publicly available.

The last part of the work presents conclusions, bibliography and appendixes.

6.2 Limitations

This thesis includes different components of the information extraction pipeline that were conceived, developed, evaluated and published over several years, during which we had different annotated datasets available and different versions of the tools developed by us. It would be interesting to evaluate all methods proposed in this thesis with the same dataset and with the last version of our tools. However, each experimental evaluation requires a considerable amount of time. A number of data transformation processes have to be carried out in order to proceed with this task.

The techniques proposed and implemented in this thesis rely on the availability of a large amount of high quality annotations. The complexity of medical reports make the definition of annotation guidelines and the effective annotation of reports a very difficult and time-consuming task. As every manual process, the annotation is error prone and is affected by inconsistency, incompleteness and incorrectness. These errors derive in implementation results that appear to be worse than what they actually are.

6.3 Contributions

We contribute by disseminating an annotation process and schema for clinical reports. The lack of standards for annotation made the guideline definition a very difficult task. So, we believe that our experience could be helpful for other researchers working in similar projects. The implementation of each of the SiMREDA modules, can be used by others to implement solutions in low or medium resource languages or in the cases where there is scarce availability of annotated data. We also propose various ideas to improve its performance. Although a NegEx adaptation for Spanish existed prior to our work [53], the dissemination of our NegEx Spanish adaptation [243, 57] was followed by a number of NegEx implementations

[1] As was mentioned previously, NegEx triggers are a list of negation and speculation terms used by the algorithm.

for Spanish [7] and by projects of negation annotation for Spanish [165, 61]. This demonstrates that it is a current area of research. We also contributed to the information extraction discipline in German clinical reports, by implementing a negation and speculation detection solution with competitive results. Furthermore, we consider that the survey of previous work in acronym expansion, de-identification and challenges in the area, and the analysis of matching criteria for evaluation of our techniques, will contribute to researchers working in comparable tasks. Overall, we believe that the report of each of the decisions taken in our medical IE pipeline will help other researchers interested in developing solutions to similar problems.

For the previously stated reasons, and because of the future publication of our annotated dataset and of the Spanish NegEx triggers, we believe that this work opens the door to further advance in information extraction from Spanish clinical reports.

6.4 Future Work

We believe that our approach could be improved in various ways.

We are planning to review some inconsistencies found in the annotated data and make the data publicly available. Cross-sentence relations will be eliminated and a simplification of the annotation of measurements will be carried out. We will evaluate the addition of attributes to relations (such as negation, speculation and conditionals).

Acronym and abbreviation expansion improve the recall of entity detection algorithms. The use of existing medical abbreviations and acronyms for Spanish [147] could help in the process of expanding abbreviations.[2] The ambiguity of abbreviations, and the lack of use of naming conventions makes it a difficult task.

Correction of text, that can involve punctuation restoration, adding diacritics according to language rules, doing spelling correction using keyboard-distance, edit distance, Soundex[3] and other techniques (see, for example, previous work mentioned in Section 2.6) could be carried out. Having a dictionary with domain specific entries would also help. Our intuition is that this would improve the results of the part of speech tagging used in the named entity detection algorithm and of the dependency parsing used for negation detection.

Also, more effort could be done into extracting semantics of findings based on the analysis of Graeco-Latin morphemes. Consider, for example *linfoadenopatia* (lymphadenopathy), a pathology in the lymph nodes, and *cardiopatía* (cardiopathy), heart disease. Both indicate a pathology and the part of the body where it occurred. With a more thorough analysis of morphemes, than the one we did, we could extract information of the location of the finding, which would enrich the named entity recognition presented in this thesis. See [271] for an idea of how to develop such analysis.

The semantic of multi-word findings could be also analyzed. Consider, for example *pyloric stenosis* and *liver cyst*. Both are findings that are not only indicating a finding (*stenosis* and *cyst*), but also its location (*pylorus* and *liver*). The analysis of the components of multi-word findings could lead to a deeper understanding in

[2] Acronyms and abbreviations provided by the National Academy of Medicine of Colombia `http://dic.idiomamedico.net/Siglas_y_abreviaturas` and by the Spanish Ministry of Health `http://www.redsamid.net/archivos/201612/diccionario-de-siglas-medicas.pdf?0` (both accessed Mar. 2018).

[3] Soundex is a phonetic algorithm, originally thought for English for encoding names according to its pronunciation, but also used to other type of entities. Homophones are to be encoded in the same way. There are Soundex implementations for Spanish [99].

the information extracted. Depending on our definition of relation, we could also argue that this is a sort of relation, since it indicates the location of a finding. Furthermore, *liver cyst* can also be referred as *hepatic cyst*. The availability and use of synonyms would also improve our recall in the entity recognition task.

We worked on the detection of clinical findings, among others. But as we previously mentioned, there are many clinical conditions that are not explicitly mentioned (e.g. splenomegaly and appendicitis can be described not only by explicit reference, but also through indirect information like the measure of the spleen or the visibility of the appendix). Thus, the detection of other entity types and characteristics, such as measures and texture, and the elaboration of rules could help in the detection of clinical conditions not explicitly mentioned. Its evaluation would need a different annotation process.

We are interested in including a module in the IE pipeline for information extraction in Spanish clinical reports for asserting relations between clinical findings and the anatomical entity, where they have been observed. We are currently working on a baseline to detect the *occurs_in* relation among findings and anatomical entities, described in Chapter 3. Therefore, we use the clinical findings and anatomical entities of our annotated dataset and we assume that the co-occurrence of a clinical finding and an anatomical entity in a sentence means that there exists an *occurs_in* relation between these two entities. Our preliminary results yield results with precision greater than 50%.

As future work, we propose to evaluate the named entity recognition, the negation and speculation detection and the classification techniques with the annotated dataset presented in Chapter 3.

Also, deep learning architectures could be used to improve named entity recognition performance.

In our dictionary-based entity recognition algorithm the proposed inverted index method has a drawback. RadLex and SNOMED CT terms are usually composed by many words. It is common of anatomical entities to be embedded in findings, for example *bladder*(AE) is embedded in *traumatic lesion of the bladder* (FI). A heuristic should be used in order to eliminate those words that we know that probably do not belong to the main entity type.

It would be also interesting to improve our classification technique among reports containing findings and reports not containing them. Therefore, our name entity detection and negation algorithms could be used. The presented method could be used as a baseline.

Finally, nowadays we are not working with image information from PACS (picture archiving and communication system), but we have the identifiers to relate the reports to their corresponding images. An independent project is being carried out by other researchers in order to relate the information extracted from reports with the associated images.

6.5 Resumen

A lo largo de esta tesis presentamos, desarrollamos y evaluamos los componentes fundamentales de un proceso de extracción de información en el dominio biomédico, específicamente para informes radiológicos escritos en español. En la primera parte del documento introdujimos y estudiamos el problema de extracción de la información en informes clínicos escritos en español. Abordamos la importancia, la dificultad del tema e hicimos un resumen de nuestras contribuciones. Luego, introdujimos el procesamiento del lenguaje natural y la minería de textos en el dominio de la biomedicina, los recursos existentes en el área, los trabajos previos en el dominio y las competencias organizadas.

En la segunda parte de este trabajo presentamos nuestras contribuciones. El capítulo 3 presenta los lineamientos que hemos creado para anotar los RR. Creamos un corpus anotado para NER, detección de factualidad y extracción de relaciones en informes radiológicos escritos en español. A nuestro buen saber y entender no existen corpus disponibles públicamente a los que se pueda acceder para realizar estas tareas.[4] El capítulo 4 presenta dos técnicas para NER. La primera, basada en la búsqueda en diccionario y en la aplicación de reglas y la segunda basada en CRF. Evaluamos ambas técnicas con criterios de coincidencia total y coincidencia parcial. Por último, introducimos un método de clasificación entre informes que contienen hallazgos clínicos fácticos y aquellos que no los contienen elaborado a partir de resultados preliminares de NER y de detección de negaciones. El capítulo 5 se ocupa de la detección de negación y especulación. Para esto adaptamos NegEx al español y además desarrollamos un método basado en reglas, que detecta patrones de negación inferidos a partir del análisis de caminos en árboles de dependencia. Dado que NegEx obtuvo los mejores resultados, lo implementamos para el alemán, que contaba con escasos recursos para resolver esta problemática. Por último, probamos una simplificación de los triggers de NegEx, que parecen hacerlo más adaptable a otros dominios y a otras lenguas indoeuropeas. La última parte del trabajo presenta conclusiones, bibliografía y apéndices.

Limitaciones

Esta tesis incluye diferentes componentes de un proceso de extracción de información en el área médica que fueron concebidos, desarrollados, evaluados y publicados a lo largo de varios años, durante los cuales tuvimos diferentes conjuntos de datos anotados disponibles y contamos con diferentes versiones de las herramientas desarrolladas por nosotros. Sería interesante evaluar todos los métodos propuestos en esta tesis con el mismo conjunto de datos y con la última versión de nuestras herramientas. Sin embargo, cada evaluación experimental requiere una cantidad considerable de tiempo. Se deben llevar a cabo varios procesos de transformación de datos para continuar con esta tarea.

Las técnicas propuestas e implementadas en esta tesis se basan en la disponibilidad de una gran cantidad de anotaciones de alta calidad. La complejidad de los informes médicos hace que la definición de las pautas de anotación y la anotación efectiva de los informes sea una tarea difícil y muy demandante en tiempo. Como todo proceso manual, la anotación es propensa a errores y se ve afectada por inconsistencias, por incompletitud y por falta de corrección. Estos errores derivan en resultados de implementación que parecen ser peores de lo que realmente son.

Contribuciones

Con este trabajo contribuimos con la implementación de parte fundamental de un proceso de extracción de información médica. Desarrollamos dos técnicas para la detección de entidades anatómicas y hallazgos clínicos para textos en español. Podemos identificar con resultados competitivos si los hallazgos identificados están negados o afirmados. También podemos clasificar a los informes en función de si contienen hallazgos afirmados o no contienen hallazgos clínicos que lo estén. Por último, contribuimos al proceso de extracción de información de informes clínicos

[4] Un corpus para detectar negaciones fue publicado al mismo tiempo en que obtuvimos el nuestro.

escritos en alemán, mediante la implementación de una solución para la detección de negaciones y términos de especulación con resultados competitivos.

Publicamos un proceso de anotación y su correspondiente esquema para RR, que es útil para informes clínicos en general. Creemos que nuestra experiencia podría ser de utilidad para oros investigadores que trabajan en proyectos similares, disminuyendo las dificultades enumeradas en esta tesis. Por otro lado, la implementación de cada uno de los módulos de SiMREDA puede ser utilizada para implementar soluciones en lenguajes de recursos bajos o medios o en los casos donde hay poca disponibilidad de datos anotados. También proponemos varias ideas para mejorar su performance.

Si bien existía una adaptación de NegEx para español antes de nuestro trabajo [53], la implementación y publicación de nuestra adaptación [243, 57] fue seguida unos años después por otra implementación [7] y por proyectos de anotación de negación para español [165, 61]. Esto demuestra que es un área de investigación con vigencia en la actualidad. Consideramos que la revisión de trabajos previos en el área de expansión de acrónimos, anonimización, de las competencias organizadas, el análisis de los criterios de evaluación de coincidencias de los resultados de los métodos, el reporte de cada una de las decisiones tomadas en nuestro proceso de IE médico y la futura publicación de nuestro conjunto de datos anotados y de los triggers de la adecuación de NegEx, será de ayuda a otros investigadores trabajando en tareas relacionadas y contribuirá a mejorar los resultados obtenidos en este trabajo, permitiendo un avance en la extracción de información de informes clínicos escritos en español.

Trabajo futuro

Creemos que este trabajo da lugar a varios otros. A continuación, mencionamos varias líneas de trabajo futuro, algunas de las cuáles estamos encarando actualmente.

Tenemos el objetivo de hacer una revisión de los problemas de calidad encontrados en el corpus anotado, hacerle mejoras a partir de nuestra experiencia con su uso y ponerlo a disposición para uso público.

La expansión de acrónimos y abreviaturas mejoraría la sensibilidad de los algoritmos de detección de entidades. Para hacerlo, se podrían considerar inicialmente compilaciones existentes de estos términos [147][5]. La ambigüedad de las abreviaturas y la falta de uso de abreviaturas estándar en los informes radiológicos dificulta la tarea.

La corrección de texto (agregado de signos de puntuación, corrección de la ortografía) (ver sección 2.6) o el uso de técnicas de coincidencia aproximada, que tienen en cuenta los problemas de ortografía podrían mejorar los resultados de la asignación de categorías gramaticales utilizada en los algoritmos de NER y del árbol de dependencias utilizado para la detección de negaciones.

Se podría continuar con el trabajo de extracción de semántica de los hallazgos clínicos en base al análisis de los morfemas grecolatinos contenidos en los mismos. Por ej., *linfoadenopatía* y *cardiopatía* hacen referencia a una patología y a la parte del cuerpo en la que ocurrió (ganglios linfáticos y corazón). Con un mayor análisis de los morfemas, se podría extraer información acerca de la ubicación del hallazgo, lo que enriquecería el NER presentado. Para esto se puede consultar [271].

[5] Abreviaturas y acrónimos provistos por la academia nacional de medicina de Colombia http://dic.idiomamedico.net/Siglas_y_abreviaturas. Abreviaturas y acrónimos provistos por el Ministerio Español de Sanidad y Consumo http://www.redsamid.net/archivos/201612/diccionario-de-siglas-medicas.pdf?0

También se podría analizar la semántica de hallazgos clínicos compuestos por más de una palabra. Por ej., en *estenosis pilórica* y *quiste hepático* son hallazgos clínicos que además del hallazgo (*estenosis* y *quiste*) indican su ubicación (*píloro* e *hígado*). El análisis de los componentes de hallazgos compuestos por más de una palabra podría ayudar a una comprensión más profunda de información extraída.

Hemos trabajado en la detección de hallazgos clínicos, entre otros, pero hay muchas condiciones clínicas que no se mencionan explícitamente. Por ej., *esplenomegalia* y *apendicitis* pueden ser descriptas no sólo por referencia explícita, pero sino también a través de información indirecta, como ser la medida del bazo o la visibilidad del apéndice. Es por esto que la detección de otras entidades, como medidas y texturas podrían colaborar, junto con la elaboración de reglas, a la detección de condiciones clínicas no mencionadas explícitamente.

Estamos interesados en incluir en nuestro proceso de IE un módulo de extracción de relaciones. Actualmente estamos trabajando en una solución que servirá de base para un desarrollo posterior[6] para la detección de la relación *ocurre en*[7] entre hallazgos clínicos y entidades anatómicas. Para esto usamos las entidades anatómicas (AE) y hallazgos clínicos (FI) de nuestro conjunto de datos anotados y asumimos que la coocurrencia de un AE y un FI en una misma oración implican que el hallazgo ocurre en la AE anotada. Nuestros resultados preliminares arrojan una precisión mayor al 50 %.

Proponemos también la evaluación de NER, detección de negaciones y especulación y la clasificación de informes con un mismo conjunto de datos: aquel presentado en el capítulo 3.

Se podrían utilizar redes neuronales profundas para intentar mejorar la performance en la detección de entidades nombradas.

El módulo de índice invertido de SiMREDA podría mejorarse, teniendo en cuenta que en muchos casos hay AE que están mencionadas en los hallazgos clínicos. Por ej., *vejiga* está incluida en el hallazgo *lesión traumática de la vejiga*. Se podría pensar en una heurística que elimine las palabras que tienen una alta probabilidad de no pertenecer a la categoría con la que el término fue clasificado.

Se podría también mejorar la técnica de clasificación de informes propuesta. Para comenzar, se podrían utilizar los resultados de las últimas versiones de los algoritmos de NER y de detección de negaciones.

Finalmente, se podrían asociar los informes con las imágenes de radiología correspondientes y a partir del análisis de las mismas enriquecer la información extraída.

[6] Nos refererimos a un *baseline*.

[7] Relación *occurs_in* descripta en el capítulo 3.

Bibliography

[1] Zubair Afzal, Ewoud Pons, Ning Kang, Miriam Sturkenboom, Martijn Schuemie, and Jan Kors. ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. *BMC Bioinformatics*, 15(1), 2014. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4264258/.

[2] Zubair Afzal, Saber A. Akhondi, Herman H.H.B.M. van Haagen, Erik M. van Mulligen, and Jan A. Kors. Biomedical concept recognition in French text using automatic translation of English terms. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, 2015. URL http://ceur-ws.org/Vol-1391/36-CR.pdf.

[3] Charu C. Aggarwal and Cheng Xiang Zhai. *Mining Text Data*. Springer Publishing Company, Incorporated, 2012. ISBN 1461432227, 9781461432227.

[4] Sima Ajami and Tayyebe Bagheri. Barriers for adopting electronic health records (EHRs) by physicians. *Acta Inform Med.*, 21(2):129–134, 2013. doi: 10.5455/aim.2013.21.129-134. URL https://www.ncbi.nlm.nih.gov/pubmed/23616868.

[5] Zharko Aleksovski. Testing RadLex for completeness using large database of radiology reports. In *Society for Imaging Informatics in Medicine. Annual Meeting*, 2014.

[6] Marc Colosimo Alexander Yeh, Alexander Morgan and Lynette Hirschman. BioCreAtIvE Task 1A: gene mention finding evaluation. *BMC Bioinformatics*, 6(1), 2005. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1869012/.

[7] Begoña Altuna, Noa P. Cruz, and Carlos Parra. Negation Detection in Spanish: Past, Present and Future. In *Proceedings of the First Workshop about Spanish Negation (NEGES).*, 2017.

[8] Edwin Saldaña Ambulódegui. *Manual de Terminología Médica N 2*. 2012.

[9] Sophia Ananiadou. *Towards a Methodology for Automatic Term Recognition. (Volumes I and II) (Term Banks)*. PhD thesis, 1988. AAID-93194.

[10] Sophia Ananiadou. A methodology for automatic term recognition. In *Proceedings of the 15th Conference on Computational Linguistics - Volume*

2, COLING '94, pages 1034–1038, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. doi: 10.3115/991250.991317. URL https://doi.org/10.3115/991250.991317.

[11] Sophia Ananiadou and John McNaught. *Text Mining for Biology And Biomedicine.* Artech House, Inc., Norwood, MA, USA, 2005. ISBN 158053984X.

[12] Sophia Ananiadou, Carol Friedman, and Jun'ichi Tsujii. Introduction: named entity recognition in biomedicine. *Journal of Biomedical Informatics*, 37(6): 393–395, 2004. doi: 10.1016/j.jbi.2004.08.011. URL https://doi.org/10.1016/j.jbi.2004.08.011.

[13] Sophia Ananiadou, Douglas B. Kell, and Jun'ichi Tsujii. Text mining and its potential applications in systems biology. *Trends in Biotechnology*, 24 (12):571–579, 2006. ISSN 0167-7799. doi: https://doi.org/10.1016/j.tibtech.2006.10.002. URL http://www.sciencedirect.com/science/article/pii/S0167779906002423.

[14] Don D. Anderson. The multilingual entity task a descriptive analysis of Enamex in Spanish. In *TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, VA, USA, May 6-8, 1996*, 1996. URL https://aclanthology.info/papers/X96-1052/x96-1052.

[15] Blanca Arias, Núria Bel, Mercé Lorente, Montserrat Marimón, Alba Milà, Jorge Vivaldi, Muntsa Padró, Marina Fomicheva, and Imanol Larrea. Boosting the Creation of a Treebank. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 9(Suppl 11):775–781, 2014.

[16] Alan R Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc.*, 17 (3):229—-236, 2010. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995713/.

[17] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December 2008. ISSN 0891-2017. doi: 10.1162/coli.07-034-R2. URL http://dx.doi.org/10.1162/coli.07-034-R2.

[18] Miguel Ballesteros, Virginia Francisco, Alberto Díaz, Jesús Herrera, and Pablo Gervás. Inferring the scope of negation in biomedical documents. In *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'12, pages 363–375, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-28603-2. doi: 10.1007/978-3-642-28604-9_30. URL http://dx.doi.org/10.1007/978-3-642-28604-9_30.

[19] Krisztian Balog, Pavel Serdyukov, and Arjen P. de Vries. Overview of the TREC 2010 entity track. In *Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010*, 2010. URL http://trec.nist.gov/pubs/trec19/papers/ENTITY.OVERVIEW.pdf.

[20] Marco Basaldella, Lenz Furrer, Carlo Tasso, and Fabio Rinaldi. Entity recognition in the biomedical domain using a hybrid approach. *Journal of biomedical semantics*, 8(1):51, 2017.

[21] Riza Batista-Navarro, Rafal Rak, and Sophia Ananiadou. Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics. *Journal of cheminformatics*, 7(S1):S6, 2015.

[22] Riza Theresa Batista-Navarro, Rafal Rak, and Sophia Ananiadou. Chemistry-specific features and heuristics for developing a CRF-based chemical named entity recogniser. In *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, volume 2, pages 55–59. Citeseer, 2013.

[23] Leonard Berlin. Communication of the urgent finding. *AJR Am J Roentgenol.*, 166(3):513–515, 1996. URL https://www.ajronline.org/doi/abs/10.2214/ajr.166.3.8623618.

[24] Leonard Berlin. Malpractice issues in radiology. *AJR Am J Roentgenol.*, 169 (4):943–946, 1999. URL https://www.ajronline.org/doi/abs/10.2214/ajr.169.4.9308440.

[25] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009. ISBN 0596516495, 9780596516499.

[26] David Blumenthal. Stimulating the adoption of health information technology. *New England journal of medicine*, 360(15):1477–1479, 2009.

[27] Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richard Farkas, Filip Ginter, and Jan Hajic. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics(TACL)*, 1(11):415–428, 2013. URL https://pdfs.semanticscholar.org/0ec5/46aa8f6f9340020ddaa3f77b7b5b2541b559.pdf.

[28] Behrouz Bokharaeian, Alberto Diaz, Mariana Neves, and Virginia Francisco. Exploring negation annotations in the DrugDDI Corpus. In *Fourth workshop on building and evaluating resources for health and biomedical text processing (BIOTxtM 2014)*, 2014.

[29] Selen Bozkurt, Kemal Gülkesen, and Daniel Rubin. Annotation for information extraction from mammography reports. In *ICIMTH*, pages 183–185, 2013. URL https://web.stanford.edu/group/rubinlab/pubs/BozkurtRubinIE2013.pdf.

[30] Claudia Bretschneider, Sonja Zillner, and Matthias Hammon. Identifying pathological findings in German radiology reports using a syntacto-semantic parsing approach. In *Proc of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013)*, pages 27–35, 2013.

[31] Albert Burger, Duncan Davidson, and Richard Baldock. *Anatomy ontologies for bioinformatics. Principles and practice*, volume 6. 01 2008.

[32] Sigfrido Burgos Caceres. Electronic health records: beyond the digitization of medical files. *Clinics*, 68(8):1077–1078, 2013.

[33] Juan C. Caicedo, Jose G. Moreno, Edwin A. Niño, and Fabio A. González. Combining visual features and text data for medical image retrieval using latent semantic kernels. In *Proceedings of the 11th ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2010, Philadelphia, Pennsylvania, USA, March 29-31, 2010*, pages 359–366, 2010. doi: 10.1145/1743384.1743442. URL http://doi.acm.org/10.1145/1743384.1743442.

[34] Fernando Campos, Fernando Plazzotta, Daniel Luna, Analía Baum, and Fernán González Bernaldo de Quirós. Developing and implementing an interoperable document-based electronic health record. *Studies in health technology and informatics*, 192:1169–1169, 2013.

[35] Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, 2004.

[36] Francisco M. Carrero, José Carlos Cortizo, José María Gómez, and Manuel de Buenaga. In the Development of a Spanish Metamap. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 1465–1466, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458335. URL http://doi.acm.org/10.1145/1458082.1458335.

[37] Philip N Cascade and Leonard Berlin. Malpractice issues in radiology. *AJR Am J Roentgenol.*, 173(6):1439–1442, 1999. URL https://www.ajronline.org/doi/abs/10.2214/ajr.173.6.10584778.

[38] Arantza Casillas, Koldo Gojenola, Alicia Pérez, and Maite Oronoz. Clinical text mining for efficient extraction of drug-allergy reactions. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pages 946–952. IEEE, 2016.

[39] André Coutinho Castilla, A. Bacic, and S. Furuie. Portuguese and Spanish Versions of RadLex: Term Browser and Report Tagging Application. Abstract. *Radiological Society of North America 2007 Scientific Assembly and Annual Meeting*, 2007. URL http://archive.rsna.org/2007/5004975.html.

[40] Elena Castro, Ana Iglesias, Paloma Martínez, and Leonardo Castaño. Automatic identification of biomedical concepts in Spanish-language unstructured clinical texts. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 751–757. ACM, 2010.

[41] Wendy W. Chapman and Kevin Brettonel Cohen. Current issues in biomedical text mining and natural language processing. *Journal of Biomedical Informatics*, 42(5):757–759, 2009.

[42] Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. Evaluation of Negation Phrases in Narrative Clinical Reports. In *Proceedings of AMIA, American Medical Informatics Association Annual Symposium*, page 105, Washington, DC, USA, 2001.

[43] Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34 (5):301–310, 2001. doi: 10.1006/jbin.2001.1029. URL http://dx.doi.org/10.1006/jbin.2001.1029.

[44] Wendy W. Chapman, Dieter Hillert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E. Chapman, Mike Conway, Melissa Tharp, Danielle L. Mowery, and Louise Deléger. Extending the NegEx Lexicon for Multiple Languages. In *Proceedings of the 14th World Congress on Medical and Health Informatics*, pages 677–681, Copenhagen, Denmark, 2013. doi: 10.3233/978-1-61499-289-9-677. URL http://dx.doi.org/10.3233/978-1-61499-289-9-677.

[45] SA Charles. Developing universal electronic medical records. *Gastroenterology & hepatology*, 4(3):193–195, 2008.

[46] Nancy Chinchor. Overview of MUC-7/MET-2. In *Proc. Message Understanding Conference MUC-7*, 1999. URL http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html.

[47] Nancy Chinchor, Lynette Hirschman, and David D. Lewis. Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3). *Association for Computational Linguistics*, 19(3):409–449, 1993. URL http://www.anthology.aclweb.org/J/J93/J93-3001.pdf.

[48] Stéphane Clinchant, Julien Ah-Pine, and Gabriela Csurka. Semantic combination of textual and visual information in multimedia retrieval. In *Proceedings of the 1st International Conference on Multimedia Retrieval, ICMR 2011, Trento, Italy, April 18 - 20, 2011*, page 44, 2011. doi: 10.1145/1991996.1992040. URL http://doi.acm.org/10.1145/1991996.1992040.

[49] N J Clinger, T B Hunter, and B J Hillman. Radiology reporting: attitudes of referring physicians. *Radiology*, 169(3):325–826, 1988. URL https://www.ncbi.nlm.nih.gov/pubmed/3187005.

[50] Aaron M. Cohen and William R. Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6 (1):57–71, 2005. URL https://pdfs.semanticscholar.org/fa64/c424c3ec1e4494ca00bb89222fa1ed5fa474.pdf.

[51] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[52] Trevor Anthony Cohn. *Scaling conditional random fields for natural language processing*. PhD thesis, University of Melbourne, Australia, 2007. URL http://hdl.handle.net/11343/39185.

[53] Roberto Costumero, Federico Lopez, Consuelo Gonzalo-Martín, Marta Millan, and Ernestina Menasalvas. An Approach to Detect Negation on Medical Documents in Spanish. In *Brain Informatics and Health*, pages 366–375, Cham, 2014. Springer International Publishing. ISBN 978-3-319-09891-3.

[54] Viviana Cotik. Information extraction of texts in the biomedical domain. In *AAAI Publications, Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4357–4358, 2015. URL https://www.aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/view/11322/11172.

[55] Viviana Cotik, Darío Filippo, and José Castaño. An Approach for Automatic Classification of Radiology Reports in Spanish. *Studies in health technology and informatics*, 216:634–638, 2015. URL http://ebooks.iospress.nl/publication/40286.

[56] Viviana Cotik, Roland Roller, Feiyu Xu, Hans Uszkoreit, Klemens Budde, and Danilo Schmidt. Negation detection in clinical reports written in German. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016)*, pages 115–124, Osaka, Japan, 2016. The COLING 2016 Organizing Committee. URL http://aclweb.org/anthology/W16-5113.

[57] Viviana Cotik, Vanesa Stricker, Jorge Vivaldi, and Horacio Rodríguez. Syntactic methods for negation detection in radiology reports in Spanish. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing (BioNLP 2016)*, pages 156–165, Berlin, Germany, 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W/W16/W16-2921.pdf.

[58] Viviana Cotik, Darío Filippo, Roland Roller, Feiyu Xu, and Hans Uszkoreit. Creation of an Annotated Corpus of Spanish Radiology Reports. In *Proceedings of 1st WiNLP*, 2017.

[59] Viviana Cotik, Darío Filippo, Roland Roller, and Hans Uszkoreit Feiyu Xu. Annotation of Entities and Relations in Spanish Radiology Reports. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 177–184, 2017. doi: 10.26615/978-954-452-049-6_025. URL https://doi.org/10.26615/978-954-452-049-6_025.

[60] Viviana Cotik, Jorge Vivaldi, and Horacio Rodríguez. Arabic medical entities tagging using distant learning in a multilingual framework. *Journal of King Saud University - Computer and Information Sciences*, 29(2):204–211, 2017. URL https://doi.org/10.1016/j.jksuci.2016.10.004.

[61] Noa Cruz, Roser Morante, Manuel J. Maña López, Jacinto Mata Vázquez, and Carlos L. Parra Calderón. Annotating Negation in Spanish Clinical Texts. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 53–58, Valencia, Spain, April 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-1808.

[62] Noa P. Cruz Díaz, Manuel Jesús Maña López, and Jacinto Mata Vázquez. Aprendizaje Automático Versus Expresiones Regulares en la Detección de la Negación y la Especulación en Biomedicina [Machine Learning versus Regular Expressions in Negation and Speculation Detection in Biomedicine]. *Procesamiento del Lenguaje Natural [Natural Language Processing]*, 45:77–85, 2010.

[63] David Crystal. *English worldwide*, pages 420–439. Cambridge University Press, 2006. doi: 10.1017/CBO9780511791154.010.

[64] H. Cunningham. GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36:223–254, 2002.

[65] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: an Architecture for Development of Robust HLT applications. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 168–175, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073112. URL http://www.aclweb.org/anthology/P02-1022.

[66] Hercules Dalianis and Sumithra Velupillai. De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields. *J. Biomedical Semantics*, 1:6, 2010. doi: 10.1186/2041-1480-1-6. URL https://doi.org/10.1186/2041-1480-1-6.

[67] Pragya A. Dang, Mannudeep K. Kalra, Thomas J. Schultz, Steven A. Graham, and Keith J. Dreyer. Informatics in Radiology: Render: An Online Searchable Radiology Study Repository. *RadioGraphics*, 29(5):1233–1246, 2009. doi: 10.1148/rg.295085036. URL https://doi.org/10.1148/rg.295085036. PMID: 19564253.

[68] David Day, John Aberdeen, Lynette Hirschman, Robyn Kozierok, Patricia Robinson, and Marc Vilain. Mixed-initiative development of language processing systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, pages 348–355, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics. doi: 10.3115/974557.974608. URL https://doi.org/10.3115/974557.974608.

[69] David Day, Chad McHenry, Robyn Kozierok, and Laurel Riek. Callisto: A configurable annotation workbench, 2004.

[70] Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. In *Journal of the American Medical Informatics Association : JAMIA*, volume 18, pages 557–562, 2011. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3168309/.

[71] Louise Deléger and Cyril Grouin. Detecting negation of medical problems in French clinical notes. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 697–702, 2012.

[72] Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772, 2009.

[73] Dina Demner-Fushman, Willie J. Rogers, and Alan R. Aronson. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *Journal of the American Medical Informatics Association*, 24(4):841—-844, 2017. URL https://doi.org/10.1093/jamia/ocw177.

[74] Adrien Depeursinge and Henning Müller. Fusion techniques for combining textual and visual information retrieval. In *ImageCLEF, Experimental Evaluation in Visual Information Retrieval*, pages 95–114. 2010. doi: 10.1007/978-3-642-15181-1_6. URL https://doi.org/10.1007/978-3-642-15181-1_6.

[75] Noa P C Diaz. Negation and speculation detection in medical and review texts. In *SPLN*, volume 13, 2014.

[76] Shipra Dingare, Malvina Nissim, Jenny Finkel, Christopher James Manning, and Claire Grover. A system for identifying named entities in biomedical text: how results from two evaluations reflect on both the system and the evaluations. *Comparative and Functional Genomics*, 6:77 – 85, 2005.

[77] Stefan Dlugolinsky, Marek Ciglan, and Michal Laclavik. Evaluation of named entity recognition tools on microposts. In *IEEE 17th International Conference on Intelligent Engineering Systems*, pages 197–202, 2013. doi: 10.1109/

INES.2013.6632810. URL http://ikt.ui.sav.sk/archive/vega15/AEC06_SCOPUS_IEEE_Dlugolinsky_NER_eval_INES.pdf.

[78] B. Do, A.S. Wu, J. Maley, S., and Biswal. Automatic retrieval of bone fracture knowledge using natural language processing. *J Digit Imaging*, 26(4):709–713, 2013.

[79] Bao H. Do, Andrew Wu, Sandip Biswal, Aya Kamaya, and Daniel L. Rubin. Informatics in Radiology: RADTF: A Semantic Search–enabled, Natural Language Processor–generated Radiology Teaching File. *RadioGraphics*, 30 (7):2039–2048, 2010. doi: 10.1148/rg.307105083. URL https://doi.org/10.1148/rg.307105083. PMID: 20801868.

[80] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ACE) program tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf. ACL Anthology Identifier: L04-1011.

[81] Rezarta Islamaj Dogan, Donald C. Comeau, Lana Yeganova, and W. John Wilbur. Finding abbreviations in biomedical literature: three BioC-compatible modules and four BioC-formatted corpora. *Database*, 2014, 2014. doi: 10.1093/database/bau044. URL https://doi.org/10.1093/database/bau044.

[82] Keith J. Dreyer, Mannudeep K. Kalra, Michael M. Maher, Autumn M. Hurier, Benjamin A. Asfaw, Thomas Schultz, Elkan F. Halpern, and James H. Thrall. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: Validation study. *Radiology*, 234(2):323–329, 2005. doi: 10.1148/radiol.2341040049. URL https://doi.org/10.1148/radiol.2341040049. PMID: 15591435.

[83] Scott L DuValla and Olga V Patterson. A hands-on introduction to natural language processing in healthcare. 2010.

[84] Khaled El Emam. Data Anonymization Practices in Clinical Research. A descriptive study. , University of Ottawa, 2006. URL http://www.ehealthinformation.ca/wp-content/uploads/2014/07/2006-Data-Anonymization-Practices.pdf.

[85] Khaled El Emam, Sam Rodgers, and Bradley Malin. Anonymising and sharing individual patient data. *BMJ*, 350, 2015. URL http://www.bmj.com/content/350/bmj.h1139.

[86] Kurt Junshean Espinosa, Riza Theresa Batista-Navarro, and Sophia Ananiadou. Learning to recognise named entities in tweets by exploiting weakly labelled data. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 153–163, 2016.

[87] Andrea Esuli and Fabrizio Sebastiani. *Evaluating Information Extraction*, volume 6360 of *Lecture Notes in Computer Science*, pages 100–111. Springer, Berlin / Heidelberg, 2010. doi: 10.1007/978-3-642-15998-5_12. URL http://nmis.isti.cnr.it/sebastiani/Publications/CLEF10.pdf.

[88] Erik Faessler and Udo Hahn. Semedico: A comprehensive semantic search engine for the life sciences. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, System Demonstrations*, pages 91–96, 2017. doi: 10.18653/v1/P17-4016. URL https://doi.org/10.18653/v1/P17-4016.

[89] Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. The CoNLL-2010 Shared Task: Learning to Detect Hedges and Their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 1–12, Uppsala, Sweden, 2010. ISBN 978-1-932432-84-8. URL http://dl.acm.org/citation.cfm?id=1870535.1870536.

[90] Pieter Fivez, Simon Suster, and Walter Daelemans. Unsupervised context-sensitive spelling correction of clinical free-text with word and character n-gram embeddings. In *BioNLP 2017, Vancouver, Canada, August 4, 2017*, pages 143–148, 2017. doi: 10.18653/v1/W17-2317. URL https://doi.org/10.18653/v1/W17-2317.

[91] Nelson W. Francis. A standard sample of present-day English for use with digital computers. 1964.

[92] Cláudia Freitas, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira, and Paula Carvalho. Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, 2010. URL http://www.lrec-conf.org/proceedings/lrec2010/summaries/412.html.

[93] Carol Friedman. Towards a comprehensive medical language processing system: methods and issues. In *Proceedings of the AMIA annual fall symposium*, page 595. American Medical Informatics Association, 1997.

[94] Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2): 161–174, 1994.

[95] Carol Friedman, George Hripcsak, Lyudmila Shagina, and Hongfang Liu. Research paper: Representing information in patient reports using natural language processing and the extensible markup language. *JAMIA*, 6(1):76–87, 1999. doi: 10.1136/jamia.1999.0060076. URL https://doi.org/10.1136/jamia.1999.0060076.

[96] Kazuki Fujikawa, Kazuhiro Seki, and Kuniaki Uehara. Negfinder: A web service for identifying negation signals and their scopes. *Information and Media Technologies*, 8(3):884–889, 2013.

[97] Maria Laura Gambarte, Alejandro Lopez Osornio, Marcela Martínez, Guillermo Reynoso, Daniel R. Luna, and Fernán González Bernaldo de Quirós. A practical approach to advanced terminology services in health information systems. In *MEDINFO 2007 - Proceedings of the 12th World Congress on Health (Medical) Informatics - Building Sustainable Health Systems, 20-24 August, 2007, Brisbane, Australia*, pages 621–625, 2007. doi: 10.3233/978-1-58603-774-1-621. URL https://www.hospitalitaliano.org.ar/multimedia/archivos/servicios_attachs/_SHTI129-0621.pdf.

[98] Michele Gentili, Sara Hajian, and Carlos Castillo. A case study of anonymization of medical surveys. In *Proceedings of the 2017 International Conference on Digital Health, London, United Kingdom, July 2-5, 2017*, pages 77–81, 2017. doi: 10.1145/3079452.3079490. URL http://doi.acm.org/10.1145/3079452.3079490.

[99] Exequiel Gershanik. Análisis Fonético en un proceso de Calidad de Datos: Estudio de la fonética del dialecto Rioplatense del Castellano y desarrollo de una técnica de matching fonético. In *PhD Thesis*, 2006.

[100] Aris Gkoulalas-Divanis and Grigorios Loukides. *Overview of Patient Data Anonymization*, pages 9–30. Springer New York, New York, NY, 2013. ISBN 978-1-4614-5668-1. doi: 10.1007/978-1-4614-5668-1_2. URL https://doi.org/10.1007/978-1-4614-5668-1_2.

[101] Aris Gkoulalas-Divanis, Grigorios Loukides, and Jimeng Sun. Publishing data from electronic health records while preserving privacy: A survey of algorithms. *Journal of Biomedical Informatics*, 50:4 – 19, 2014. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2014.06.002. URL http://www.sciencedirect.com/science/article/pii/S1532046414001403. Special Issue on Informatics Methods in Medical Privacy.

[102] Lorraine Goeuriot, Gareth J. F. Jones, Liadh Kelly, Henning Müller, and Justin Zobel. Medical information retrieval: introduction to the special issue. *Inf. Retr. Journal*, 19(1-2):1–5, 2016. doi: 10.1007/s10791-015-9277-8. URL https://doi.org/10.1007/s10791-015-9277-8.

[103] Koldo Gojenola, Maite Oronoz, Alicia Pérez, and Arantza Casillas. IxaMed: Applying Freeling and a Perceptron Sequential Tagger at the Shared Task on Analyzing Clinical Texts. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014.*, pages 361–365, 2014. URL http://aclweb.org/anthology/S/S14/S14-2061.pdf.

[104] Ira Goldstein and Özlem Uzuner. Specializing for predicting obesity and its co-morbidities. *Journal of Biomedical Informatics*, 42(5):873–886, 2009. doi: 10.1016/j.jbi.2008.11.001. URL https://doi.org/10.1016/j.jbi.2008.11.001.

[105] L. Grabenbauer, R. Fraser, J. McClay, N. Woelfl, CB Thompson, J. Cambell, and J. Windle. Adoption of electronic health records: a qualitative study of academic and private physicians and health administrators. *Appl Clin Inform.*, 2(2):165–176, 2011. doi: 10.4338/ACI-2011-01-RA-0003. URL https://www.ncbi.nlm.nih.gov/pubmed/23616868.

[106] Ralph Grishman. Information extraction: Techniques and challenges. In *International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, SCIE '97, pages 10–27, London, UK, UK, 1997. Springer-Verlag. ISBN 3-540-63438-X. URL http://dl.acm.org/citation.cfm?id=645856.669801.

[107] Ralph Grishman and Beth Sundheim. Message understanding conference-6: A brief history. In *16th International Conference on Computational Linguistics, Proceedings of the Conference, COLING 1996, Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996*, pages 466–471, 1996. URL http://aclweb.org/anthology/C96-1079.

[108] Oliver Gros and Manfred Stede. Determining Negation Scope in German and English Medical Diagnoses. In *Nonveridicality and Evaluation*, pages 113–126, 2013.

[109] Eleonora Guzzi, Mariona Taulé, and M. Antonia Martí. Criterios para la detección del foco de la negación en español. In *Proceedings of the First Workshop about Spanish Negation (NEGES).*, 2017.

[110] T. Takagi H. Ao. ALICE: an algorithm to extract abbreviations from MEDLINE. *Journal of the American Medical Informatics Association*, 5(12):576–86, 2005.

[111] Ferris M. Hall. Language of the radiology report: primer for residents and wayward radiologists. *AJR Am J Roentgenol.*, 175(5):1239–1242, 2000. URL https://www.ajronline.org/doi/pdfplus/10.2214/ajr.175.5.1751239?src=recsys&.

[112] Byron Hamilton. *History and Evolution of Electronic Health Records*, pages 1–33. Mc Graw Hill Education, 2013.

[113] Xianpei Han and Le Sun. Global distant supervision for relation extraction. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2950–2956. AAAI Press, 2016. URL http://dl.acm.org/citation.cfm?id=3016100.3016315.

[114] Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. ConText: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports. *Journal of biomedical informatics*, 42(5):839–851, October 2009. ISSN 1532-0480. doi: 10.1016/j.jbi.2009.05.002. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2757457&tool=pmcentrez&rendertype=abstract.

[115] Marti A. Hearst. Untangling text data mining. In *27th Annual Meeting of the Association for Computational Linguistics, University of Maryland, College Park, Maryland, USA, 20-26 June 1999.*, 1999. URL http://www.aclweb.org/anthology/P99-1001.

[116] William Hersh and Ravi Teja Bhupatiraju. TREC Genomics Track Overview. *Text REtrieval Conference (TREC) 2003 Proceedings*, pages 14–23, 2003. URL http://trec.nist.gov/pubs/trec12/papers/GENOMICS.OVERVIEW3.pdf.

[117] William Hersh and Ellen Voorhees. TREC Genomics Special Issue Overview. *Inf. Retr.*, 12(1):1–15, February 2009. ISSN 1386-4564. doi: 10.1007/s10791-008-9076-6. URL http://dx.doi.org/10.1007/s10791-008-9076-6.

[118] William Hersh, Ravi Teja Bhuptiraju, Laura Ross, Phoebe Johnson, Aaron Cohen, and Dale Kraemer. TREC 2004 Genomics Track Overview. *Text REtrieval Conference (TREC) 2004 Proceedings*, pages 1–19, 2004. URL http://trec.nist.gov/pubs/trec13/papers/GEO.OVERVIEW.pdf.

[119] William Hersh, Aaron Cohen, Jianji Yang, Ravi Teja Bhupatiraju, Phoebe Roberts, and Marti Hearst. TREC 2005 Genomics Track Overview. *Text REtrieval Conference (TREC) 2005 Proceedings*, pages 1–26, 2005. URL http://trec.nist.gov/pubs/trec14/papers/GEO.OVERVIEW.pdf.

[120] William Hersh, Aaron Cohen, Phoebe Roberts, and Hari Krishna Rekapalli. TREC 2006 Genomics Track Overview. *Text REtrieval Conference (TREC) 2006 Proceedings*, pages 52–78, 2006. URL http://trec.nist.gov/pubs/trec15/papers/GEO06.OVERVIEW.pdf.

[121] William Hersh, Aaron Cohen, Lynn Ruslen , and Phoebe Roberts. TREC 2007 Genomics Track Overview. *Text REtrieval Conference (TREC) 2007 Proceedings*, pages 1–23, 2007. URL http://trec.nist.gov/pubs/trec16/papers/GEO.OVERVIEW16.pdf.

[122] Alex Hoechsmann. Centralized electronic health records benefit emergency medicine. *Canadian Medical Association Journal*, 184(1):74–74, 2012.

[123] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 541–550, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL http://dl.acm.org/citation.cfm?id=2002472.2002541.

[124] Chung-Chi Huang and Zhiyong Lu. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief Bioinform*, 17(1): 132–144, 2016.

[125] Yang Huang and Henry J Lowe. A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. *Journal of the American Medical Informatics Association*, 14(3):304–311, 2007.

[126] Lawrence Hunter and K. Bretonnel Cohen. Biomedical Language Processing: Perspective What's Beyond PubMed? . *Mol Cell.*, 21(5):589–594, 2006. doi: 10.1016/j.molcel.2006.02.012.

[127] Nancy Ide and James Pustejovsky. *Handbook of Linguistic Annotation*. Springer, 2017.

[128] Ana Iglesias, Elena Castro, Rebeca Pérez, Leonardo Castaño, Paloma Martínez, José Manuel Gómez-Pérez, Sandra Kohler, and Ricardo Melero. Mostas: Un etiquetador morfo-semántico, anonimizador y corrector de historiales clínicos. *Procesamiento del lenguaje Natural*, 41, 2008.

[129] B. Cochran J. Pustejovsky, J. Castaño and M. Morrell. Automatic extraction of acronym-meaning pairs from MEDLINE databases. In *Stud Health Technol Inform*, volume 84, pages 371–75, 2001.

[130] Jingchi Jiang, Yi Guan, and Chao Zhao. WI-ENRE in CLEF eHealth Evaluation Lab 2015: Clinical Named Entity Recognition Based on CRF. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, Toulouse, France, 2015. URL http://ceur-ws.org/Vol-1391/22-CR.pdf.

[131] Ridong Jiang, Rafael E. Banchs, and Haizhou Li. Evaluating and combining named entity recognition systems. In *Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL*, pages 21–27, 2016. URL https://aclweb.org/anthology/W/W16/W16-2703.pdf.

[132] Salud María Jiménez-Zafra, M Teresa Martín-Valdivia, L Alfonso Ureña-López, M Antónia Martí, and Mariona Taulé. Problematic cases in the annotation of negation in Spanish. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*, pages 42–48, 2016. URL http://www.aclweb.org/anthology/W16-5006.

[133] Salud María Jiménez-Zafra, Roser Morante, and L Alfonso M. Teresa Martín-Valdivia andUreña López. Spanish corpora annotated with negation. In *Proceedings of the First Workshop about Spanish Negation (NEGES).*, 2017.

[134] Salud María Jiménez-Zafra, Mariona Taulé, M Teresa Martín-Valdivia, L Alfonso Ureña-López, and M Antónia Martí. Sfu reviewsp-neg: a spanish corpus annotated with negation for sentiment analysis. a typology of negation patterns. *Language Resources and Evaluation*, pages 1–37, 2017.

[135] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Leo Szolovits, Peter Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 2016. URL https://www.nature.com/articles/sdata201635.

[136] María-Luz Terrada José María López Piñero. *Introducción a la terminología médica*. Masson S.A., 2005. URL https://books.google.com.ar/books?id=MVpa_NMSdCMC&printsec=frontcover#v=onepage&q&f=false.

[137] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2000. ISBN 0130950696.

[138] Helena Karsten and Hanna Suominen. Mining of clinical and biomedical text and data: editorial of the special issue. *Int J Med Inform*, 78(12):786–787, 2009.

[139] Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta, and Jun'ichi Tsujii. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, pages 1–8. Association for Computational Linguistics, 2002. URL https://dl.acm.org/citation.cfm?id=1118150.

[140] J. Kim. GENIA corpus- a semantically annotated corpus for bio-textmining. *Bioinformatics*, 10(1):180–182, 2003.

[141] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9, Boulder, Colorado, 2009. ISBN 978-1-932432-44-2. URL http://www.aclweb.org/anthology/W09-1401.

[142] Youngjun Kim, John Hurdle, and Stéphane Meystre. Using UMLS lexical resources to disambiguate abbreviations in clinical text. *Annual Symposium proceedings*, 2011:715–722, 2011. ISSN 1942-597X.

[143] Ari Klein, Abeed Sarker, Masoud Rouhizadeh, Karen O'Connor, and Graciela Gonzalez. Detecting personal medication intake in twitter: An annotated corpus and baseline classification system. In *BioNLP 2017, Vancouver, Canada,*

August 4, 2017, pages 136–142, 2017. doi: 10.18653/v1/W17-2316. URL https://doi.org/10.18653/v1/W17-2316.

[144] Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M van Mulligen, and Dietrich Rebholz-Schuhmann. A multilingual gold-standard corpus for biomedical concept recognition: the mantra gsc. *J Am Med Inform Assoc.*, 22(5), 2015. URL https://www.ncbi.nlm.nih.gov/pubmed/25948699.

[145] C.J. Kuo, M. H. T. Ling, K. T. Lin, and C. N. Hsu. BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature. *BMC bioinformatics*, 7(10 (S15)), 2009.

[146] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289, 2001.

[147] Javier Yetano Laguna. Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias. 2003. URL https://www.amirsalud.com/informacion-relevante/diccionario-siglas-medicas.pdf.

[148] Kenneth H. Lai, Maxim Topaz, Foster R. Goss, and Li Zhou. Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics*, 55:188–195, 2015. doi: 10.1016/j.jbi.2015.04.008. URL https://doi.org/10.1016/j.jbi.2015.04.008.

[149] Paras Lakhani and Curtis P. Langlotz. Automated detection of radiology reports that document non-routine communication of critical or significant results. *J Digit Imaging*, 23(6):647–57, 2009.

[150] R. Leaman and G. Gonzalez. BANNER: An executable survey of advances in biomedical named entity recognition. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 13, pages 652–663, 2008. URL https://www.ncbi.nlm.nih.gov/pubmed/18229723.

[151] Robert Leaman, Rezarta Islamaj Dogan, and Zhiyong Lu. Dnorm: disease name normalization with pairwise learning to rank. In *Bioinformatics*, 2013.

[152] Dennis Lee, Ronald Cornet, Francis Lau, and Nicolette de Keizer. A survey of SNOMED CT implementations. *Journal of Biomedical Informatics*, 46(1):87–96, 2013. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2012.09.006. URL http://www.sciencedirect.com/science/article/pii/S1532046412001530.

[153] Dennis Lee, Nicolette de Keizer, Francis Lau, and Ronald Cornet. Literature review of SNOMED CT use. *J Am Med Inform Assoc.*, 21, 2014. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3957381/pdf/amiajnl-2013-001636.pdf.

[154] Geoffrey Leech. *Adding Linguistic Annotation*, pages 17–29. M. Wynne. Oxbow Books, Oxford, 2004. doi: 10.1007/978-3-642-15998-5_12. URL https://ota.ox.ac.uk/documents/creating/dlc/index.htm.

[155] D. A. Lindberg, B. L. Humphreys, and A. T. McCray. The Unified Medical Language System. *Methods Inf Med.*, 32(4):281–291, 1993. URL https://www.ncbi.nlm.nih.gov/pubmed/8412823.

[156] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher Lin, Xiao Ling, and Daniel Weld. Effective crowd annotation for relation extraction. In *Proceedings of NAACL-HLT 2016*, pages 897–906, 2016.

[157] Hongfang Liu and Carol Friedman. Mining terminological knowledge in large biomedical corpora. In *Proceedings of the 8th Pacific Symposium on Biocomputing, PSB 2003, Lihue, Hawaii, USA, January 3-7, 2003*, pages 415–426, 2003. URL http://psb.stanford.edu/psb-online/proceedings/psb03/liu.pdf.

[158] Leonardo Campillos Llanos, Paloma Martínez, and Isabel Segura Bedmar. A preliminary analysis of negation in a Spanish clinical records dataset. In *Proceedings of the First Workshop about Spanish Negation (NEGES).*, 2017.

[159] Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, volume 1, pages 63–70, Philadelphia, Pennsylvania, 2002. doi: 10.3115/1118108.1118117. URL http://dx.doi.org/10.3115/1118108.1118117.

[160] Bingfeng Luo, Yansong Feng, Zheng Wang, Zhanxing Zhu, Songfang Huang, Rui Yan, and Dongyan Zhao. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 430–439. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1040. URL http://aclanthology.coli.uni-saarland.de/pdf/P/P17/P17-1040.pdf.

[161] Bradley Malin, David Karp, and Richard H. Scheuermann. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *Journal of Investigative Medicine*, 58(1):11–18, 2010. ISSN 1081-5589. doi: 10.2310/JIM.0b013e3181c9b2ea. URL http://jim.bmj.com/content/58/1/11.

[162] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[163] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014. URL https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf.

[164] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Comput. Linguist.*, 19(2):313–330, June 1993. ISSN 0891-2017. URL http://dl.acm.org/citation.cfm?id=972470.972475.

[165] Montserrat Marimon, Jorge Vivaldi, and Núria Bel. Annotation of negation in the IULA Spanish Clinical Record Corpus. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 43–52, Valencia,

Spain, April 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-1807.

[166] R F McLoughlin, C B So, R R Gray, and R Brandt. Radiology reports: how much descriptive detail is enough? *AJR Am J Roentgenol.*, 165 (4):803–806, 1995. URL https://www.ncbi.nlm.nih.gov/pubmed/?term=radiology+report+how+much+descriptive+detail+is+enough.

[167] Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C Max Schmidt, Hongfang Liu, et al. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *Journal of biomedical informatics*, 54:213–219, 2015.

[168] Roberta Merchant, Mary Ellen Okurowski, and Nancy Chinchor. The multilingual entity task (MET) overview. In *TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, VA, USA, May 6-8, 1996*, 1996. URL https://aclanthology.info/papers/X96-1049/x96-1049.

[169] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: A review of recent research. In *Yearb Med Inform*, pages 128–144, 2008. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.5779&rep=rep1&type=pdf.

[170] Timothy Miller, Steven Bethard, Hadi Amiri, and Guergana Savova. Unsupervised domain adaptation for clinical negation detection, 2017. URL http://aclanthology.coli.uni-saarland.de/pdf/W/W17/W17-2320.pdf.

[171] Guillermo Moncecchi, Jean-Luc Minel, and Dina Wonsever. The influence of syntactic information on hedge scope detection. In *Ibero-American Conference on Artificial Intelligence*, pages 83–94. Springer, 2014.

[172] Sungrim Moon, Serguei V. S. Pakhomov, Nathan Liu, James O. Ryan, and Genevieve B. Melton. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *JAMIA*, 21(2):299–307, 2014. doi: 10.1136/amiajnl-2012-001506. URL https://doi.org/10.1136/amiajnl-2012-001506.

[173] Roser Morante. Descriptive analysis of negation cues in biomedical texts. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, 2010. URL http://www.lrec-conf.org/proceedings/lrec2010/summaries/335.html.

[174] Roser Morante and Eduardo Blanco. Sem 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 265–274, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[175] Roser Morante and Walter Daelemans. A Metalearning Approach to Processing the Scope of Negation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 21–29, Boulder, Colorado, 2009. ISBN 978-1-932432-29-9. URL http://dl.acm.org/citation.cfm?id=1596374.1596381.

[176] Roser Morante and Walter Daelemans. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 28–36. Association for Computational Linguistics, 2009.

[177] Roser Morante and Walter Daelemans. Annotating modality and negation for a machine reading evaluation. 2011.

[178] Roser Morante and Walter Daelemans. ConanDoyle-neg: Annotation of negation in Conan Doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul.* Citeseer, 2012.

[179] Roser Morante and Caroline Sporleder, editors. *NeSp-NLP '10: Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL https://www.clips.uantwerpen.be/NeSpNLP2010/nespnlp2010-proceedings.pdf.

[180] Roser Morante and Caroline Sporleder. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2):223–260, 2012.

[181] Roser Morante, Anthony Liekens, and Walter Daelemans. Learning the scope of negation in biomedical texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 715–724. Association for Computational Linguistics, 2008.

[182] Antonio Moreno, Susana López, Fernando Sánchez, and Ralph Grishman. Developing a syntactic annotation scheme and tools for a Spanish treebank. In *Treebanks*, pages 149–163. Springer, 2003. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10.6457&rep=rep1&type=pdf.

[183] C. Morioka, F. Meng, R. Taira, J. Sayre, P. Zimmerman, D. Ishimitsu, J. Huang, L. Shen, and S. El-Saden. Automatic Classification of Ultrasound Screening Examinations of the Abdominal Aorta. *J Digit Imaging*, 29(6): 742–748, 2016.

[184] Danielle L. Mowery, Brett R. South, Lee M. Christensen, Jianwei Leng, Laura-Maria Peltonen, Sanna Salanterä, Hanna Suominen, David Martínez, Sumithra Velupillai, Noémie Elhadad, Guergana Savova, Sameer Pradhan, and Wendy W. Chapman. Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth Challenge 2013, Task 2. *J. Biomedical Semantics*, 7:43, 2016. doi: 10.1186/s13326-016-0084-y. URL https://doi.org/10.1186/s13326-016-0084-y.

[185] D. S. Patterson M.R. Ramaswamy, L. Yin, and B.W. Goodacre. MoSearch: a radiologist-friendly tool for finding-based diagnostic report and image retrieval. *RadioGraphics*, 16(4):923–933, 1996. URL https://doi.org/10.1148/radiographics.16.4.8835980.

[186] Henning Müller and Jayashree Kalpathy-Cramer. The imageclef medical retrieval task at ICPR 2010 - information fusion to combine visual and textual information. In *Recognizing Patterns in Signals, Speech, Images and Videos - ICPR 2010 Contests, Istanbul, Turkey, August 23-26, 2010, Contest Reports*, pages 99–108, 2010. doi: 10.1007/978-3-642-17711-8_11. URL https://doi.org/10.1007/978-3-642-17711-8_11.

[187] Pradeep Mutalik, Aniruddha M. Deshpande, and Prakash M. Nadkarni. Research paper: Use of general-purpose negation detection to augment concept indexing of medical documents: A quantitative study using the UMLS. *JAMIA*, 8(6):598–609, 2001. doi: 10.1136/jamia.2001.0080598. URL https://doi.org/10.1136/jamia.2001.0080598.

[188] Agnieszka Mykowiecka, Małgorzata Marciniak, and Anna Kupść. Rule-based information extraction from patients' clinical data. *Journal of Biomedical Informatics*, 42(5):923 – 936, 2009. Biomedical Natural Language Processing.

[189] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 1(30):3–26, 2007. ISSN 0378-4169. doi: 10.1075/li.30.1.03nad. URL http://nlp.cs.nyu.edu/sekine/papers/li07.pdf.

[190] Aurélie Névéol, Cyril Grouin, Xavier Tannier, Thierry Hamon, Liadh Kelly, Lorraine Goeuriot, and Pierre Zweigenbaum. CLEF eHealth Evaluation Lab 2015 Task 1b: Clinical Named Entity Recognition. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*, 2015. URL http://ceur-ws.org/Vol-1391/inv-pap5-CR.pdf.

[191] Aurélie Névéol, Cyril Grouin, Xavier Tannier, Thierry Hamon, Liadh Kelly, Lorraine Goeuriot, and Pierre Zweigenbaum. CLEF eHealth Evaluation Lab 2015 Task 1b: Clinical Named Entity Recognition. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*, 2015.

[192] Aurélie Névéol, Cyril Grouin, Kevin B Cohen, Thierry Hamon, Thomas Lavergne, Liadh Kelly, Lorraine Goeuriot, Grégoire Rey, Aude Robert, Xavier Tannier, and Pierre Zweigenbaum. Clinical information extraction at the CLEF eHealth evaluation lab 2016. In *Proc of CLEF eHealth Evaluation lab*, Evora, Portugal, September 6th 2016.

[193] Aurélie Névéol, Robert N Anderson, K. Bretonnel Cohen, Cyril Grouin, Thomas Lavergne, Grégoire Rey, Aude Robert, Claire Rondet, and Pierre Zweigenbaum. ICD-10 coding of death certificates in multiple languages: the CLEF eHealth 2016 and 2017 shared tasks. In *NLP WG*, Washington, DC, 2017.

[194] Mariana Neves and Ulf Leser. A survey on annotation tools for the biomedical literature. *Briefings in Bioinformatics*, 15(2):327–340, 2014. URL https://academic.oup.com/bib/article/15/2/327/210719/A-survey-on-annotation-tools-for-the-biomedical.

[195] Natalya F. Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L. Rubin, Margaret-Anne Storey, Christopher G. Chute, and Mark A. Musen. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.*, 37: 170–173, 2009. doi: 10.1093/nar/gkp440. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2703982/.

[196] Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. The Quaero French medical corpus: A ressource for medical entity recognition and normalization. In *In Proc BioTextM, Reykjavik*, 2014.

[197] Maite Oronoz, Arantza Casillas, Koldo Gojenola, and Alicia Perez. Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names. *Lecture Notes in Computer Science, 8259. Progress in Pattern Recognition, ImageAnalysis, ComputerVision, and Applications 18th Iberoamerican Congress, CIARP 2013*, pages 36–543, 2013.

[198] Maite Oronoz, Koldo Gojenola, Alicia Pérez, Arantza Díaz de Ilarraza, and Arantza Casillas. On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics*, 56:318 – 332, 2015.

[199] Özlem Uzuner, Peter Szolovits, and Isaac Kohane. i2b2 workshop on natural language processing challenges for clinical records. In *Proceedings of the Fall Symposium of the American Medical Informatics Association.* Washington, DC., 2006.

[200] Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *J Biomed Inform*, 17(5), 2010.

[201] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–556, 2011.

[202] Abel Laerte Packer. Scielo-an electronic publishing model for developing countries. In *ELPUB*, 1999.

[203] Sergey V. Pakhomov. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 160–167, 2002. URL http://www.aclweb.org/anthology/P02-1021.pdf.

[204] Serguei Pakhomov, Ted Pedersen, and Christopher G Chute. Abbreviation and acronym disambiguation in clinical discourse. In *AMIA Annual Symposium Proceedings*, pages 589–593, 2005. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1560669/.

[205] Venkataraman Palabindala, Amaleswari Pamarthy, and Nageshwar Reddy Jonnalagadda. Adoption of electronic health records and barriers. *Journal of community hospital internal medicine perspectives*, 6(5):32643, 2016.

[206] Thierry Poibeau and Leila Kosseim. Proper name extraction from non-journalistic texts. In *Computational Linguistics in the Netherlands 2000, Selected Papers from the Eleventh CLIN Meeting, Tilburg, November 3, 2000*, pages 144–157, 2000.

[207] Christopher Potts. On The Negativity of Negation. In *Proceedings of Semantics and Linguistic Theory*, volume 20, pages 636–659, New Brunswick, New Jersey, 2011.

[208] Sameer Pradhan, Noémie Elhadad, Brett R. South, David Martínez, Lee M. Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana K. Savova. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. In

Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013., 2013.

[209] Sameer Pradhan, Noémie Elhadad, Brett R South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana Savova. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 2014. ISSN 1067-5027.

[210] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning.* O'Reilly Media, Inc., 2012.

[211] Johan Reitan, Jørgen Faret, Björn Gambäck, and Lars Bungum. Negation scope detection for twitter sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 99–108, Lisboa, Portugal, 2015. Association for Computational Linguistics.

[212] Lior Rokach, Roni Romano, and Oded Maimon. Negation Recognition in Medical Narrative Reports. *Journal of Information Retrieval*, 11(6):1–50, 2008. URL http://www.ise.bgu.ac.il/faculty/liorr/RV2.pdf.

[213] Roland Roller, Feiyu Xu Hans Uszkoreit, Laura Seiffe, Michael Mikhailov, Oliver Staeck, Klemens Budde, Fabian Halleck, and Danilo Schmidt. A fine-grained corpus annotation schema of German nephrology records. *Proceedings of the Clinical Natural Language Processing Workshop*, 28(1):69–77, 2016.

[214] Roland Roller, Nils Rethmeier, Philippe Thomas, Marc Hübner, Hans Uszkoreit, Oliver Staeck, Klemens Budde, Fabian Halleck, and Danilo Schmidt. Detecting Named Entities and Relations in German Clinical Reports. In *Lecture Notes in Computer Science*, volume 10713, pages 115–124, 2018. URL https://link.springer.com/chapter/10.1007/978-3-319-73706-5_12.

[215] Barbara Rosario and Marti Hearst. Multi-way relation classification: Application to protein-protein interaction. In *HLT-NAACL 2005*, 2005.

[216] Barbara Rosario and Marti A. Hearst. Classifying semantic relations in bioscience text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 2004. URL https://doi.org/10.3115/1218955.1219010.

[217] Wael Salloum, Greg Finley, Erik Edwards, Mark Miller, and David Suendermann-Oeft. Deep learning for punctuation restoration in medical reports. In *BioNLP 2017, Vancouver, Canada, August 4, 2017*, pages 159–164, 2017. doi: 10.18653/v1/W17-2319. URL https://doi.org/10.18653/v1/W17-2319.

[218] Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118853.1118877. URL http://dx.doi.org/10.3115/1118853.1118877.

[219] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1119176.1119195. URL http://dx.doi.org/10.3115/1119176.1119195.

[220] Sara Santiso, Arantza Casillas, Alicia Pérez, and Maite Oronoz. Medical entity recognition and negation extraction: Assessment of NegEx on health records in Spanish. In *Bioinformatics and Biomedical Engineering - 5th International Work-Conference, IWBBIO 2017, Granada, Spain, April 26-28, 2017, Proceedings, Part I*, pages 177–188, 2017. doi: 10.1007/978-3-319-56148-6_15. URL https://doi.org/10.1007/978-3-319-56148-6_15.

[221] Roser Saurí. *A factuality profiler for eventualities in text*. Brandeis University, 2008.

[222] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc.*, 17 (5):507—-513, 2010. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995668/.

[223] A. Schwartz and M. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 452–462, 2003.

[224] B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, COLING-04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[225] Burr Settles. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192, 2005. doi: 10.1093/bioinformatics/bti47. URL https://www.ncbi.nlm.nih.gov/pubmed/15860559.

[226] Hagit Shatkay, W. John Wilbur, and Andrey Rzhetsky. Annotation guidelines. 2005. URL https://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/AnnotationGuidelines.pdf. [Online; accessed 28-04-2017].

[227] Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew-Lim Tan. Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13*, pages 49–56. Association for Computational Linguistics, 2003.

[228] Gary F. Simons and Charles D. Fennig. *Ethnologue: Languages of the World, Twenty-first edition*. SIL International, 2018. URL http://www.ethnologue.com.

[229] Matthew S. Simpson and Dina Demner-Fushman. *Mining Text Data*, chapter 14. Biomedical Text Mining: A Survey of Recent Progress. Springer, 2012.

[230] Sunayana Sitaram. Pronunciation modelling for synthesis of low resources languages. In *PhD Thesis*, 2015.

[231] Maria Skeppstedt. Negation Detection in Swedish Clinical Text: An Adaption of NegEx to Swedish. *Journal of Biomedical Semantics*, 2 (Suppl 3):S3, January 2011. ISSN 2041-1480. doi: 10.1186/2041-1480-2-S3-S3. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3194175&tool=pmcentrez&rendertype=abstract.

[232] Maria Skeppstedt, Maria Kvist, Gunnar H. Nilsson, and Hercules Dalianis. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics*, 49:148 – 158, 2014. ISSN 1532-0464.

[233] Genevieve L. Smith. *Quick medical terminology.* John Wiley & Sons, 1984. URL https://www.amazon.com/Quick-Medical-Terminology-Self-teaching-Guides/dp/0471884510.

[234] Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A Struble, Richard J Povinelli, Andreas Vlachos, Jr William A Baumgartner, Lawrence Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Ma na López, Jacinto Mata, and W John Wilbur. Overview of BioCreative II gene mention recognition. *Genome Biol.*, 9(2), 2008. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2559986/.

[235] Jeffrey L. Sobel, Marjorie L. Pearson, Keith Gross, Katherine A. Desmond, Ellen R. Harrison, Lisa V. Rubenstein, William H. Rogers, and Katherine L. Kahn. Information content and clarity of radiologists' reports for chest radiography. *Academic Radiology*, 3(9):709 – 717, 1996. ISSN 1076-6332. doi: https://doi.org/10.1016/S1076-6332(96)80407-7. URL http://www.sciencedirect.com/science/article/pii/S1076633296804077.

[236] Sunghwan Sohn, Donald C Comeau, Won Kim, and W John Wilbur. Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics*, 9(402), 2008. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-402.

[237] Sunghwan Sohn, Stephen Wu, and Christopher G Chute. Dependency parser-based negation detection in clinical narratives. *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2012:1–8, 2012. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3392064/pdf/1-joint_summit_c2012.pdf.

[238] Parikshit Sondhi. A survey on named entity extraction in the biomedical domain. 2008.

[239] William W. Stead, John R. Searle, Henry E. Fessler, Jack W. Smith, and Edward H. Shortliffe. Biomedical informatics: Changing what physicians need to know and how they learn. *J Academic Medicine*, 86(4):429–434, 2011. ISSN 0891-2017. doi: 10.1097/ACM.0b013e3181f41e8c. URL https://journals.lww.com/academicmedicine/Fulltext/2011/04000/.

[240] M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang. SNOMED clinical terms: overview of the development process and project status. In *Proc AMIA Symp*, pages 662–666, 2001. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243297/.

[241] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April 2012. Association for Computational Linguistics.

[242] Vanesa Stricker. Detección de Negaciones en Informes Radiológicos escritos en Español. In *Undergraduate Thesis. Supervisor: Viviana Cotik*, 2016.

[243] Vanesa Stricker, Ignacio Iacobacci, and Viviana Cotik. Negated findings detection in radiology reports in Spanish: an adaptation of NegEx to Spanish. In *IJCAI - Workshop on Replicability and Reproducibility in Natural Language Processing: adaptative methods, resources and software*, Buenos Aires, Argentina, 2015.

[244] Amber Stubbs. Developing specifications for light annotation tasks in the biomedical domain. In *Proceedings of the Workshop on Building and Evaluationg Resources for Biomedical Text Mining, LREC*, 2012.

[245] Amber Stubbs and Özlem Uzuner. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58:20 – 29, 2015. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2015.07.020. URL http://www.sciencedirect.com/science/article/pii/S1532046415001823. Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

[246] Amber Stubbs and Özlem Uzuner. Annotating risk factors for heart disease in clinical narratives for diabetic patients. *J Biomed Inform*, 58, 2015. ISSN 1532-0464.

[247] Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. Annotating temporal information in clinical narratives. *J Biomed Inform*, 46, 2013. ISSN 1532-0464.

[248] Hanna Suominen. CLEFeHealth2012 - The CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, 2012. URL http://ceur-ws.org/Vol-1178/CLEF2012wn-CLEFeHealth-Suominen2012.pdf.

[249] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Found. Trends Mach. Learn.*, 4(4):267–373, April 2012. ISSN 1935-8237. doi: 10.1561/2200000013. URL https://www.nowpublishers.com/article/Details/MAL-013.

[250] György Szarvas, Richárd Farkas, and Róbert Busa-Fekete. State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework. In *J Am Med Inform Assoc.*, volume 14, pages 574–580, 2007. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1975791/.

[251] György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367, 2012.

[252] Koichi Takeuchi and Nigel Collier. Bio-medical entity extraction using support vector machines. *Artificial Intelligence in Medicine*, 33(2):125–137, 2005. URL https://www.sciencedirect.com/science/article/pii/S0933365704001307.

[253] Almas Tasneem and Archana B. A survey on biomedical named entity extraction. *Asian Journal of Engineering and Technology Innovation*, 4(7):25–28, 2016.

[254] Paul Thompson, Sophia Ananiadou, and Jun'ichi Tsujii. *The GENIA Corpus: Annotation Levels and Applications*, pages 1395–1432. Springer Netherlands, Dordrecht, 2017. ISBN 978-94-024-0881-2. doi: 10.1007/978-94-024-0881-2_54. URL https://doi.org/10.1007/978-94-024-0881-2_54.

[255] Richard Tzong-Han Tsai, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang, Ting-Yi Sung, and Wen-Lian Hsu. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7 (92), 2006. doi: 10.1186/1471-2105-7-92. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1402329/.

[256] Yoshimasa Tsuruoka and Jun'ichi Tsujii. Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13*, pages 41–48. Association for Computational Linguistics, 2003.

[257] Yu Usami, Han-Cheol Cho, Naoaki Okazaki, and Jun'ichi Tsujii. Automatic acquisition of huge training data for bio-medical named entity recognition. In *Proceedings of BioNLP 2011 Workshop*, pages 65–73. Association for Computational Linguistics, 2011. URL http://www.aclweb.org/anthology/W11-0208.

[258] Özlem Uzuner, Yuan Luo, and Peter Szolovits. Viewpoint paper: Evaluating the state-of-the-art in automatic de-identification. *JAMIA*, 14(5):550–563, 2007. doi: 10.1197/jamia.M2444. URL https://doi.org/10.1197/jamia.M2444.

[259] Özlem Uzuner, Tawanda C. Sibanda, Yuan Luo, and Peter Szolovits. A de-identifier for medical discharge summaries. *Artificial Intelligence in Medicine*, 42(1):13–35, 2008. doi: 10.1016/j.artmed.2007.10.001. URL https://doi.org/10.1016/j.artmed.2007.10.001.

[260] Özlem Uzuner, Xiaoran Zhang, and Tawanda Sibanda. Machine Learning and Rule-based Approaches to Assertion Classification. *Journal of the American Medical Informatics Association : JAMIA*, 16(1):109–115, 2009. ISSN 1067-5027. doi: 10.1197/jamia.M2950.

[261] Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, 2010.

[262] Özlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R. South. Evaluating the state of the art in coreference resolution for electronic medical records. *JAMIA*, 19(5):786–791, 2012. doi: 10.1136/amiajnl-2011-000784. URL https://doi.org/10.1136/amiajnl-2011-000784.

[263] Erik M. van Mulligen, Zubair Afzal, Saber A. Akhondi, Dang Vo, and Jan A. Kors. Erasmus MC at CLEF eHealth 2016: Concept Recognition and Coding in French Texts. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, 2016. URL http://ceur-ws.org/Vol-1609/16090171.pdf.

[264] CM van Son, CWJ van Miltenburg, R Morante Vallejo, et al. Building a dictionary of affixal negations. 2016.

[265] Sumithra Velupillai, Hercules Dalianis, Martin Hassel, and Gunnar H. Nilsson. Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial. *I. J. Medical Informatics*, 78(12):19–26, 2009. doi: 10.1016/j.ijmedinf.2009.04.005. URL https://doi.org/10.1016/j.ijmedinf.2009.04.005.

[266] David Vilares, Miguel Alonso, and Carlos Gómez Rodríguez. Syntactic treatment of negation for monolingual and multilingual sentiment analysis. In *Proceedings of the First Workshop about Spanish Negation (NEGES).*, 2017.

[267] Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(11), 2008. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2586758/.

[268] Veronika Vincze, György Szarvas, György Móra, Tomoko Ohta, and Richárd Farkas. Linguistic scope-based and biological event-based speculation and negation annotations in the BioScope and Genia Event corpora. *Journal of Biomedical Semantics*, 2(5):S8, 2011.

[269] Jorge Vivaldi. Extracción de candidatos a término mediante combinación de estrategias heterogéneas. In *PhD Thesis. Supervisor: Maria Teresa Cabré Castellví i Horacio Rodríguez Hontoria*, 2001.

[270] Jorge Vivaldi and Horacio Rodríguez. Some notes about the evaluation of terms and term extraction systems. In *IV International Conference on Language Resources and Evaluation. (LREC 2006)*, pages 12–17, 2006. ISBN 2-9517408-2-4.

[271] R. Estopà; J. Vivaldi and M. T. Cabré. Use of Greek and Latin forms for term detection. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2000)*, 78:855–859, 2000.

[272] Ellen M. Voorhees. The TREC medical records track. In *ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics. ACM-BCB 2013, Washington, DC, USA, September 22-25, 2013*, page 239, 2013. doi: 10.1145/2506583.2506624. URL http://doi.acm.org/10.1145/2506583.2506624.

[273] Ellen M. Voorhees and William R. Hersh. Overview of the TREC 2012 medical records track. In *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012*, 2012. URL http://trec.nist.gov/pubs/trec21/papers/MED12OVERVIEW.pdf.

[274] David Fernández Vítores. El español: una lengua viva. Informe 2017. , Instituto Cervantes, 2017. URL https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2017.pdf.

[275] Hanna M. Wallach. Conditional random fields: an introduction. , University of Pennsylvania CIS Technical REport MS-CIS-04-21, 2004. URL http://dirichlet.net/pdf/wallach04conditional.pdf.

[276] Paul Christopher Webster. Centralized, nationwide electronic health records schemes under assault, 2011.

[277] R. Weegar, A. Casillas, A. Diaz de Ilarraza, M. Oronoz, A. Pérez, and K. Gojenola. The impact of simple feature engineering in multilingual medical NER. *Proceedings of the Clinical Natural Language Processing Workshop*, pages 1–6, 2016. URL http://www.aclweb.org/anthology/W16-4201.

[278] Ben Wellner, Matt Huyck, Scott A. Mardis, John S. Aberdeen, Alexander A. Morgan, Leonid Peshkin, Alexander S. Yeh, Janet Hitzeman, and Lynette Hirschman. Research paper: Rapidly retargetable approaches to de-identification in medical records. *JAMIA*, 14(5):564–573, 2007. doi: 10.1197/jamia.M2435. URL https://doi.org/10.1197/jamia.M2435.

[279] Joachim Wermter and Udo Hahn. An Annotated German-Language Medical Text Corpus as Language Resource. In *Proc 4th Intl LREC Conf*, pages 473–476, 2004.

[280] Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP'10, pages 60–68, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[281] W. John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 356(7), 2006. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1559725/.

[282] John R. Wilcox. The written radiology report. *Applied Radiology*, pages 33–37, 2006. URL http://cdn.agilitycms.com/applied-radiology/PDFs/Issues/AR_07-06_Wilcox.pdf.

[283] Dina Wonsever, Aiala Rosá, and Marisa Malcuori. Factuality annotation and learning in Spanish texts. In *LREC*, 2016.

[284] Andrew S. Wu, Bao H. Do, Jinsuh Kim, and Daniel L. Rubin. Evaluation of negation and uncertainty detection and its impact on precision and recall in search. *Journal of digital imaging*, 24(2): 234–242, April 2011. ISSN 1618-727X. doi: 10.1007/s10278-009-9250-4. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3056979&tool=pmcentrez&rendertype=abstract.

[285] Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PloS one*, 9(11):e112774, 2014.

[286] Hua Xu, Peter D. Stetson, and Carol Friedman. A Study of Abbreviations in Clinical Notes. In *AMIA 2007, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 10-14, 2007*, 2007. URL http://knowledge.amia.org/amia-55142-a2007a-1.623841/t-001-1.624626/f-001-1.624627/a-159-1.624649/a-160-1.624646.

[287] Lana Yeganova, Donald C. Comeau, and W. John Wilbur. Machine learning with naturally labeled data for identifying abbreviation definitions. *BMC Bioinformatics*, 6(12 (S3)), 2011.

[288] Daekeun You, Sameer K. Antani, Dina Demner-Fushman, Md. Mahmudur Rahman, Venu Govindaraju, and George R. Thoma. Automatic identification of ROI in figure images toward improving hybrid (text and image) biomedical document retrieval. In *Document Recognition and Retrieval XVIII, part of the IS&T-SPIE Electronic Imaging Symposium, San Jose, CA, USA, January 26-27, 2011, Proceedings*, page 78740K, 2011. doi: 10.1117/12.873434. URL https://doi.org/10.1117/12.873434.

[289] Hong Yu, George Hripcsak, and Carol Friedman. Research paper: Mapping abbreviations to full forms in biomedical articles. *JAMIA*, 9(3):262–272, 2002. doi: 10.1197/jamia.M0913. URL https://doi.org/10.1197/jamia.M0913.

[290] Huiwei Zhou, Xiaoyan Li, Degen Huang, Yuansheng Yang, and Fuji Ren. Voting-based ensemble classifiers to detect hedges and their scopes in biomedical texts. *IEICE TRANSACTIONS on Information and Systems*, 94(10): 1989–1997, 2011.

[291] Pierre Zweigenbaum and Dina Demner-Fushman. *Advanced Literature-Mining Tools*, pages 347–380. Springer New York, New York, NY, 2009. ISBN 978-0-387-92738-1. doi: 10.1007/978-0-387-92738-1_17. URL https://doi.org/10.1007/978-0-387-92738-1_17.

[292] Pierre Zweigenbaum and Natalia Grabar. Automatic acquisition of morphological knowledge for medical language processing. In *Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making*, AIMDM '99, pages 416–422, London, UK, UK, 1999. Springer-Verlag. ISBN 3-540-66162-X. URL http://dl.acm.org/citation.cfm?id=646051.677260.

[293] Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen. Frontiers of biomedical text mining: current progress. *Brief Bioinform.*, 8(5): 358–375, 2007. doi: 10.1093/bib/bbm045.

APPENDIX A

Appendix A - Abbreviations and acronyms

Table A.1 presents the main abbreviations and acronyms used throughout the work.

term	**expansion in English**
ACE	Automatic Content Extraction
AE	anatomical entity
AMIA	American Medical Informatics Association
API	application programming interface
ASCII	American Standard Code for Information Interchange
BioNLP	NLP applied to the biomedical domain
CRF	conditional random fields
CUI	concept unique identifier
EHR	electronic health report
EL	entity linking
ELRA	European Language Resources Association
EMEA	European Medical Agency
EPO	European Patent Office
EU	European Union
FI	clinical finding
FN	false negative
FP	false positive
GATE	General Architecture for Text Engineering
GS	gold standard
HIPAA	Health Insurance Portability Accountability Act
HMM	Hidden Markov Models
hNLP	Health Natural Language Processing Center
IAA	inter-annotator agreement
IBM	International Business Machines
ICD-10	International Statistical Classification of Diseases and Related Health Problems 10th Revision
IE	information extraction
IR	information retrieval
IULA	Institute for Applied Linguistics (Pompeu Fabra University)
LDA	Latent Dirichlet allocation

LDC	Linguistic Data Consortium
LO	location
LOINC	Logical Observation Identifiers Names and Codes
ME	measure
MEDLARS	Medical Literature Analysis and Retrieval System
MeSH	Medical Subject Headings
ML	Machine Learning
MUC	Message Understanding Conferences
NB	Naive Bayes
NCBI	US National Center for Biotechnology Information
NCBO	US National Center for Biomedical Ontology
NER	named entity recognition
NERC	named entity recognition and classification
NLM	US National Library of Medicine
NLP	natural language processing
NLTK	Natural Language Toolkit
PACS	Picture Archiving and Communication System
PAHO	Pan American Health Organization
PoS	part of speech
PP	prepositional phrase
RadLex	radiology lexicon
RR	radiology report
RSNA	Radiological Society of North America
SciELO	Scientific Electronic Library Online
SCTID	SNOMED CT identifier
SEPLN	Spanish Society of Natural Language Processing
SMS	short message service (text message)
SN	SNOMED CT
SNOMED CT	Standard Nomenclature of Medicine - Clinical Terms
SVM	support vector machines
TM	text mining
TN	true negative
TP	true positive
TREC	Text Retrieval Conference
UIMA	Unstructured Information Management applications
UMLS	Unified Medical Language System
UMLS STY	UMLS Semantic Types
US	United States of America
WHO	World Health Organization
WSD	Word Sense Disambiguation

Table A.1: Acronyms and abbreviations.

APPENDIX B

Appendix B - Additional material

This appendix provides additional material. First, in Section B.1, events organized in BioNLP area in 2016 and 2017 are shown. Section B.2 gives further details of some annotation projects carried out as previous work. Then, in Section B.3 a portion of Freeling Spanish tagset is presented and in Section B.4 implementation details about how SNOMED CT terms were retrieved are shown. Section B.5 shows the CRF feature set used. Finally, the generic trigger set for Spanish NegEx is shown in Section B.6.

B.1 Events organized in the BioNLP area in 2016 and 2017

Only in 2016 and 2017 following events related to the BIONLP area have been organized:

- ACM 10th and 11th International Workshops on Data and Text Mining in Biomedical Informatics (DTMBio)[1], co-located with CIKM (ACM Conference on Information and Knowledge Management).[2]
- Negation and Speculation Detection in Biomedical Texts Tutorial[3] and *BioNLP: Biomedical Natural Language Processing Workshop*[4], co-located with RANLP (Recent Advances in Natural Language Processing) 2017.[5]
- 2017 MEDINFO (World Congress on Medical and Health Informatics).[6]
- *International Workshop on Digital Disease Detection using Social Media* workshop in the 2017 IJCNP (International Joint Conference on Natural Language Processing)[7].
- Clinical Natural Language Processing Workshop (ClinicalNLP)[8], *Fifth Work-*

[1] DTMBio 2016: http://dtmbio.net/dtmbio2016/, DTMBio 2017: http://dtmbio.net/dtmbio2017/ (both accessed Nov. 2017).

[2] http://dl.acm.org/event.cfm?id=RE302 (accessed Nov. 2017).

[3] http://lml.bas.bg/ranlp2017/tutorials.php#cruz (accessed Jan. 2018).

[4] http://lml.bas.bg/ranlp2017/bioNLP2017/index.html (accessed Jan. 2018).

[5] RANLP 2017: http://lml.bas.bg/ranlp2017/bioNLP2017/index.html (accessed Jan. 2018).

[6] MEDINFO 2017: http://medinfo2017.medmeeting.org/en (accessed Jan. 2018).

[7] http://ijcnlp2017.org/site/page.aspx?pid=901&sid=1133&lang=en

[8] ClinicalNLP Workshop 2016: http://text-machine.cs.uml.edu/clinical-nlp-2016 (accessed Jan. 2018).

shop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)[9] and the invited talk in the domain: *NLP to support clinical tasks and decisions*, given by Dina Demner-Fushman (U.S. National Library of Medicine) in COLING (International Conference on Computational Linguistics) 2016.[10]

- *Natural Language Processing for Precision Medicine* Tutorial[11] and *BioNLP: 16th Workshop on Biomedical Natural Language Processing Workshop* in ACL (Conference of the Association of Computational Linguistics) 2017,[12] and 15th BioNLP[13] in ACL 2016[14].
- ACM 11th International Workshop on Data and Text Mining Biomedical Informatics (DTMBio) Workshop presented in CIKM (International Conference on Information and Knowledge Management) 2017.
- Taller de NEGación en ESpañol (NEGES) (Workshop of Negation in Spanish Language)[15] in the XXXIII Congreso Internacional de la sociedad española para el procesamiento del lenguaje natural (SEPLN 2017) (33th NLP International Conference of the Spanish Society).[16]

B.2 Details of annotation projects

Details of previous annotation projects, which can be used as an idea for future annotation of biomedical texts are provided below.

B.2.1 Thyme project

As mentioned in Section 3.5, embedded and overlapping entities are annotated. Below we show examples of both types of annotations. Also examples of the relation *location_of* are shown.

As example of embedded entities, the phrase *renal cell carcinoma* should be annotated as *[renal cell]*(AE) and *[renal cell carcinoma]*(FI).

An example of overlapping entities is shown next. The phrase *right lower leg swelling caused by edema*, should be annotated as *[right lower leg swelling]* (SI),[17] *[right lower leg]* (AE), *[leg swelling]* (SI) and *[edema]* (FI).

Some examples of the relation *location of* (taken from the annotation guidelines) are *The patient has gout in the olecranon bursa*: [olecranon bursa] *location_of* [gout], *He was admitted with right leg swelling*: [right leg] *location_of* [leg swelling], *She was diagnosed with breast cancer*: [breast] *location_of* [breast cancer], *The patient had a skin tumor removed from behind his left ear*: [skin] *location_of* [skin tumor], [behind his left ear] *location_of* [skin tumor].

[9] BioTxtM2016: http://www.nactem.ac.uk/biotxtm2016/ (accessed Jan. 2018).

[10] COLING 2016: http://coling2016.anlp.jp/ (accessed Jan. 2018).

[11] ACL 2017 tutorials: http://acl2017.org/tutorials/, Tutorial precision medicine: https://www.microsoft.com/en-us/research/wp-content/uploads/2017/07/1707_tutorial.pdf (both accessed Jan. 2018).

[12] ACL 2017: http://acl2017.org/workshops/ (accessed Jan. 2018).

[13] BioNLP Workshops: http://www.aclweb.org/aclwiki/index.php?title=BioNLP_Workshop (accessed Jan. 2018).

[14] ACL 2016: https://www.aclweb.org/portal/content/acl-2016-annual-meeting-association-computational-linguistics-0 (accessed Jan. 2018).

[15] NEGES Workshop SEPLN 2017: http://sepln2017.um.es/neges.html (accessed Jan. 2018).

[16] SEPLN 2017: http://sepln2017.um.es/index.html (accessed Jan. 2018).

[17] *SI* refers to sign or symptom.

B.2.2 2010 i2b2 annotation guidelines

The 2010 i2b2 annotation guideline for concepts[18] provides, among others following guidelines: a concept to be annotated can include up to one prepositional phrase (PP) following it, if the PP indicates a body part or can be rearranged to eliminate the PP. For instance, "pain in the chest" can be rearranged to "chest pain", so "pain in the chest" should be annotated. In the same way: "changes in mental status" can be rearranged to "mental status changes" and the whole phrase should be annotated. Verbs that describe the outcome of an event, such as "growing" in "the tumor was growing" have to be annotated. Terms that fit semantic rules but that are only used as modifiers in a noun phrase should not be marked. For example in "She developed diabetes", "diabetes" should be annotated, but in "she takes diabetes medication", "diabetes" should not be annotated. Modifiers (except for assertion modifiers) of annotated concepts appearing in the same phrase should be annotated. For instance, "recurrent angina" and "chronic hepatitis" should be annotated and not only "hepatitis" or "angina".

The annotation guidelines for relations[19] provided the annotation instructions used for the evaluation of the relation classification task, whose goal was to determine the type of relation that exists between two concepts in a sentence of the text. The relations annotated for the i2b2/VA challenge were: 1) medical problem-treatment, 2) medical problem-test and 3) medical problem-medical problem. Medical problem-treatment relations are classified into following relationships: a) treatment improves medical problem, b) treatment worsens medical problem, c) treatment causes medical problem, d) treatment is administered for medical problem and e) treatment is not administered because of medical problem. Medical problem-test relations were classified in a) test reveals medical problem and b) test conducted to investigate medical problem. Finally medical problem-medical problem relations imply that a medical problem indicates another medical problem.

B.3 Freeling Spanish tagset

A subset of Freeling Spanish tagset can be seen below:[20]
AQ: qualifying adjective
CC : coordinating conjunction
DA: determiner article
DI: determiner indefinite
DP: determiner possessive
Fpa: pos:punctuation; type:parenthesis; punctenclose:open
Fpt: pos:punctuation; type:parenthesis; punctenclose:close
Fz: pos:punctuation; type:other
NC : Noun common
NP: Noun proper
PR:pronoun relative
RG : adverb general
RN : adverb negative
SP: adposition preposition

[18] 2010 i2b2/ VA challenge evaluation. Concept annotation guidelines https://www.i2b2.org/NLP/Relations/assets/Concept%20Annotation%20Guideline.pdf (accessed Jan. 2018).

[19] 2010 i2b2/VA challenge evaluation. Relation annotation guidelines https://www.i2b2.org/NLP/Relations/assets/Relation%20Annotation%20Guideline.pdf (accessed Jan. 2018).

[20] Freeling Spanish tagset: https://talp-upc.gitbooks.io/freeling-user-manual/content/tagsets/tagset-es.html (accessed Nov. 2017).

VMI: verb main indicative
VMP: verb main participle
VMS: verb main subjunctive

B.4 Retrieval of SNOMED CT terms. Implementation details.

In Section 4.4.7 we described how we retrieved SNOMED CT Spanish descriptions of the chosen SNOMED CT ids.

SNOMED CT descriptions are documented in a *description file*. The *typeId* field present in this file identifies whether the description is a fully specified name (FSN), a synonym or has another description type. Fully specified names of a concept and their synonyms have the same SCTID. In order to retrieve only original terms (FSN), we look for terms with typeId = '900000000000003001', since this typeId identifies descriptions that correspond to FSNs and not to synonyms or to other description types. For some terms and typeIds there is more than one description. In those cases we chose the one that was closer to the date of retrieval of SNOMED terms.

Term descriptions include the category they correspond to (for example, "(body structure)"). These terms were removed. So, if we had "liver (body structure)", we keep only "liver". Finally, SNOMED CT terms include *entire* as part of the names. The Spanish version translates *entire* as *como un todo -as a whole-*. This substring was removed from the original terms, so, for example "181081007 nervio tibial anterior [como un todo]" ("Entire deep peroneal nerve"), was transformed to "181081007 nervio tibial anterior" (deep peroneal nerve).

B.5 CRF features

This section contains the different features tried for our CRF++ implementations. The templates are written in CRF++ format.[21]

Tables B.5.1, B.5.2 and B.5.3 show excerpts of training files to understand the features used in each case.

B.5.1 Baseline features

Features used for the baseline and extracted from an example of CRF++ implementation include the current word, the current PoS tag, and the context of the current word and PoS tag (considering two tokens before, one before, the current token, one token behind, two tokens behind, one token before and the current token and the current token and one behind; for PoS tags also two PoS tags before and one PoS tag before, and one and two PoS tags behind).

```
# Unigram
#word level
U00:%x[-2,0] #two words before (columns 0)
U01:%x[-1,0] #one word before (columns 0)
U02:%x[0,0] #the current word (columns 0)
U03:%x[1,0] #one word behind (columns 0)
U04:%x[2,0] #two words behind (columns 0)
U05:%x[-1,0]/%x[0,0] #one word before AND the current word (columns 0)
```

[21] How to understand CRF++ feature templates can be seen in https://taku910.github.io/crfpp/#templ. See *Preparing feature templates*.

Implante	NC	B-AE
hepatico	AQ	I-AE
con	SP	O
alteracion	NC	B-FI
de	SP	I-FI
la	DA	I-FI
eceostructura	NC	I-FI

Table B.1: Example of file encoding for baseline features taken from CRF++.

```
U06:%x[0,0]/%x[1,0] #current word AND one word behind (columns 0)

#POS level
U10:%x[-2,1] #two POS before (columns 1)
U11:%x[-1,1] #one POS before (columns 1)
U12:%x[0,1] #the current POS (columns 1)
U13:%x[1,1] #one POS behind (columns 1)
U14:%x[2,1] #two POS behind (columns 1)
U15:%x[-2,1]/%x[-1,1] #two POS before AND one POS before (columns 1)
U16:%x[-1,1]/%x[0,1] #one POS before AND the current POS (columns 1)
U17:%x[0,1]/%x[1,1] #current POS AND one POS behind (columns 1)
U18:%x[1,1]/%x[2,1] #one POS behind AND two POS behind (columns 1)

U20:%x[-2,1]/%x[-1,1]/%x[0,1] #two POS before AND one POS before AND current POS (columns 1)
U21:%x[-1,1]/%x[0,1]/%x[1,1] #one POS before AND current POS AND one POS behind (columns 1)
U22:%x[0,1]/%x[1,1]/%x[2,1] #current POS AND one POS behind AND two POS behind (columns 1)

# Bigram
B
```

B.5.2 Our features

Implante	implante	NC	8	ante	mplante	OnlyLetter	False	Impl	B-AE
hepatico	hepatico	AQ	8	tico	epatico	OnlyLetter	False	hepa	I-AE
con	con	SP	3	con	con	OnlyLetter	False	con	O
alteracion	alteracion	NC	10	cion	eracion	OnlyLetter	False	alte	B-FI
de	de	SP	2	de	de	OnlyLetter	False	de	I-FI
la	el	DA	2	la	la	OnlyLetter	False	la	I-FI
eceostructura	eceostructura	NC	13	tura	ructura	OnlyLetter	False	eceo	I-FI

Table B.2: Example of file encoding for our feature set.

```
# Unigram

# Word lemma
```

```
U00:%x[-2,1]
U01:%x[-1,1]
U02:%x[0,1]
U03:%x[1,1]
U04:%x[2,1]
U05:%x[-1,1]/%x[0,1]
U06:%x[0,1]/%x[1,1]

# Reduced POS tags
U10:%x[-2,2]
U11:%x[-1,2]
U12:%x[0,2]
U13:%x[1,2]
U14:%x[2,2]
U15:%x[-2,2]/%x[-1,2]
U16:%x[-1,2]/%x[0,2]
U17:%x[0,2]/%x[1,2]
U18:%x[1,2]/%x[2,2]
U20:%x[-2,2]/%x[-1,2]/%x[0,2]
U21:%x[-1,2]/%x[0,2]/%x[1,2]
U22:%x[0,2]/%x[1,2]/%x[2,2]
U23:%x[-3,2]/%x[-2,2]/%x[-1,2]/%x[0,2]
U24:%x[-2,2]/%x[-1,2]/%x[0,2]/%x[1,2]
U25:%x[-1,2]/%x[0,2]/%x[1,2]/%x[2,2]
U26:%x[0,2]/%x[1,2]/%x[2,2]/%x[3,2]

# Length of the current token
U40:%x[0,3]

# 4 letter suffix
U62:%x[0,4]

# 7 letter suffix
U67:%x[0,5]

#orthographic features
%Only letters, only number letter and numbers or none of the above
U80:%x[0,6]

# whether all characters in the current token are capital letters
U90:%x[0,7]

# 4 letter prefix
U95:%x[-2,8]
U96:%x[-1,8]
U97:%x[0,8]
U98:%x[1,8]
U99:%x[2,8]

# Bigram
B
```

B.5.3 Wi-ENRE features

Implante	implante	Impl	ante	NC	8	False	True	OnlyLetter	B-AE
hepatico	hepatico	hepa	tico	AQ	8	False	False	OnlyLetter	I-AE
con	con	con	con	SP	3	False	False	OnlyLetter	O
alteracion	alteracion	alte	cion	NC	10	False	False	OnlyLetter	B-FI
de	de	de	de	SP	2	False	False	OnlyLetter	I-FI
la	la	la	la	DA	2	False	False	OnlyLetter	I-FI
eceostructura	eceostructura	eceo	tura	NC	13	False	False	OnlyLetter	I-FI

Table B.3: Example of file encoding for WI-ENRE features.

```
# Unigram
# lower case of the current token
U00:%x[0,1]

# first 4 characters of the current token
U10:%x[0,2]
# first 4 characters of two previous tokens
U11:%x[-2,2]/%x[-1,2]
# first 4 characters of two next tokens
U12:%x[1,2]/%x[2,2]

# last four characters of the current token
U20:%x[0,3]
# last four characters of two previous tokens
U21:%x[-2,3]/%x[-1,3]
# last four characters of two next tokens
U22:%x[1,3]/%x[2,3]

# POS of the current token
U30:%x[0,4]
# POS of two previous tokens
U31:%x[-2,4]/%x[-1,4]
# POS of two next tokens
U32:%x[1,4]/%x[2,4]

# length of the current token
U50:%x[0,5]

# is all CAPS
U60:%x[0,6]

# starts with caps
U70:%x[0,7]

# has only letters, only digits or letters and digits
U80:%x[0,8]
```

# Bigram
B

B.6 Spanish NegEx generic triggers

Table B.6 shows the generic trigger set for Spanish NegEx.

trigger	**translation**	label
:	:	CONJ
abajo de	below	PSEU
arriba de	above	PSEU
aunque	although	CONJ
con	with	PSEU
debido a	due to, because of	PSEU
en	in	PSEU
incluso	even	PSEU
junto a	next to	PSEU
ni	nor	CONJ
no	not	PREN
pero	but	CONJ
porque	because	CONJ
si	if	CONJ
sin	without	PREN
y	and	CONJ

Table B.4: Generic trigger set for Spanish NegEx.

APPENDIX C

Appendix C - Glossary

We provide the definition of some terms, that are useful for understanding lexical semantics and language resources that are used throughout the work. Some of the lexical semantic terms were taken from Jurafsky and Martin [137] and others are based on Wikipedia articles.

Homonymy and polysemy can lead to ambiguity. To discover the sense of a lexeme, word sense disambiguation has to be done. Homophones can lead to spelling errors and problems in speech recognition systems.

term	**definition**
antonyms	lexems that have the contrary meaning.
corpus	(plural *corpora*), a collection of texts or speech used for a specific purpose, which may be enriched with some type of annotation.
dictionary	ordered list of lexemes and their meaning (usually in function of other lexemes).
gazetteers	set of lists containing different terms, such as days of the week, cities and names of persons. Gazetteers are usually used to find occurrences of them in texts.
gold standard	dataset annotated by specialists, that can be used as a reference to evaluate software tools.
homonymy	refers to a relation that holds between words, that have the same form with unrelated meaning (for example: *bank* -financial institution and accumulation of material in the bed of a river-; in Spanish, it's translation, *banco*, also means a particular type of seat).
homophones	words with the same pronunciation and different spellings (for example, *would* and *wood* and *vaso-glass-* and *bazo -spleen-* in Spanish), cien (hundred), sien (temple -part of the head-), sun and son
homopraphs	lexemes with the same orthographic form and different meaning (for example, *bass* the musical instrument and the type of fish).
hypernyms	lexems that are a superclass of others.
hyponyms	lexems that are a subclass of others.
indexing	the act of describing or classifying a document by index terms or other symbols in order to indicate what the document is about, to summarize its content or to increase its findability.[1]
knowledge base	repository of information.
lexeme	is a minimal unit of meaning, independent of the inflectional endings that words related to it may have.
lexicon	repository of words. Entries of lexicons are lexemes.
ontology	set of concepts, their properties and relations of some subject area or domain. Relation among concepts can be vertical (is-a), of inclusion (part-of) and related-to, among others.
polysemy	refers to a lexeme with different meanings (for example: *bank* -cell bank and financial institution-). It is not always distinguished from homonomy. Etymology can help make this distinction.
synonyms	lexems that have the same meaning.
taxonomy	organization of elements in a tree-like structure.
terminology	set of terms used in a particular domain.
thesaurus	repositories of words of one particular area or domain, that include hierarchical relations and equivalence relations (synonyms, homonyms, sometimes antonyms and polysemy among terms).
word sense disambiguation (WSD)	deals with deciding which sense of a lexeme applies in a given text.

Table C.1: Glossary