eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Mapping of Electronic Health Records in Spanish to the Unified Medical Language System Metathesaurus

Author: Naiara Perez

Advisors: Montse Cuadros and German Rigau

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

September 2017

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Laburpena

Lan honetan, espainieraz idatzitako mediku-txosten elektronikoak Unified Medical Languge System Metathesaurus deituriko terminologia biomedikoarekin etiketatzeko lehen urratsak eman dira. Prototipoak Apache Lucene® erabiltzen du Metathesaurus-a indexatu eta mapatze hautagaiak sortzeko. Horrez gain, anbiguotasunak UKB bidez ebazten ditu. Ebaluazioari dagokionez, prototipoaren eta MetaMap-en arteko adostasuna neurtu da bi ingelera-gaztelania corpus paralelotan. Corpusetako bat artikulu zientifikoetako izenburu eta laburpenez osatutako dago. Beste corpusa mediku-txosten pasarte batzuez dago osatuta.

Abstract

This work presents a preliminary approach to annotate Spanish electronic health records with concepts of the Unified Medical Language System Metathesaurus. The prototype uses Apache Lucene® to index the Metathesaurus and generate mapping candidates from input text. In addition, it relies on UKB to resolve ambiguities. The tool has been evaluated by measuring its agreement with MetaMap in two English-Spanish parallel corpora, one consisting of titles and abstracts of papers in the clinical domain, and the other of real electronic health record excerpts.

Contents

List of Figures

List of Tables

1 Introduction

Medical Text Mining focuses on the application and development of biomedical text mining technologies, which are becoming a key tool for the efficient exploitation of information contained in unstructured data repositories, including scientific literature, Electronic Health Records (EHRs), patents, biobank metadata, clinical trials and social media. Thus, Natural Language Processing (NLP) is gaining increasing attention in biomedical research, as it can help unlock the information conveyed in free text and structure it in ways that can be used by computer processes to facilitate knowledge discovery, exchange, and reuse.

One such way of unearthing structure from biomedical text is generating projections or mappings from biomedical terminologies of reference to the free texts. These terminologies are very convenient, because they capture and structure in detail the knowledge that health care practitioners have about the domain.

In particular, this work pursues processing clinical documents written in Spanish looking for its specific domain terms. More specifically, it presents a preliminary application that maps EHRs in Spanish to the Unified Medical Language System (UMLS) Metathesaurus, which is a complex thesaurus resulting from joining various biomedical terminologies.

1.1 Objectives

The principal objective of this project has been to develop a preliminary functioning application that takes Spanish EHR text excerpts as input and enriches them with annotations of concepts contained in a specific biomedical thesaurus, namely, the UMLS Metathesaurus. In order to meet this objective, the following sub-goals have been identified:

- To adapt the UMLS to meet our needs
- To explore methods for recognizing biomedical terms in clinical narrative and for generating mapping candidates from the UMLS
- To explore ways of dealing with term ambiguity
- To measure the performance of our system and understand, as much as possible, what are the causes of the errors that it makes

1.2 Methodology

There exist two main approaches to develop term normalization systems, as is common in all automation problems, namely supervised and unsupervised approaches. The former requires big amounts of annotated data in order to approximate a function that, given unknown input, produces the desired output; the latter usually leverage domain knowledge, typically in the form of rules and/or knowledge bases designed by experts. At the moment, no corpus of clinical texts exists annotated with biomedical concepts —not in Spanish, nor in any other language. What do exist in vast amounts are biomedical terminology resources that capture the knowledge that experts have about the domain. Thus,

unsupervised techniques have dominated this field right until today. The preliminary pipeline we propose is built on unsupervised techniques.

In short, it first identifies and expands abbreviations and acronyms in the input text; then it creates candidate spans to be annotated; these spans are used to query an index of the UMLS Metathesaurus, whose purpose is to return candidate mappings from the Metathesaurus that are lexically similar to the given span; then, we rank the candidate mappings and choose the best mapping for each span consulted. That is, we approach term normalization as a string matching problem.

The problem of term ambiguity is addressed by mean of UKB (Agirre and Soroa, 2009), a collection of programs to perform unsupervised disambiguation based on knowledge graphs.

As for the evaluation of the system, we measure its agreement to an existing tool for biomedical term normalization in English, MetaMap, with two parallel corpora created for this evaluation, and provide an error analysis.

1.3 Contents

The remaining of this work is divided into four main sections: In Section 2, **State of the Art**, we introduce some background notions about NLP in the biomedical domain (i.e., tasks, resources, and so on) and the challenges posed by clinical narrative for NLP in general, and then provide an State of the Art of biomedical term normalization in particular. Section 3, **Description of the system**, is a detailed description of the prototype developed for this work. In Section 4, **Evaluation**, we evaluate the system by comparing it to an existing tool for biomedical term normalization in English, MetaMap, and provide an error analysis. Finally, in Section 5, **Conclusions and Future Work**, we present some concluding ideas and some actions that could be taken in the short and longer term. Additionally, this work has some appendices. Appendixes A and B contain data that complements some subjects discussed in different sections of the work. Most importantly, Appendix C, **Demonstrator**, presents an interactive interface developed to demonstrate the prototype presented in this work, and Appendix D, **UMLS webservice**, describes a webservice API of the Unified Medical Language System Metathesaurus (UMLS) that has been developed to give support to the demonstrator and facilitate consultation of the UMLS.

2 State of the Art

In this section we introduce the notions that feed this work —and that compose the title itself—: health records, mapping, and the the Unified Medical Language System (UMLS). First we give a brief introduction to NLP for health care: we review the tasks that are being studied in the research community, the community challenges that have been organized, and the terminological resources available. Next we describe one of those resources in detail, the UMLS, which is central to this work. Next, we explain what Electronic Health Records (EHR) are and the role that they play in the Spanish National Health System. We also provide an analysis of the narrative text contained in EHRs, in order to illustrate the difficulty of processing this type of texts. Finally, we introduce the task of term normalization, and some existing applications for biomedical term normalization.

2.1 NLP for health care

Natural Language Processing (NLP) in the biomedical domain has two, clearly distinct general applications. One aims at providing support to health care professionals and patients, typically by mining patient records. According to Friedman (2009), this research field started in the late 1980s, when various works demonstrated that NLP is feasible in the clinical domain, and that it can actually improve clinical care. The second started in the late 1990s, with researchers attempting to mine information from journal articles in the biomolecular domain. The term "bioNLP" is used interchangeably in the literature to refer to both trends or just the second one. This work is about NLP for health care. In what follows, we provide some information about the tasks that have been undertaken in this domain, the community challenges that have taken place related to this research field, and the resources available.

2.1.1 Clinical NLP tasks

The NLP community dedicated to the health care domain has undertaken tasks with various degrees of specificity. We can group them into four major classes:

Low-level processing is concerned with adapting existing tools or creating new ones for tokenization, part-of-speech tagging and shallow parsing to the biomedical domain.

Information Retrieval (IR) is concerned with returning from a set of known documents all the relevant documents that might be useful to a user, using as a hint the "question" or query that the user formulates. In clinical NLP, IR is useful to health practitioners because, given the vast amounts of biomedical documents that exist and are generated every day , it can facilitate finding case studies and health records that are relevant to a specific research subject or the care process of a particular patient.

Information Extraction (IE) is concerned with finding facts and events in particular documents; in other words, it is a targeted skimming of texts. In the biomedical domain, IE has been used to find very diverse information: negation, speculation, abbreviations, adverse drug reactions, temporal relations, and so on. IE is relevant to health care practitioners because it helps unlock the information conveyed in texts, structure it and quantify it. Term normalization, the subject of this work, is a sub-task of IE.

Higher-level processing involves tasks that feed from the results of IR and IE. Among the most popular in biomedical NLP are summarization, question answering (Q&A), and anonymization. Summarization is highly valued among practitioners for the same reason that IR is: the volume of data available to them. Reading shorter versions of relevant documents can simply help them save time. Q&A is mainly targeted towards patient Q&A, that is, to help patients understand their own health records. Finally, anonymization is key in biomedical NLP. Its goal is to erase or substitute the pieces of data in health records that could help identify the patient to which the health record belongs, in order to create datasets of health records usable by the research community.

2.1.2 Clinical NLP community challenges

The interest in NLP for health care has grown steadily over the years. A clear indicator is the amount of community challenges that have taken place, as shown in Figure 1.

The first challenge that involved NLP and clinical narrative took place in 2006 and was organized by i2b2[1]. There were two tasks in the challenge: one consisted in anonymizing or de-indetifying ("de-id") the content in EHRs; the second consisted in classifying patients as smokers or non-smokers based on their health records. Since 2006, i2b2 has organized 7 more challenges in a similar fashion. Some tasks organized include classifying patients as obese or as having a high risk of suffering a heart failure, sentiment analysis and co-reference resolution, and Research Domain Criteria classification ("RDoC", determining symptom severity in a domain for a patient).

In 2011, Text REtrieval Conference[2] (TREC) organized its first challenge of NLP for health care, after various others focused on the biomolecular domain. The task of the challenge was to find a population or cohort over which comparative effectiveness studies can be done by means of content-based access to the free-text fields of electronic medical records. The challenge was repeated in 2012. During years 2014 through 2017, TREC has organized challenges related to IR for precision medicine and clinical decision support.

In 2013, the first ShaRe/CLEF eHealth challenge[3] took place. The challenges involved normalizing disease terms with the UMLS Metathesaurus ("dnorm") in English clinical texts, normalizing acronyms and abbreviations, and retrieval of web pages based on queries generated when reading the clinical reports. Subsequent challenges have involved above all IE (specifically, identifying attributes that modify given annotated

[1] https://www.i2b2.org/
[2] http://trec.nist.gov/
[3] https://sites.google.com/site/shareclefehealth/

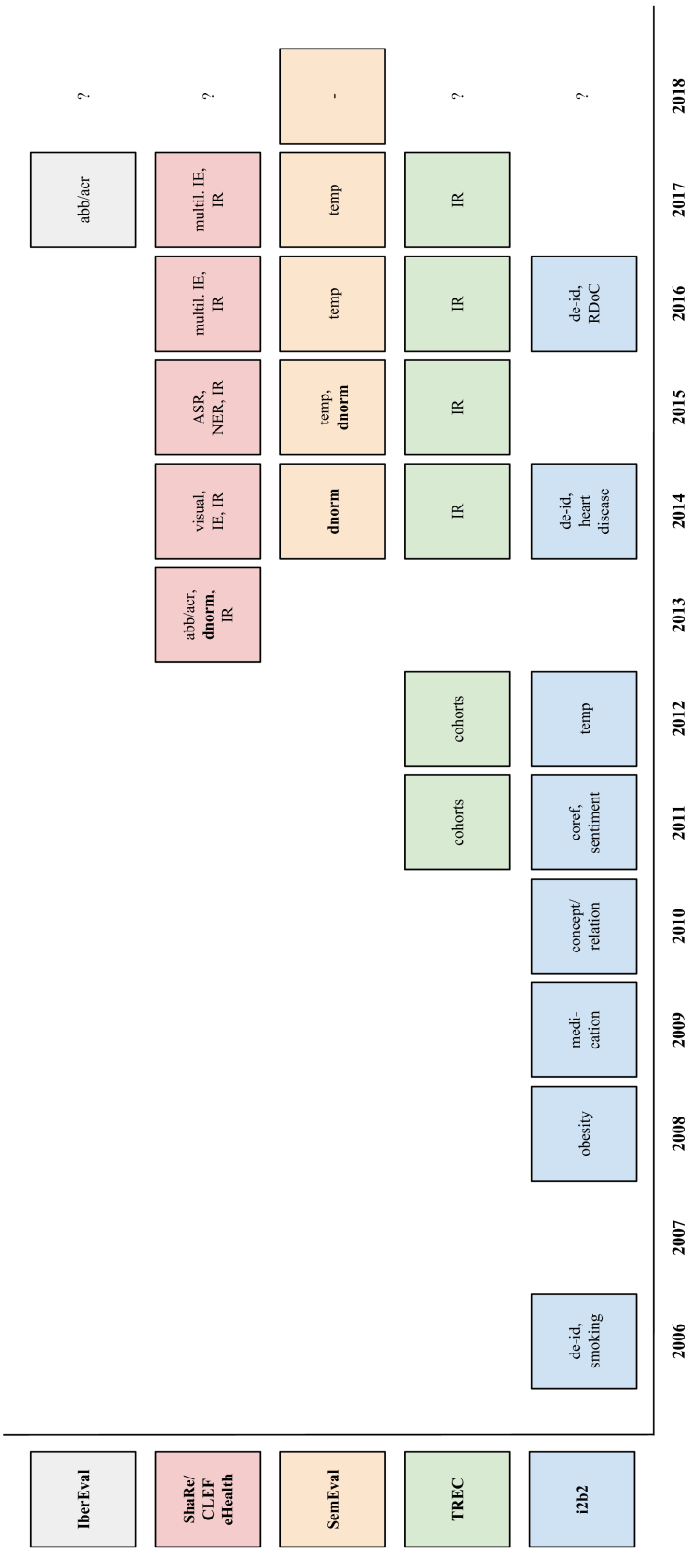


Figure 1: Clinical NLP challenges in chronological order [adapted and completed after Huang and Lu (2016)]

diseases in clinical text, such as negation, uncertainty, and so on) and user-centered health IR. In the 2016 and 2017 editions, multilingual (English/French) IE challenges have been organized. Other tasks have been about visualization of data and interactive search, and automatic speech recognition (ASR).

In 2014, SemEval included a task of disease normalization[4], following the ShaRe/-CLEF eHealth 2013 task about the same problem. The challenge was repeated in 2015, and included an additional task: extraction of temporal relations, that is, ordering in a timeline the relevant events mentioned in clinical records. This task was presented by i2b2 for the first time in 2012, but had never been proposed again. In the years 2016 and 2017, SemEval has only included the task of temporal relation extraction. In 2018, no task related to clinical NLP has been organized.

In 2017 for the first time, a challenge that involves processing clinical text in Spanish has been organized as part of IberEval[5]. The task has consisted in abbreviation and acronym recognition and normalization in clinical texts.

2.1.3 Clinical NLP resources

In this section we present the most important terminological resources that exist for the biomedical domain.

The Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT® or SCT®)[6] is the most comprehensive collection of medical terms. It was developed by the College of American Pathologists (CAP) and the United Kingdom's National Health Service, but is maintained and distributed by the Health Terminology Standards Development Organisation (IHTSDO). It has been translated to several languages, including Danish, Dutch, Spanish and Swedish . According to J. Carnicero in González and Luna (2014), it might become the standard nomenclature for EHRs in Spain —it is already in the U.S. and Australia—, given the fact that SNOMED CT®'s coverage of medical concepts has proven to be better than other controlled terminologies in various clinical domains (Chute et al., 1996; Campbell et al., 1997; Humphreys et al., 1997; Langlotz and Caldwell, 2002; Chiang et al., 2005).

The International Classification of Diseases (ICD)[7] is designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics. The reported conditions are translated into medical codes through the use of a classification structure, and selection and modification rules. The ICD was developed and is maintained by the World Health Organization (WHO). It is revised annually to incorporate changes in the medical field. The latest version of the ICD is the Tenth Revision (ICD-10), and is used by more than 100 countries around the world, including Spain. As stated in the Royal Decree 69/2015 de Sanidad, Servicios Sociales, e

[4] http://alt.qcri.org/semeval2014/task7/
[5] http://sepln2017.um.es/ibereval.html
[6] http://www.snomed.org
[7] http://www.who.int/classifications/icd/en/

Igualdad (2015), by which the Spanish Ministry for Health, Social Services and Equality regulates the Registration of Specialized Health Care Activities, the main and secondary diagnoses and medical procedures of every encounter, among other information, must be encoded with ICD-10-ES, which is an extension of the Spanish translation of ICD-10. ICD-11 is due by 2018 and will be aligned with SNOMED CT®.

The Medical Subject Headings (MeSH®)[8] thesaurus is a controlled vocabulary used for indexing, cataloging, and searching articles from 5,400 of the world's leading biomedical journals for the MEDLINE®/PubMED® database. It consists of sets of terms naming descriptors in a 13-level hierarchical structure that permits searching at various levels of specificity. MeSH is developed and maintained by the U.S. National Library of Medicine (NLM). It is continually updated by NLM health science subject specialists.

Current Procedural Terminology (CPT®)[9] is a product of the American Medical Association (AMA). CPT® codes are the U.S.'s standard to document and report medical, surgical, radiology, laboratory, anesthesiology, and evaluation and management services. They are then used by insurers to determine the amount of reimbursement that a practitioner will receive for the services provided. It is updated annually by the AMA.

International Classification of Primary Care (ICPC)[10] is an epidemiological tool used to classify data about three elements of the health care encounter: reasons for the encounter, diagnosis, and process of care. The classification is divided by body systems into 17 chapters that represent the localisation of the problem or disease. Each chapter is in turn divided into 7 components to deal with *i)* symptoms and complaints, *ii)* diagnostics, screening and preventive procedures, *iii)* medication, treatment and procedures, *iv)* test results, *v)* administration, *vi)* referrals and other reasons for encounter, and *vii)* diseases. It is complementary to the ICD, in that it pays much attention to the patient's symptoms and complaints. It was produced by the World Organization of Family Doctors (Wonca) and is maintained and updated by the International Classification Committee of Wonca (WICC).

Logical Observation Identifiers Names and Codes (LOINC® or LNC®)[11] provides formal names and standardized codes for laboratory and other clinical observations. The data cover laboratory terminology, vital signs, hemodynamics intake/output, EKG, obstetric ultrasound, cardiac echo, urological, imaging, gastroendoscopic procedures, pulmonary ventilator management, and selected survey instruments. LOINC® was developed and is maintained by the Regenstrief Institute, Inc., a non-profit medical research organization associated with Indiana University. LOINC® code translations are available from the Regenstrief Institute in Spanish, and Simplified Chinese. German,

[8] https://www.nlm.nih.gov/mesh/

[9] https://www.ama-assn.org/practice-management/cpt

[10] http://www.globalfamilydoctor.com/groups/WorkingParties/wicc.aspx

[11] https://loinc.org

French and Italian versions are available for the 3,800 most commonly used European terms. The current French translation is from Switzerland and includes short names only.

Medical Dictionary for Regulatory Activities (MedDRA)[12] is an international multilingual medical terminology that applies to all phases of drug development, excluding animal toxicology, and to the health effects and multifunction devices. MedDRA terms are arranged hierarchically in 5 levels from very specific to very general. It was developed by the International Conference on Harmonisation (ICH) and is owned by the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA).

Metathesaurus Minimal Standard Terminology Digestive Endoscopy (MTHMST) includes anatomy, findings, diagnosis, procedures, and adverse events terms related to endoscopy. It is based on information found in the Minimal Standard Terminology Digestive Endoscopy, which was developed by the European Society of Gastrointestinal Endosocpy (ESGE). The MTHMST was produced by the NLM.

WHO Adverse Drug Reaction Terminology (WHO-ART) is a 4-level hierarchical terminology that is used for coding clinical information related to adverse drug reactions. It is mainly used by drug regulatory agencies and pharmaceutical manufacturers. It was developed and is maintained by the Uppsala Monitoring Center, the WHO Collaborating Center for International Drug Monitoring. The terminology is updated every three months.

These are only a few of the terminological resources that exist. Health care centers use these terminologies to code their reports in order facilitate the management and information exchange; but they are only facilitating insofar all the users employ the same terminologies. This is easy at a health center's department level; the added value of these terminologies decreases exponentially as the stakeholders' circle is broadened to a whole health care center, to all the centers in a region or country, and to the international health care community. In an attempt to ensure interoperability between all the biomedical terminology resources, NLM developed and still maintains **the Unified Medical Language System (UMLS)**, its main component being the Metathesaurus, which brings together the terminological resources mentioned and many others. In this work, we use the UMLS Metathesaurus (2016AA version) as the knowledge base to perform term normalization. We devote the following section to describe it in depth.

2.2 The Unified Medical Language System

The Unified Medical Language System (UMLS) was created in 1986 at the U.S. National Library of Medicine (NLM) and is maintained quarterly by the same institution. It is essentially a collection of files and software. The files form a large, multi-purpose and multi-lingual knowledge database of the clinical domain that consist of various termi-

[12] `https://www.meddra.org`

nology systems of different sources put together through mapping relations. Because the entries in the vocabularies are arranged by concepts or meaning, it can be viewed as a comprehensive thesaurus or an ontology of biomedical concepts. It is intended to be used by developers of systems in medical informatics.

The main pieces of the UMLS are the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon. The SPECIALIST Lexicon[13] is a large syntactic lexicon of biomedical and general English. In what follows, we provide some information about the Metatheusarus, the vocabulary sources that it contains, and the Semantic Network.

2.2.1 Structure of the Metathesaurus

The Metathesaurus is the central and most powerful component of UMLS. It brings together several vocabulary sources of the clinical domain. These sources differ in size, language, and nature: there are thesauri, classifications, statistics, and so on. In order to standardize these sources into a single format automatically exploitable, the Metathesaurus is built by observing the following conventions:

- Each occurrence of a term in a vocabulary source is an *atom*. Each atom has an Atom Unique Identifier (AUI).
- Atoms that are lexically equivalent exemplify a *term*. Each term has a Lexical Unique Identifier (LUI).
- Terms can denote one or more *concepts*. Each atom in a vocabulary source is assigned a Concept Unique Identifier (CUI); thus, the group of atoms that have the same CUI, that is, that realize the same concept, either exemplify the same term or are strictly synonym terms.

Figure 2 illustrates the basic relation between terms, atoms and concepts. As the UMLS® Reference Manual[14] puts it:

> In the Metathesaurus, every CUI (concept) is related to at least one AUI (atom). Every AUI (atom) is linked to a single LUI (term), and a single CUI (concept). Each LUI (term) can be linked to many AUIs (atoms), and more than one CUI (concept) — although the typical case is one CUI.

An interesting property of this structure is that, theoretically, terms in the same language that express the same concept can be taken as close synonyms; furthermore, terms in different languages that express the same concept can be taken as translations of each other; finally, terms that denote more than one concept are ambiguous.

All this information is captured and distributed in a big pipe-separated file (MRCONSO.RFF), where each line represents one atom. Figure 3 shows an excerpt of the file. As can bee seen, it contains more information in addition to CUIs, LUIs and AUIs.

[13] `https://lsg3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/web/index.html`
[14] `https://www.ncbi.nlm.nih.gov/books/NBK9676/`

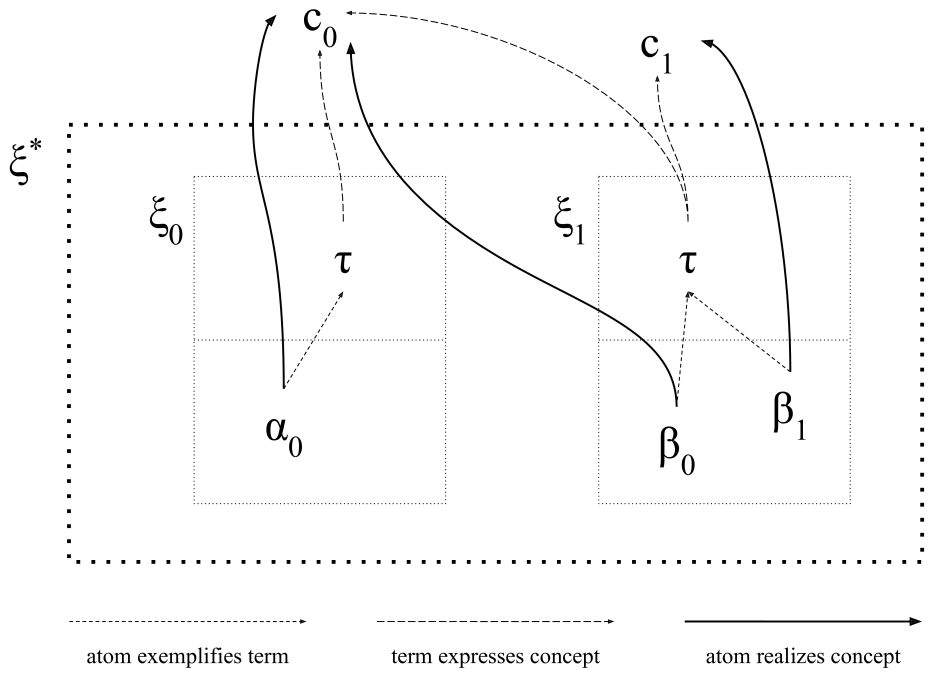


Figure 2: Basic relations among terms (τ), atoms (α, β), and concepts (c) in different vocabulary sources (ξ) [from Merrill (2009)]

Term status indicates whether the term is the preferred one ("P") or not ("S") for the CUI it is related to. *Suppressible* indicates the degree to which an entry is suppressible: "O" means the entry is obsolete; "E" or "Y" mean the entry is not obsolete but is considered suppressible; "N" means that none of the previous applies. *Source* and *language* are pretty self-explanatory.

The Metathesaurus captures *relationships* other than strict synonymy as well. In order to easily acommodate the different relationship types stated in the vocabulary sources, relationships are defined in two steps: the first assigns a broader label (see Table 1), while the second is the source asserted relationship itself as an attribute of the first. Each relation has a Relation Unique Identifier (RUI).

Information about relationships is distributed in various files. Figure 4 is an excerpt of the file MRREL.RRF, which only captures "distance-1" relationships, that is, immediate parents, children, siblings, and so on.

As the image shows, relations are not among concepts, as one could expect, but among atoms. This is because relations are asserted at the level of vocabulary sources, and a concept might be composed of atoms of different sources. In order to be able to distinguish between the information obtained from one source and another, it must be linked at the atom level. Figure 5 is an attempt at providing a visual example of the structure of the Metathesaurus as explained so far.

The upper side of the figure displays a partial representation of the concept identified as C0040185. We show 6 terms that express the concept. One of them is exemplified by

Figure 3: UMLS Metathesaurus MRCONSO.RRF excerpt

target CUI
relation type
source CUI
relation attribute
RUI
target AUI
source AUI
source

```
C0743992|A1618980|AUI|RQ|C0040185|A1619295|AUI|co-occurs_with|R00793142||CCPSS|||N||
C0749497|A1619712|AUI|RQ|C0040185|A1619295|AUI|co-occurs_with|R00793537||CCPSS|||N||
C0159870|A23284465|SCUI|PAR|C0040185|A19327870|SCUI|inverse_isa|R66735954|386040026|SCTSPA|||N|O||
C0159874|A23221483|SCUI|PAR|C0040185|A19327870|SCUI|inverse_isa|R66855040|386043029|SCTSPA|||N|O||
C1542178|A23223675|SCUI|CHD|C0040185|A19327870|SCUI|isa|R139041877|200036029|SCTSPA||0|Y|O||
C1542178|A8293161|SCUI|CHD|C0040185|A19327870|SCUI|isa|R153121983|496774602 4|SCTSPA|||Y|N||
C1997436|A19319814|SCUI|CHD|C0040185|A19327870|SCUI|isa|R96750685|321924502 8|SCTSPA|||Y|N||
```

Figure 4: UMLS Metathesaurus MRREL.RRF excerpt

Relation	Description
AQ	Allowed qualifier
CHD	has children relationship in a Metathesaurus source vocabulary
DEL	Deleted concept
PAR	has parent relationship in a Metathesaurus source vocabulary
QB	can be qualified by
RB	has a broader relationship
RL	the relationship is similar or "alike". The two concepts are similar or "alike". In the current edition of the Metathesarus, most relationships with this attribute are mapping provided by the source; hence concepts linked by this relationship may be synonymous, i.e. self-referential: CUI1 = CUI2.
RN	has a narrower relationship
RO	has relationship other than synonymous, narrower, or broader
RQ	related and possibly synonymous
RU	related, unspecified
SIB	has sibling relationship in a Metathesaurus source vocabulary
SY	source asserted synonymy
XR	not related, no mapping
	empty relationship

Table 1: Relationship types in the UMLS Metathesaurus

two atoms, which come from the same vocabulary source or terminology (MDRSPA). The bottom half of the figure shows some relations of one of the atoms that realize concept C0040185, namely, "fractura de tibia". We can see that it has three parents (or hypernyms) and one child (or hyponym). Furthermore, two of the parent atoms realize the same concept, C1542178.

Other information pieces, such as definitions, that are not part of basic structure explained, are taken to be *attributes*. They can modify concepts, atoms or relationships. There are many types of attributes and files that contain them, but they are beyond the scope of this project and will not be revised. We refer the reader to the UMLS® Reference Manual for more information: `https://www.ncbi.nlm.nih.gov/books/NBK9684/#ch02.sec2.5`.

2.2.2 Metathesaurus vocabulary sources

The UMLS Metathesaurs gathers and maps to each other more than 130 terminologies or vocabulary sources. We described the most important ones in Section 2.1.3. All of them were originally produced in English by British or American institutions, but some of them have been translated since then: the UMLS 2016AA Metathesaurus includes sources in 26 languages. Table 2 shows, for each language, how many vocabulary sources are available overall.

The Spanish subset of the UMLS 2016AA Metathesaurus contains the translations

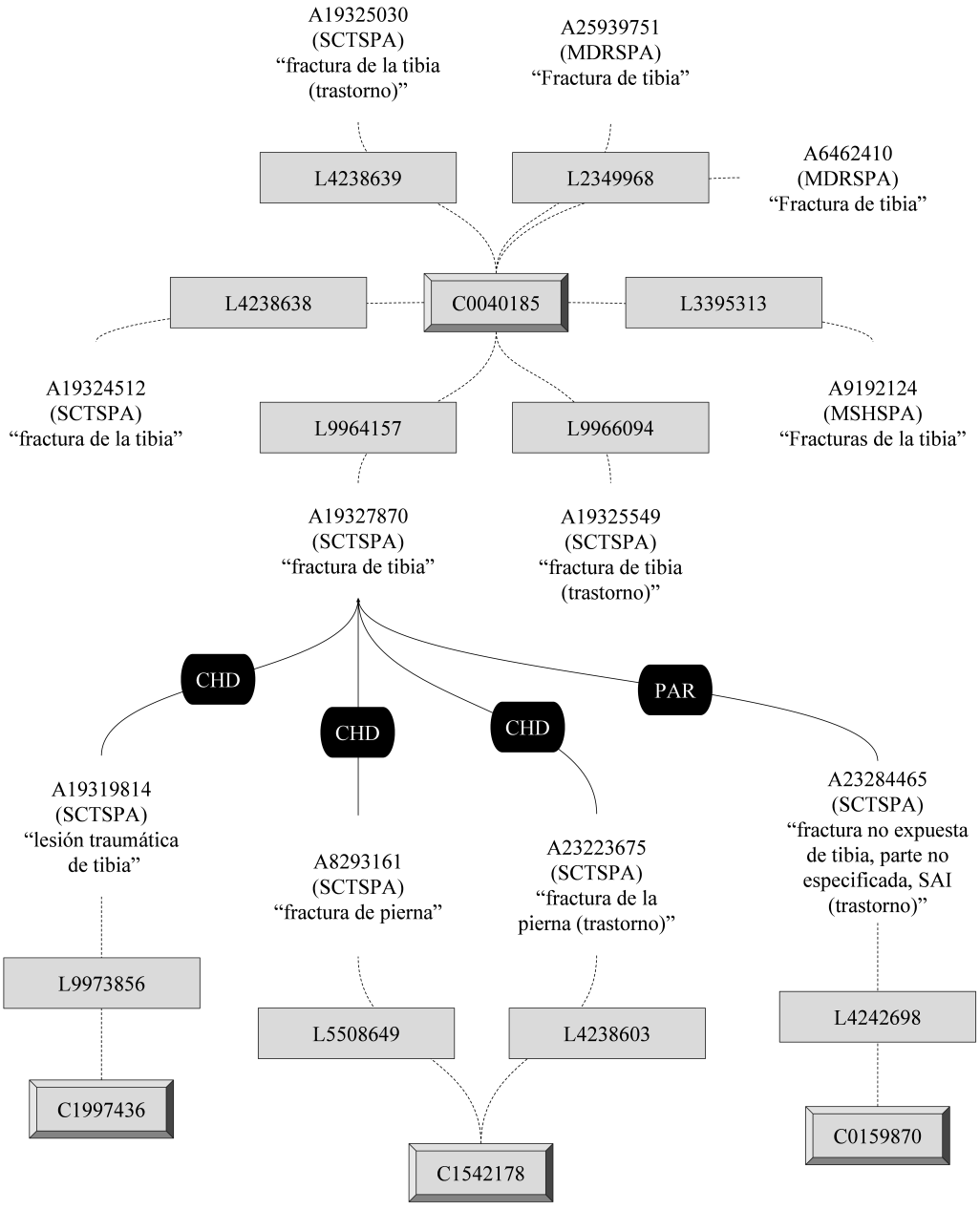


Figure 5: Partial graph of the concept “fractura de tibia” and some of its relations

of CPT®, ICPC, LOINC®, MedDRA, MeSH®, SNOMED CT®, and WHO-ART. Unfortunately, neither the Spanish version of ICD-10 nor its extension are included in the Metathesaurus of the UMLS.

Table 3 shows, for each of the source that is available in Spanish, how many concepts and terms there are in Spanish and English.

	Sources	CPT®	ICD-10	ICPC	LNC®	MedDRA	MeSH®	MTHMST	SCT®	WHO-ART
cs	2 (2)					✓	✓			
da	1 (0)			✓						
de	9 (5)		✓	✓	✓	✓	✓			✓
el	1 (1)				✓					
en	131 (76)	✓	✓	✓	✓	✓	✓	✓	✓	✓
es	9 (6)	✓		✓	✓	✓	✓		✓	✓
et	1 (1)				✓					
eu	1 (0)			✓						
fi	2 (0)		✓			✓				
fr	9 (6)			✓	✓	✓	✓	✓		✓
he	1 (0)			✓						
hr	1 (1)						✓			
hu	2 (1)		✓		✓					
it	6 (4)			✓	✓	✓	✓	✓		
ja	2 (2)					✓	✓			
ko	3 (1)				✓					
lv	1 (1)						✓			
nl	7 (2)		✓	✓	✓	✓	✓			
no	2 (1)			✓			✓			
pl	1 (1)						✓			
pt	5 (3)			✓	✓	✓	✓			✓
ru	2 (2)				✓		✓			
sv	2 (1)			✓			✓			
tr	1 (1)				✓					
zh	1 (1)				✓					

Table 2: UMLS 2016AA Metathesaurus vocabulary sources by language. The number between parenthesis indicates the amount of sources updated in the last 5 years.

	English		Spanish	
	Concepts	**Terms**	**Concepts**	**Terms**
All sources	3,250,158	9,080,352	412,831	1,042,229
CPT®	39,152	61,923	2,720	2,484
ICPC	748	1,017	722	688
LOINC®	157,645	390,425	48,609	48,631
MedDRA	51,961	78,528	45,488	61,103
MeSH®	359,116	837,305	35,970	64,804
SCT®	357,448	1,115,865	306,539	746,600
WHO-ART	3,175	3,831	2,566	3,102

Table 3: UMLS 2016AA Metathesaurus counts for English and Spanish subsets

2.2.3 The Semantic Network

The Semantic Network consist of 134 general concepts or Semantic Types (ST) and relationships or Semantic Relatinos of 54 types between the STs . The Network has two root STs: "entity" and "event"; all the other STs are descendants of one of these two. Each concept in the UMLS Metathesaurus is classified into one ST at least, always to the most specific one available (granularity varies across the Network).

2.3 Electronic Health Records

This project is about mapping UMLS concepts to narrative text in Electronic Health Records (EHR). The Spanish Association for Health Informatics, SEIS, defines *health records* as "sets of documents containing the data, assessments and information of any other nature about a patient's situation and clinical development throughout their care process. These documents might be textual or graphical, and address the patient's health and disease episodes, and the clinical activities performed as a consequence of those episodes" (SEIS, 2003, translated). Health records serve functions along five axes:

- **Care**: they are repositories where all the information about a patient and the medical acts is stored in order to guarantee the continuity of the care process.
- **Education**: they are valuable information sources for the learning of clinical cases.
- **Research**: both in clinical and epidemiological research, health records are an important source for the elaboration of analyses and retrospective studies.
- **Management**: health records serve as foundation for the billing of medical acts; they are also useful for the assessment and management of health resources and the quality of the services provided.
- **Law**: health records are proof of the medical acts and services offered.

Thus, being able to structure the information captured in the free text of a health record has potential applications at least for each one of those domains.

Health records have traditionally been stored in paper, and so they have presented several difficulties, among which SEIS (2003) mentions: disorder and lack of homogeneity, illegible information, filing errors, questionable availability and access to the information, and questionable privacy protection. For these reasons, health care agents throughout the world have devoted much effort in the last decades to computarize their health care management systems. Computerized versions of health records have many different names in the literature: "Electronic Health Records" (EHR), "Electronic Medical Records" (EMR) and "Computer-based Patient Records" (CPR) are just a few. We will henceforth use the acronym "EHR". Dick et al. (1997) describe them as follows:

> [they replace] the paper medical record as the primary record of care, meeting all clinical, legal, and administrative requirements. A [EHR] system provides reminders and alerts, linkages with knowledge sources for decision

> support, and data for outcomes research and improved management of health care delivery. [...] a [EHR] system is an evolving concept that responds to the dynamic nature of the health care environment and takes advantage of technological advances.

From the point of view of NLP research, the main advantage of computerizing health records is that accessing, manipulating and interacting with the data is much more easier. However, in practice, it has not alleviated the major bottleneck for the advancement of NLP in the medical field: the struggle in accessing real clinical texts due to privacy protection issues. This is specially worrying because NLP has steadily shifted towards techniques that require larger volumes of text. De-identification or anonymization tasks have gained popularity among researchers for this reason; even so, many institutions are still skeptical about disseminating the data for research purposes even if they are anonymized, according to Friedman (2009). As a consequence, many of the applications that have been developed rely on manually curated knowledge sources, such as the UMLS Metathesaurus and SNOMED CT®. In any case, what little research has been done with real health records is usually not comparable and can never be reproduced because the authors are not allowed to share their data.

2.3.1 EHRs in the Spanish Health Care System

Recent attention to the computarization of health records has been given by the Spanish government as a core element of the modern reformation of health care, driven above all by the necessity of providing a solution for the Spanish citizens that require medical assistance outside their autonomous community but cannot access their health data for that very reason, given that the Spanish National Health System has a decentralized structure.

In the framework defined by the project *Historia Clínica Digital* ("Digital Clinical History" —yet another name for EHRs—) of the Spanish National Health Care System[15], EHRs are composed of the following sections:

- Patient summary
- Primary Care report
- Emergency Department report
- Discharge report
- Outpatient Specialties report
- Nursery Care report
- Laboratory Tests' Result report
- Image Tests' Result report
- Other Diagnostic Tests' Result report

According to the Ministry of Health, Social Services and Equality of Spain (MSSI, 2017), 179.575.717 such electronic documents have been generated and stored up until August of 2017. However, none of these are openly available for the research community due to

[15] `https://www.msssi.gob.es/profesionales/hcdsns/home.html`

privacy issues. This is the major bottleneck for the advancement of NLP in the medical field, as has been explained.

Fortunately, we have had access to a small database of real EHRs to develop this work thanks to the collaboration in the SEMANHIS project (Gaitek 2015, IG-2015/0001027) between Clínica de la Asunción at Tolosa and Vicomtech-IK4.

2.3.2 Characteristics of clinical narrative text

Let us now focus on the research object of this project: the textual content in EHRs, that is, clinical narrative text. These texts serve diverse purposes (as illustrated by the list in the previous section), they differ in their content and level of detail. In general, they are aimed at other health care professionals or the authors themselves, so editing the texts to facilitate comprehension by a wide audience is not a main concern, as is the case other genres of texts in the same domain, such as biomedical scientific publications. Most importantly, health care professionals typically have limited time devoted to the task of writing; as a consequence, they use a myriad of abbreviations and acronyms, while hardly ever caring for spelling correctly nor respecting the grammatical standards of their language. As J. Carnicero points out in SEIS (2003), the situation has worsened since EHRs were implemented in health centers, because it is harder for practitioners having to navigate through the new complex interfaces, clicking constantly and filling forms to which they are not accustomed yet.

As a consequence, clinical narrative text is unlike general domain language in so many ways, which makes its process an extremely difficult and challenging problem for NLP researchers. Table 4 shows real and concrete examples in Spanish of these difficulties.

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

Another characteristic of clinical narrative text is atypical grammar. The most striking feature related to grammar is the amount of non-standard ellipsis found in the texts. It is also common to find unusual part-of-speech tag combinations. In this regard, it can be said that the style of clinical texts is similar to that in the titles of newspapers —extremely synthetic.

Third, despite the reductive grammar, descriptions contained in the texts are actually very rich. The same structures can be used to refer to a variety of textual subjects, such as a patient, a body part of a patient, a relative of a patient, a health care professional, a health care facility, a health care procedure, and so on. Furthermore, clinical narrative is rich because it is a product of a very specialized domain activity. As every specialized domain, health care has an ever evolving terminology.

Finally, clinical narrative is plagued with misspellings and typographical errors.

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

Table 4: Illustrative examples of common challenges in processing text from clinical narratives [adapted from Leaman et al. (2015)]

2.4 Term normalization of clinical narrative text

In this section we provide an overview of biomedical term normalization, which is the main objective of this work. Let us start with some definitions.

Terms are expressions that have a denotation in the real world. In linguistics, they are typically taken to be nouns or noun phrases. Clinical terms, then, are nouns or noun phrases that denote disorders, clinical procedures, symptoms, anatomical structures, and so on. These are the terms we presented in the UMLS Metathesaurus structure. *Term normalization* is a NLP task that consists in identifying key clinical terms mentioned in texts, and assigning them a unique entry in an ontology or controlled vocabulary. This information can then be used by other applications to provide a higher level processing, making their result highly dependent on the quality of the normalization results obtained.

There is a bit of a confusion in the literature with respect to this task. We found at least seven different ways of referring to it in the domain of NLP for health care: "term identification", "term normalization", "term mapping", "concept mapping", "concept recognition", "concept identification", and "semantic mapping". In the more general domain, it is also known as "entity linking". It has also been somewhat misleadingly called "Named Entity Recognition" (Demner-Fushman et al., 2009; Savova et al., 2010); as Funk et al. (2014) point out, there do exist sophisticated named entity recognition tools that address specific categories of terms, such as genes or diseases, but these tools require annotated training data and cannot generically be applied to recognize arbitrary terms for large, fine-grained vocabularies (Hunter and Cohen, 2006). Term normalization, on the other hand, remains a very open research problem.

2.4.1 Standard workflow for term normalization

Term normalization typically takes three steps: term recognition, term classification, and the assignment of a mapping or identity. *Term recognition*, also known as *boundary detection*, is the process by which sequences of words are recognized as clinical terms; it sets the boundaries that separate terms and non-terms. *Term classification* consists in establishing the general category of the terms recognized, that is, saying whether they are names of diseases, drugs, devices, and so on. The actual *mapping* links terms to referent data sources like ontologies or thesaurus; by doing so, it assigns to each term a unique identity. These steps can be merged and reordered (Krauthammer and Nenadic, 2004): some view term normalization as a named entity recognition and classification task plus a disambiguation step; but, if term recognition is based on a dictionary or database lookup, then the corresponding term identities can be obtained directly from the matching entries, blurring the distinction between term classification and mapping.

2.4.2 Why is biomedical term normalization difficult

The challenges explained in Section 2.3.2 and displayed in Table 4 pose huge problems for anyone attempting to apply NLP techniques to clinical text. If that were not enough, the task of term normalization poses additional problems, namely "the extensive variability of lexical term representations, term synonymy (when a concept is represented with

several terms), and term homonymy (when a term has several meanings)" (Krauthammer and Nenadic, 2004). These are the two defining characteristics of terms, as we have already mentioned when desribing the UMLS Metathesaurus: terms can exemplify more than one concept, in which case we say they are ambiguous terms —we must deal with **ambiguity**; concepts can be realized by several terms, in which case we say that those terms are synonymous —we must deal with **lexical variability**. It is often the case that what is captured in terminological resources does not match the reality found in free text. Thus, a tool that performs term normalization must be able to assign correct senses to the terms it recognizes in order to achieve the best precision possible, and it must be able to recognize terms in their varying forms in order to achieve the best recall possible.

This work addresses the problem of ambiguity but not lexical variability. In the next section, we present some approaches that have been employed to tackle ambiguity in the biomedical domain.

2.4.3 Handling term ambiguity in the biomedical domain

The task of choosing the correct sense of a word or expression in a given context is known as Word Sense Disambiguation (WSD). Approaches to WSD can be divided into supervised or semi-supervised and knowledge-based (unsupervised) methods. The former two learn statistical models from real or synthetic data in order to assign a concrete meaning to a term based on the context of its occurrence. They usually outperform knowledge-base approaches.

For example, Pustejovsky et al. (2001) used a simple word-based vector space model to disambiguate acronyms. First, they gathered abstracts that contained the ambiguous acronyms and ordered them in collections by the meaning of the acronym. Then, new abstracts with ambiguous acronyms were compared to each of these groups, by using the *tf*tdf* weighting and cosine similarity. Finally, the meaning represented by the group with higher similarity is assigned to all the occurrences of the acronym in the new abstract. This approach resulted in 97.6% accuracy.

In Pakhomov (2002), Pakhomov used a maximum-entropy classifier on the sentence level to assign the correct interpretation to an ambiguous acronym by using a context of two tokens to the left and two to the right. Precision was in average almost 90%.

Savova et al. (2010) experimented with learning a classifier for WSD using stochastic gradient decent as training algorithm and a modification of Huber's loss as the loss function. All the evaluations applying different feature sets to resolve 83 ambiguities improved the majority sense baseline.

However, in many situations —such as this project—, it is impossible to gather or create the data needed to train good models. Knowledge-based techniques, on the other hand, do not require labeled training data. Among knowledge-based techniques for WSD based on UMLS, we can mention the works by Agirre et al. (2010), McInnes (2008) and Garla and Brandt (2013). The first is specially relevant, because here we implement a little variation of their work. Let us explain their approach in some detail:

Personalized PageRank Agirre et al. (2010) use *Personalized PageRank* (Haveliwala, 2002), which builds on the notion of *PageRank* (Brin and Page, 1998). PageRank was originally conceived to rank web pages according to their relative structural importance. It uses a random walk model, where a random "surfer" starts at an arbitrary web page and, at each step, clicks at random on a hyperlink of the page or navigates to a completely unrelated web page. The PageRank score of a web page yields the probability that the random surfer is found in that page, assuming that the random walk continues indefinitely.

PageRank is formalized by modeling web pages and the links between them as a graph G that has N vertices or web pages $(v_1, \ldots, v_N)$ and, for a given vertex or webpage v_i, $B(v_i)$ is the set of vertices that point to it, that is, the set of web pages that contain a link to v_i. The PageRank, *Rank*, of vertex v_i is then defined as:

$$Rank(v_i) = \frac{1-\alpha}{N} + \alpha \sum_{v_j \in B(v_i)} \frac{Rank(v_j)}{N_j} \tag{1}$$

where α is the so-called *damping factor*, a scalar value between 0 and 1, and and N_j is the out-degree of vertex v_j.

Personalized PageRank (Haveliwala, 2002) computes the structural importance of the vertices in a graph when some vertices are more relevant than others for the task at hand. Let M be a $N \times N$ stochastic matrix corresponding to the graph G, where a matrix entry m_{ij} has the value $1/N_j$ if a link from v_j to v_i exists, and zero otherwise. Let $\boldsymbol{v}$ be a stochastic normalized $N \times 1$ vector whose elements are in $1/N$. Then, the *PageRank Vector* $\boldsymbol{PPV}$ over the graph G is given by

$$\boldsymbol{PPV} = (1-\alpha)\boldsymbol{v} + \alpha M \boldsymbol{PPV} \tag{2}$$

The key to personalizing PageRank is that $\boldsymbol{v}$ can be non-uniform and assign stronger probabilities to certain vertices.

Agirre and Soroa (2009) exploit Personalized PageRank to perform sense disambiguation of a word in a given context by using a knowledge graph: they assign greater importance to the vertices represented by the context words; then, the correct sense of the word must be that represented with the vertex that is more activated among those that represent the senses of the word. Their application is called UKB and is open source[16].

Agirre et al. (2010) showed that UKB can be used for WSD in the biomedical domain using the UMLS as knowledge graph. They annotated with MetaMap —a reference tool for biomedical term normaliation in English, introduced in the next section — a corpus of biomedical article abstracts, without having MetaMap perform disambiguation but returning annotation candidates; then, they initialized the graph for each ambiguous word with a context consisting of the CUIs in a window of 20 terms, and choosing the sense of the ambiguous words with the highest activation. In doing so, they outperformed

[16] `https://github.com/ixa-ehu/ukb`

other, more elaborate, graph-based algorithms (Navigli and Lapata, 2007; Sinha and Mihalcea, 2007; Tsatsaronis et al., 2007).

2.4.4 Existing tools and applications for biomedical term normalization

What follows is a description of some of the best -known tools for term normalization in English and their reported evaluation results, if available, and also of some works for term normalization in Spanish texts.

MetaMap (MM) Aronson (2001, 2006) is a highly configurable program that maps biomedical text to the UMLS Metathesaurus, developed by Alan R. Aronson at the U.S. National Library of Medicine (NLM). In particular, it was originally meant to improve MEDLINE® citation retrieval by enriching the articles' abstracts and titles with UMLS concept annotations. MM natively works with UMLS, but the optional data file builder[17] allows MM to use any ontology, as long as they are formatted as UMLS database tables.

MM parses input text into noun phrases and generates variants (alternate spellings, abbreviations, synonyms, inflections and derivations) from these using a "knowledge intensive" technique: it feeds from the SPECIALIST Lexicon[18], a large syntactic lexicon of biomedical and general English. A candidate set of Metathesaurus terms containing one of the variants is formed, and scores are computed on the strength of mapping from the variants to each candidate term (see Section 3.3.5 for the actual function).

As of 2006, MM performs WSD to choose among candidates that score equally well. According to Aronson (2006), "disambiguation is done by choosing the concept or concepts having the most likely semantic type for the context in which the ambiguity arises". It is not clear, however, how it does it nor what happens when the candidates have the same semantic type.

Aronson and Lang (2010) name among the strengths of MM its aggressive generation of word variants, the linguistically principled approach to the syntactic and lexical analysis, and the evaluation metric for ranking the candidates. Its weaknesses are that it only applies to English text, its speed —it has been said to be relatively slow—, and the reduced accuracy in the presence of ambiguity.

As for the performance of MM, the earliest evaluations consisted of standard information retrieval experiments. In Aronson et al. (1994), MEDLINE® documents were indexed with and without the concepts that MM found in them; retrieval of the documents indexed with the additional information improved by 4% as measured by 11-point average precision; in Aronson and Rindflesch (1997), MM was used to augment the queries as well as to aid the indexing process, improving retrieval by 14%.

In 2003, Pratt and Yetisgen-Yildiz (2003) compared the annotations of MM to those by 6 people using 60 titles of articles from MEDLINE® as dataset. They report a precision and recall of 52.8% and 27.7%, respectively, when taking exact-matches only as true positives; when allowing partial matches, recall increased by 40% and precision

[17] `https://metamap.nlm.nih.gov/DataFileBuilder.shtml`

[18] `https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/web/index.html`

by 17%. The main reason for disagreement between MM and the human annotators was the UMLS Metathesaurus' coverage; humans were not forced to restrict their annotations to concepts in the UMLS Metathesaurus.

The next year, Divita et al. (2004) performed a pilot study that evaluated the performance of MM against human annotators on two documents about genetic condition. MM found about 53% of the concepts annotated by humans. The authors identified 13 reasons for MM missing the annotations, among which the most common were: information being encoded implicitly in the text (i.e., MM does not make inferences from contextual information), underspecification (i.e., annotators assigned broad terms for specific ideas when narrower terms were not present in the UMLS Metathesaurus), definitional phrases, co-reference, and coordinating conjunctions. That is, most of the failures stemmed from MM's lack of understanding of the texts.

In 2005, Meystre and Haug (2005) evaluated MM with 160 clinical documents of diverse nature (radiology reports, exam reports, and so on). MM's results were compared to annotations by 8 physicians; the reported precision and recall for detecting a set of 80 diseases were 76% and 74%, respectively.

One must bear in mind, in any case, that both MetaMap and the UMLS Metathesaurus are in constant change, so the same experiments with more recent versions of the application and the knowledge base would most likely yield different results.

MedLEE (Friedman, 2000; Friedman et al., 1994) is one of the earliest English term mapping system for the clinical domain, alongside MM. It was created by Carol Friedman in collaboration with the Department of Biomedical Informatics at Columbia University. It was originally meant to process radiology reports only, but it was extended later to accommodate other types of clinical text as well. It works in three steps:

First, it parses the input text with a Prolog-interpretable context-free grammar and a semantic lexicon to determine its structure and reduce the stylistic variation found in natural language expressions. Then, it performs "phrase-regularization": using a mapping knowledge base automatically generated and maintained, contiguous and non-contiguous expressions are combined and standardized to the appropriate regularized forms. Finally, in the encoding phase, the standard forms are mapped "one-to-one" to the controlled vocabulary by means of a synonym knowledge base. It appears, then, that MedLEE tackles both the problems of linguistic variability and ambiguity through a careful formulation of the grammars and the knowledge bases it depends upon.

In Friedman (2000), MedLEE is evaluated by measuring its precision and recall at detecting the presence of four diseases in a collection of health records; the results were 70% recall and 87% accuracy.

NCBO Annotator is a web service provided by the National Center for Biomedical Ontology (NCBO) that annotates textual data with ontology terms from the UMLS and BioPortal ontologies. The input text is fed into a concept recognition tool, MGREP, and annotations are produced. MGREP was developed at the University of Michigan. The details of how it works are not clear (Stewart et al., 2012), publications on it being

limited to the conference poster by Dai et al. (2008). Apparently, "it is much simpler than MetaMap" (Aronson and Lang, 2010), and "it implements a novel radix-tree-based data structure that enables fast and efficient matching of text against a set of dictionary terms" (Jonquet et al., 2009). It has been compared to MM in several studies:

Shah et al. (2009) experimented with the task of large-scale indexing of online biomedical resources. MM recognized more concepts but with a lower precision than MGREP. MGREP also turned to be faster than MM. Bhatia et al. (2009) conclude as well that MGREP has a clear edge over MM for large-scale applications.

Stewart et al. (2012) use MM and MGREP to process archives of the Pediatric Pain Mailing List[19]. They observed again that MM makes more annotations (2,381 to 1,350), while MGREP has a significantly higher precision (76.1% to 58.8%).

cTakes or clinical Text Analysis and Knowledge Extraction System (Savova et al., 2010) is a comprehensive platform for performing many clinical information extraction tasks, including mapping text to the UMLS Metathesaurus. They use dictionary lookup techniques to recognize and identify clinical entities. As dictionary, they use SNOMED CT® and RxNORM[20] —a clinical drugs terminology— enriched with UMLS synonyms and a Mayo-maintained list of terms. They report that mapping to the UMLS accuracy is high for exact span matches.

GALEN Carrero et al. (2008a,b) proposed a"Spanish MetaMap" that combines machine translation techniques with the use of MM. They proposed to first process the Spanish texts with MM and a custom database that includes Spanish as well as English terms of the UMLS Metathesaurus; then, substitute the concepts found with their respective CUIs and translate this new text to English using Google Translate; finally, replace the CUIs inserted with their string representation in English, and find concepts in it with the regular MM. Unfortunately, they did not apply this system to any task, so performance scores cannot be reported.

Castro et al. (2010) in 2010, introduced a system very similar to MM but for Spanish documents with the aim of retrieving SNOMED CT® concepts based on a input phrase. The system is a component of a bigger application, the Morpho-Semantic Tagging system or MOSTAS (Iglesias et al., 2008). Term normalization is done by querying a Lucene index of SNOMED-CT and reranking the candidates with a function of their own, that is presented in Section 3.3.5. In order to evaluate the performance of this system, they obtained a set of 100 health records manually tagged by two specialists with concepts exclusively belonging to the "disruptions" or "procedures" branches in SNOMED CT®. For the exact-matching assessment, they report an average precision of 39% and a recall of 0.65%. Partial matching increases precision to 71%, but recall is still 0.75% .

[19] `http://pediatric-pain.ca/resources/pediatric-pain-mailing-list/`
[20] `https://www.nlm.nih.gov/research/umls/rxnorm/`

FreelingMed Oronoz et al. (2013) also aimed at processing clinical text in Spanish. For this purpose, they used the Freeling analyzer (Carreras et al., 2004) and extended its linguistic data with various knowledge sources: SNOMED CT®, a list of medical abbreviations (Yetano, 2003), Bot PLUS[21] —yet another clinical drugs terminology, but for Spanish— and ICD-9. Although this system does appear to assign unique identifiers to some of the terms, the actual task that the tool is meant to perform is term recognition, not term mapping. Actually, it does not perform sense disambiguation but returns the possible identities of the terms recognized. Thus, it was accordingly evaluated: a group of doctors and pharmacologists tagged manually drug names, diseases and substances in 100 health records that where split into training, developing and testing sets; the system was assessed against the human annotations, counting as true positives approximate matches of recognitions, not identities, allowed to differ by six positions to left and to the right, as is the standard approach to these evaluations; the final result was 0.90 per the F-measure.

[21] `http://www.portalfarma.com/inicio/botplus20`

3 Description of the system

The major goal of this project has been to build a preliminary system (prototype) that enriches a given text written in Spanish with annotations of concepts in the UMLS Metathesaurus, taking into account that the texts given will most likely be excerpts of medical narrative. This section is devoted to describe the system proposed: its different modules, the third-party tools and resources that it depends upon, the inputs expected and outputs generated, and other relevant details.

3.1 An overview

This section provides a general description of the prototype. Let us start with some brief technical remarks:

- the entire program has been written in Java 8
- it deploys various third-party libraries and tools, among which we must highlight:
  - Apache Lucene™ [22]
  - ixa-pipes (Agerri et al., 2016), a set of tools for NLP
  - UKB (Agirre and Soroa, 2009), a collection of programs to perform unsupervised word sense disambiguation based on a given knowledge base
- it expects plain text or XML files as input, and produces as output NAF (Fokkens et al., 2014) or JSON files that contain, among other data, term normalization annotations
- it can be run in several ways: as a TCP service or as a single-run process, reading input and writing output from/to files or the standard input/output; it can also be run as a web service

The system proposed consist of components executed in sequence to process the clinical narrative. The overall preliminary architecture for the prototype is schematized in Figure 6. Before explaining how each module works separately in the subsequent sections, let us define broadly how they interact, so that the system is able to receive plain text and output normalized terms.

3.1.1 A tentative workflow

Let us describe the proposed processing flow by means of an example; take the input text to be the following:

"lesión grave en rodilla dcha"

[22] `https://lucene.apache.org/core/`

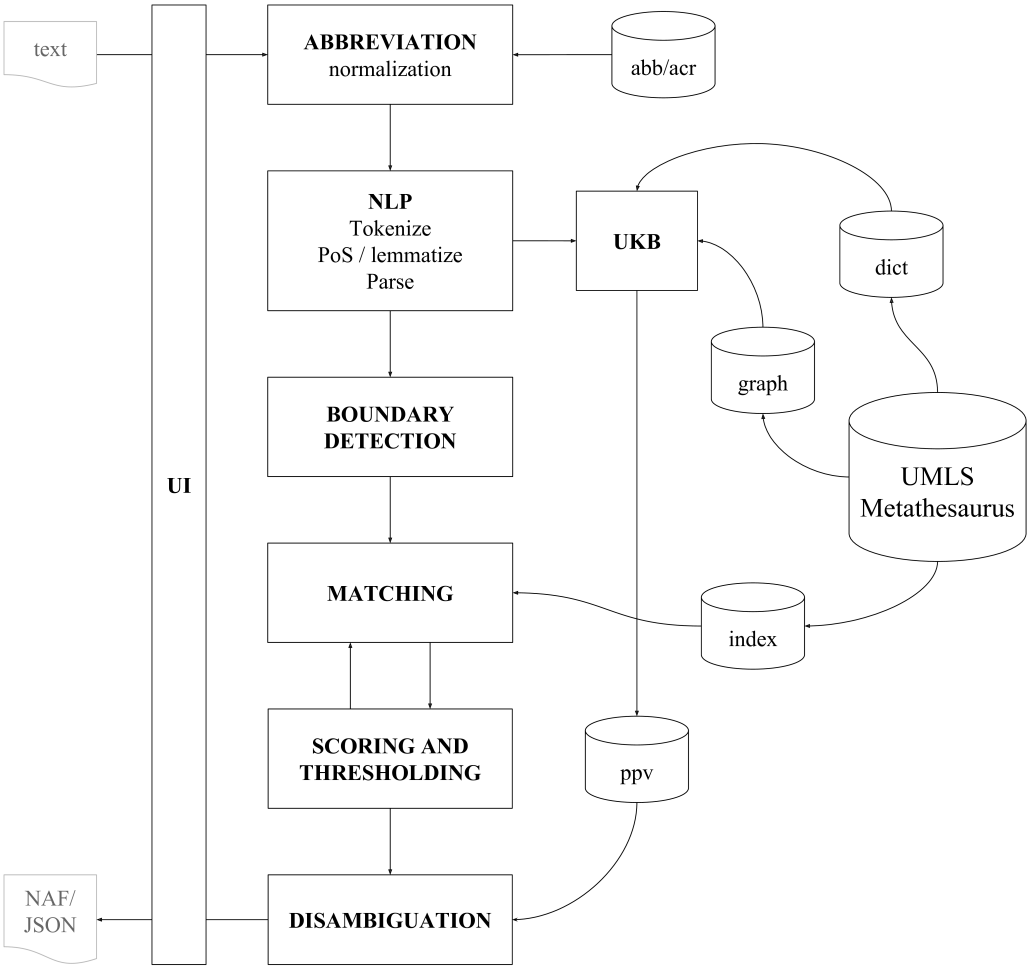


Figure 6: System architecture schema

First, the text received is analyzed in search of **abbreviations and acronyms**, which are expanded to their corresponding full expressions. This is the only step that aims at normalizing the text; we do not yet deal with misspellings, abnormal capitalization nor punctuation, and so. In our example, this step would produce

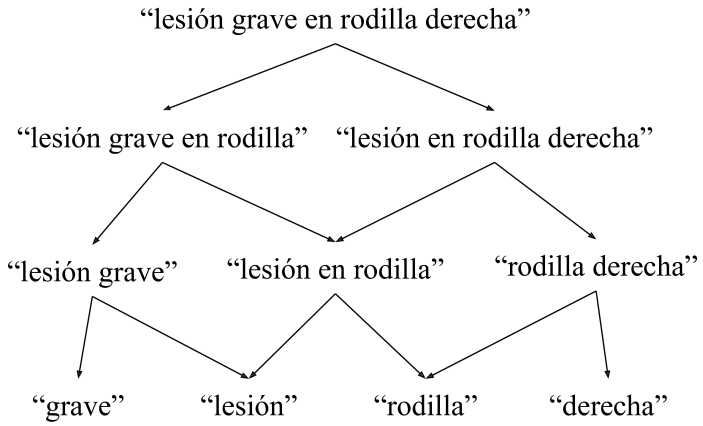
"lesión grave en rodilla derecha"

Next, the system carries out low-level **linguistic processing** of the expanded text: tokenization, part-of-speech tagging and, depending on the configuration chosen, constituent parsing. It uses ixa-pipes tools (Agerri et al., 2016) to do so.

The linguistic information obtained serves then as basis to perform **boundary detection**, that is, to recognize in the text spans or sequences of tokens that are likely to be mapped to a medical concept. We explore two ways of doing so: by means of simply extracting ngrams, or detecting nominal phrases using rules. For the sake of clarity, we will henceforth refer to the nominal phrases and ngrams simply as "spans".

After identifying spans, the system attempts to find mapping candidates with the UMLS Metathesaurus terms by lexical proximity —just not for all of the spans: the **mapping candidate generation algorithm** has been designed so that longer spans are prioritized and, in case of not finding satisfactory candidates, smaller spans are tested. To be more specific,

1. the system first orders the spans by subsumption creating oriented trees:



2. Then, it generates candidate mappings for each root of the trees and its direct children. Following the example, it would generate candidate mappings for “lesión grave en rodilla derecha”, “lesión grave en rodilla”, and “lesión en rodilla derecha”.

3. If any of the children get a better mapping than their parent, then the candidates found for the parent span are ruled out, and the algorithm is repeated recurrently for the children nodes. That is, if either “lesión grave en rodilla” or “lesión en rodilla derecha” has a mapping better than those for “lesión grave en rodilla derecha”, then the system would discard the candidates for the latter and search new ones for “lesión grave”, “lesión en rodilla”, and “rodilla derecha”.

4. If a parent has a mapping better than any of its children’s, the mappings found for the parent are accepted as candidates and the system does not attempt to map any of its descendants. Then, if “lesión grave en rodilla” had a better mapping than those of its children, namely “lesion grave” and “lesión en rodilla”, the system would discard the candidates of the latter two and would not attempt to map their respective children “grave”, “lesión” nor “rodilla”.

Following this algorithm, spans that overlap can be annotated with different concepts, but not spans that are nested within a bigger one. But, how does the system generate candidate mappings? What makes one candidate better than another?

Candidate generation is performed by the **matching** module. We have created an Apache Lucene™ index of the 2016AA version of UMLS Metathesaurus that the system queries with spans; the result of a query is a collection of candidate Metahesaurus terms,

which are in turn related to one or more CUIs and a relevance score provided by Lucene: the higher this score, the more relevant the result is with respect to the query, thus a better candidate according to Lucene.

The **scoring** module re-ranks the candidates retrieved, that is, it assigns new scores to the candidates using a function other than that of Lucene's. We explore two such functions: the one by Castro et al. (2010), and the one by Aronson (2001) implemented in MetaMap. Furthermore, a **threshold** can be applied to discard candidates with low scores.

At this point, a span can have zero, one or multiple mapping candidates. Then,

a) if no candidate is available, one must conclude that either the span in question was never a term in the first place, or that it is a term but does not have an explicit or convincing enough mapping to the UMLS Metathesaurus.

b) if one candidate is available, the system takes its CUI as a mapping for the span.

c) if more than one is available, the system takes the CUI of the one scored highest.

d) a trickier situation is when more than one candidate become tied in first position; the system needs to carry out a **disambiguation** process in order to choose the correct mapping. It resorts to UKB for this purpose. For comparison purposes, the system can also perform disambiguation randomly, by choosing the first of the candidate list, or do not perform it at all, that is, do not assign any mapping to ambiguous spans.

Finally, the mappings are gathered in the output file and displayed to the user.

3.1.2 Parametrization

There are several settings of the application that users can control through a PROPERTIES file or through call parameters when run as a web service:

- whether to detect and normalize abbreviations,
- whether to consider ngrams or nominal phrases as spans,
- the maximum length (in tokens) of the spans,
- the maximum amount of Lucene results considered per span,
- whether to perform re-ranking of candidates (and with which function),
- the threshold that invalidates candidates,
- whether to perform concept disambiguation (and how), and
- the format of output files.

The following table shows the accepted values:

Param	Description	Values
`abbr`	Enables abbreviation normalization	`true`, `false`
`bound`	Determines which boundary detection technique is used	`ngram`, `phrase`
`length`	Specifies the maximum length (in tokens) of the ngrams or phrases	int > `0`
`max`	Specifies the maximum amount of Lucene results considered per query	int > `0`
`score`	Determines which reranking method is used — if none is chosen, the system uses the score given by Lucene	`aronson`, `castro`, `perez`, `-`
`thresh`	Specifies the minimun weight a candidate must have to be accepted	float ≥ `.0`
`disamb`	Determines which disambiguation method is used — if none is chosen, the system simply skips ambiguous ngrams or phrases	`ukb`, `rand`, `first`, `-`
`format`	Determines the format of the output files	`naf`, `json`

Table 5: System parameters and possible values

3.1.3 The output: NAF and JSON files

Abbreviations and UMLS annotations are encoded in NAF files as external references of `term` elements. Figure 7 shows the NAF output produced when parsing "lesión grave de rodilla derecha".

Abbreviations contain their expanded meaning as the value of the `reference` attribute. UMLS annotations have the CUI as the value of the `reference` attribute and the UMLS source of the annotation as the value of the `source` attribute. When annotations are made of terms that do not exist in the `terms` layer after part-of-speech tagging, a new term is inserted.

The output encoded in JSON files contain less information than NAF files do; it was implemented mainly to facilitate creating the demonstration webpage (Appendix C).

3.2 Resources

In this section, the reader will learn about the three main resources that the application depends upon: the Lucene index of the UMLS Metathesaurus, and the dictionary and graph for UKB disambiguation.

3.2.1 The UMLS index

We use Apache Lucene™ in order to be able to make fast searches in the UMLS Metathesaurus. An index has been created of the 2016AA version of the Metathesaurus, where

```
<terms>
	<!--lesión-->
	<term id="t1" type="open" lemma="lesión" pos="N" morphofeat="NCFS000">
		<span>
			<target id="w1" />
		</span>
	</term>
	<!--grave-->
	<term id="t2" type="open" lemma="grave" pos="G" morphofeat="AQ0CS0">
		<span>
			<target id="w2" />
		</span>
	</term>
	<!--en-->
	<term id="t3" type="close" lemma="en" pos="P" morphofeat="SPS00">
		<span>
			<target id="w3" />
		</span>
	</term>
	<!--rodilla-->
	<term id="t4" type="open" lemma="rodilla" pos="N" morphofeat="NCFS000">
		<span>
			<target id="w4" />
		</span>
	</term>
	<!--dcha-->
	<term id="t5" type="open" lemma="dcho" pos="G" morphofeat="AQ0FS0">
		<span>
			<target id="w5" />
		</span>
		<externalReferences>
			<externalRef resource="Yetano.2003" reference="derecha" />
		</externalReferences>
	</term>
	<!--lesión en rodilla-->
	<term id="t6" lemma="lesión_en_rodilla">
		<span>
			<target id="w1" />
			<target id="w3" />
			<target id="w4" />
		</span>
		<externalReferences>
			<externalRef resource="UMLS-2016AA" reference="C0022744" source="MDRSPA" />
		</externalReferences>
	</term>
	<!--rodilla dcha-->
	<term id="t7" lemma="rodilla_derecha">
		<span>
			<target id="w4" />
			<target id="w5" />
		</span>
		<externalReferences>
			<externalRef resource="UMLS-2016AA" reference="C0230431" source="SCTSPA" />
		</externalReferences>
	</term>
</terms>
```

Figure 7: Example of enriched `terms` layer in a NAF file

```
{
	"terms": [
		{"id": "t1", "wf": "lesión", "lemma": "lesión", "pos": "N", "morphofeat": "NCFS000",
			"offset": 0, "length": 6, "sent": 1, "para": 1},
		{"id": "t2", "wf": "grave", "lemma": "grave", "pos": "G", "morphofeat": "AQ0CS0",
			"offset": 7, "length": 5, "sent": 1, "para": 1},
		{"id": "t3", "wf": "en", "lemma": "en", "pos": "P", "morphofeat": "SPS00",
			"offset": 13, "length": 2, "sent": 1, "para": 1},
		{"id": "t4", "wf": "rodilla", "lemma": "rodilla", "pos": "N", "morphofeat": "NCFS000",
			"offset": 16, "length": 7, "sent": 1, "para": 1},
		{"id": "t5", "wf": "dcha", "lemma": "dcho", "pos": "G", "morphofeat": "AQ0FS0",
			"offset": 24, "length": 4, "sent": 1, "para": 1}
	],
	"bioConcepts": [
		{
			"id": "bc1",
			"ci": "C0022744",
			"source": "MDRSPA",
			"references": [[ "t1", "t3", "t4" ]]
		},
		{
			"id": "bc2",
			"ci": "C0230431",
			"source": "SCTSPA",
			"references": [[ "t4", "t5" ]]
		}
	],
	"abbreviations": [
		{
			"id": "abb1",
			"meanings": ["derecha"],
			"references": [["t5"]]
		}
	]
}
```


Figure 8: Example of output in JSON format

each entry represents a term of the Metathesaurus and contains the following information: the term itself, the normalized term, its CUI and its source.

The normalized string is obtained after erasing spurious parenthetical content, punctuation, and stopwords. The list of the spurious parenthetical content has been added as an appendix (A); it has been curated manually after studying the Metathesaurus and can certainly be optimized. As for the stopwords, we consider 303 common Spanish words except "no", "sin" and "con" (*no*, *without*, and *with*, respectively) because they alter the polarity of expressions, which is important to capture in the medical domain.

Not all the terms and concepts in the Metathesaurus have been indexed, only those that

1. are in Spanish,
2. do not belong to LOINC®,
3. are shorter than 15 tokens,
4. are not obsolete or suppressible according to the Metahesaurus,
5. do not consist of a single character,
6. do not consist of just numbers, and
7. do not consist of only stopwords.

LOINC® terms look typically like "especie de Thrichomonas:número areico:punto en el tiempo:sedimento urinario:cuantitativo:microscopia.de luz.campo de gran aumento", so they are not suited for the task at hand.

The original complete Metathesaurus (version 2016AA) has 3,250,226 concepts and 10,586,865 terms. The subset in Spanish consist of 412,831 concepts and 982,565 terms. After applying the filters listed, we are left with 352,075 concepts and 546,309 terms. Let us analyze the amount of ambiguity (or homonymy) and lexical variability (or synonymy) captured in our index.

Terms can occur more than once in the index. Each occurrence represents a distinct sense of the term: it relates the term to a different CUI. That is, terms that occur more than once are ambiguous. The index contains such 2,147 terms (that is, 99.6% of the terms are unambiguous):

# of senses	**# of terms**
1	544,162
2	2,041
3	75
4	14
5	4
6	3
7	3
8	2
9	1
10	2
15	1
16	1
total:	546,309

Table 6: Sense counts of terms indexed

As can be seen, most ambiguous terms have 2 senses, but there are also terms related to 10 or more different CUIs. These are the most ambiguous terms:

# of senses	**term**
16	limitación funcion/minusvalía
15	con injerto autólogo (incluye obtención del injerto)
10	con manipulación
10	con anestesia
9	con aloinjerto
8	diámetro de la lesión de 1.1 a 2.0 cm
8	diámetro de la lesión de 0.6 a 1.0 cm

Table 7: The 7 most ambiguous terms in the UMLS index

None of these terms looks ambiguous in the traditional sense; most of them actually seem to express highly accurate ideas. A closer look reveals that the source of all these terms is the Spanish translation of the Current Procedural Terminology (CPT®), and that the English counterparts are not ambiguous at all. Let us illustrate this with an example; below are 3 of the 9 senses for the term "con aloinjerto" (*with allograft*) – we give the CUI that represents each sense, and some terms in English that exemplify them:

"con aloninjerto"	
CUI	**English terms**
C0370857	"Excision or curettage of bone cyst or benign tumor of clavicle or scapula; *with allograft*", "Removal of bone lesion"
C0370860	"Excision or curettage of bone cyst or benign tumor of proximal humerus; *with allograft*", "Removal of humerus lesion"
C0370901	'Excision or curettage of bone cyst or benign tumor, humerus; *with allograft*", "Remove/graft bone lesion"

As can be seen, these concepts do have different referents in fact, but for some reason that we have not been able to clarify, in the Spanish translation all of them have been represented with a single term that only captures what all them have in common: "with allograft". In theory, terms in different languages that realize the same concept in the Metathesaurus should be translations of each other; this example shows that it is not always the case. It also shows that much of the ambiguity that has to be dealt with when working with the UMLS has been introduced artificially: ambiguity in natural language is usually easy to resolve by humans, we do it all the time without even realizing in the majority of cases; but a situation where someone infers from "with allograft" that the speaker is referring to the therapeutic procedures excision or curettage, specifically of bone cyst or a benign tumor, and in the humerus as opposed to the clavicle or scapula is hard to imagine.

This is not to say that all the ambiguities in the index are of the type just explained. Most of the ambiguous terms are assigned a few senses, typically not more than two or three, and are more of the kind that one would expect when dealing with ambiguity in natural language. Here are a few examples:

"recto"	
CUI	**English terms**
C0034896	"rectum"
C0445291	"straight"
C0370860	"belching", "burping", "eructation"

"abultamiento"	
CUI	**English terms**
C0038999	"Part of body puffy", "Swelling"
C0370860	"Abdomen feels bloated", "TYMPANITES"

"boca"	
CUI	**English terms**
C0230028	"Mouth region", "Oral part of face"
C0370860	"Bucal cavity", "Cavitas orias"

Looking at the normalized versions of the terms, the count of distinct terms goes down to 538,026 but ambiguous terms increase up to 9,973, that is, 2% of the normalized terms are ambiguous (Table 8). This occurs because removing spurious parenthetical content and stop words from the terms can erase the differences between terms that were previously distinct.

# of senses	**# of terms**
1	528,077
2	9,563
3	320
4	47
5	13
6	9
7	8
8	4
9	2
10	3
11	2
16	2
	Total: 538,026

Table 8: Sense counts of normalized terms indexed

Regarding lexical variability, the term-to-concept ratio in the index is 1.55, that is, for every concept indexed we have 1.55 terms that refer to it. The actual distribution of terms per concept is shown in Table 9. Again, most of the concepts have just one term associated (66.6% of the concepts) or only a few, but there are concepts realized by more than 20 different terms, up to a maximum of 59. Just out of curiosity, the concept with 59 terms is C0028470, which is defined as "Agents capable of exerting a harmful effect on the body"; some of the terms indexed for this concept are "Agente Biológico Nocivo", "Agente Biológico Perjudicial para la Salud", "Agente Etiológico Físico", "Agente Físico

Nocivo", "Agente Biológico Nocivo", "Agentes Patógenos Biológicos", and so on. That is, they result from the combination of four or five phrases in their singular and plural form. As such, we can say that this concepts is quite well covered for the possible ways in which it can appear in narrative text. Unfortunately, we have seen that this is a rare case in our knowledge base.

3.2.2 UKB graph and dictionary

UKB is a collection of programs to perform unsupervised word sense disambiguation based on a given knowledge base (KB) in the form of a graph, where the vertices are concepts and edges are relations between the concepts. Additionally, UKB needs a dictionary that associates terms to one or more concept of the KB.

The KB for this project contains all the relations in the 2016AA Metathesaurus that have as origin and target concepts included in our UMLS index. For each relation, we indicate the source CUI, target CUI, the direction of the relation, and its type, but do not assign any weight at the moment to relations. Overall, the graph consists of 352,075 vertices and 8,381,482 edges. That is, all the concepts indexed participate in one relation at least. As for the dictionary, it simply maps each term in the UMLS index to their CUI or CUIs, in the case of those that are ambiguous.

3.3 Modules

This section describes the separate modules that constitute the system proposed.

3.3.1 Abbreviation and acronym expansion

The tool employed to identify abbreviation- or acronym-like elements in texts was developed by Montoya (forthcoming) as part of his Master's Thesis. We will not elaborate on the details of this tool since it is not the focus of this project, but refer the reader to Montoya's work for more information; let it suffice to say here that it exploits a set of rules and a 2,312-item long list of abbreviation/acronym and corresponding expansions, curated after manual annotations by health care professionals. The abbreviations and acronyms in the list are taken to be unambiguous, that is, each has a unique expansion and is substituted directly with that expansion whenever detected in a text.

3.3.2 NLP pipeline

The NLP module performs tokenization, part-of-speech tagging and constituent parsing. It uses ixa-pipes tools (Agerri et al., 2016) to do so.

A set of 1,080 sentences of clinical narrative have been annotated with ixa-pipes-pos and manually corrected in order to assess its performance. Table 10 shows the results. As can be seen, accuracy drops by more than 15% as compared to the results obtained when tested on a subset of the Ancora Spanish 3.0[23] corpus. It has been observed that

[23] `http://clic.ub.edu/corpus/en`

# of terms	# of concepts
1	243,459
2	68,865
3	21,223
4	8,401
5	3,821
6	2,095
7	1,540
8	900
9	533
10	352
11	255
12	180
13	128
14	84
15	53
16	45
17	29
18	29
19	20
20	16
21	12
22	8
23	6
24	6
25	1
26	2
27	1
28	3
31	1
32	1
33	2
34	1
35	1
36	1
59	1
	Total: 352,075

Table 9: Term counts per concept indexed

the tagger has problems specially with uppercase expressions and sentences that start with noun phrases without articles.

	Ancora	**Clinical**
# of sentences	3,383	1,080
# of tokens	101,385	10,821
# of tags	256	148
accuracy	95.82%	80.29%

Table 10: Evaluation of the ixa-pipes part-of-speech tagger in clinical text

3.3.3 Boundary detection

The purpose of this module is to extract from the text spans or sequences of tokens that potentially can be mapped to a medical concept. It expects as input the linguistic information obtained about the text from the NLP module; it returns token sequences or "spans" as output. In this work we explore two ways of doing this: the first simply consists in extracting ngrams of different sizes; the second is a rule-based algorithm to extract nominal phrases. The motivation for implementing such an algorithm is to maximize recall by means of extracting discontinuous spans as well. That is, its aim is to extract, for example, the spans in Figure 9 (repeated here for convenience) when given the text "acude por lesión grave en rodilla derecha". It does so by traversing the constituent trees received.

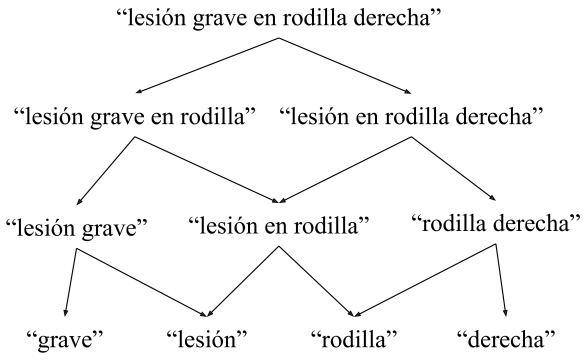


Figure 9: Example of a span tree

3.3.4 Matching

The aim of this module is to retrieve for a given span relevant results from the UMLS index presented in Section 3.2.1, that is, to generate candidate mappings. A relevant result is one which contains a normalized term that is lexically similar to the span without stopwords. A query can return zero, one or more results. Each of these results

retrieved contains the following information: a CUI, the term, and the source in the UMLS Metathesaurus. Additionally, Lucene assigns a score to each result; the higher the score, the more relevant the result is supposed to be.

For instance, given the span "lesión grave", the results obtained would be:

score	CUI	term
11.04	C1282312	lesión craneoencegálica grave
6.50	C0588018	prevención de una lesión permanente grave de la salud física/mental de la embarazada

In the case of "lesión de rodilla", the top 15 results would be these:

score	CUI	term
11.92	C0022744	Lesión de rodilla
10.00	C0022744	lesión traumática de la rodilla
10.00	C0160991	lesión traumática por aplastamiento de la rodilla
10.00	C0160991	Lesión por aplastamiento de la rodilla
10.00	C0187904	resección de lesión en articulación de rodilla
10.00	C0347548	lesión superficial de la rodilla
10.00	C0410093	lesión de ligamentos de la región de la rodilla
10.00	C0410095	lesión de la cápsula de la región de la rodilla
10.00	C0433113	lesión cerrada por aplastamiento, rodilla
10.00	C0433147	lesión por desollamiento de rodilla
10.00	C0451979	lesión de varias estructuras de la rodilla
10.00	C0877583	lesión traumática de ligamento de la rodilla
8.79	C0160989	lesión traumática por aplastamiento, rodilla y pierna
8.79	C0188300	destrucción local de lesión de articulación de la rodilla
8.79	C0410084	lesión de partes blandas de la región de la rodilla

If the span were "fin de semana", the result obtained would be:

score	CUI	term
17.33	C0556334	bebedor de fin de semana
17.33	C0581045	visita durante el fin de semana
10.20	C0269662	hiperemesis gravídica antes del fin de la semana 22 de gestación con depleción de carbohidratos
10.20	C0269663	hiperemesis gravídica antes del fin de la semana 22 de gestación con deshidratación
10.20	C0269664	hiperemesis gravídica antes del fin de la semana 22 de gestación con la desequilibrio electrolítico

Notice how Lucene assings a much higher score to "bebedor de fin de semana" when queried with "fin de semana" than to "Lesión de rodilla" when queried with "lesión de rodilla". Lucene's score does *not* measure the lexical similarity between the indexed entries and the query; it *does* measure the relevance of an indexed entry with respect to the query and in contrast to the rest of the entries in the index.

3.3.5 Candidate scoring

This module overrides the scores given by Lucene to a collection of candidate mappings for a span. It also applies a threshold given by the user in order to discard candidates with scores lower than desired. As a result, three scenarios are possible: that none of the candidates passes the filter, that only one passes the filter, or that more than one pass it.

In this work we implement two scoring functions of varying complexity. The first one is used in Castro et al. (2010); we will henceforth refer to this function as $score_{Castro}$ or simply *Castro*. The second one is by Aronson (2001), the one implemented in MetaMap; we will henceforth refer to this function as $score_{Aronson}$ or *Aronson*.

Before describing each function, let us first define some notation that will come handy, since it is used in all of them:

q is the text span used to query the index
r is the Metathesaurus term retrieved from the index
γ is the amount of tokens that match between q and r
w/s means "without stopwords"

If q were "clear cell" and r were "clear cell cystadenocarcinoma of the ovary", γ would be 2 because two tokens participate in both terms: "clear" and "cell". $r_{w/s}$ would typically look like "clear cell cystadenocarcinoma ovary".

Castro. The function is based on that proposed by Patrick et al. (2007). In their study, the authors retrieved concepts directly from SNOMED-CT and proposed a formula where the score was equal to the number of tokens used in all matches divided by the number of tokens in the total input stream. Castro et al. where concerned that this formula did not take into account the length of the retrieved string, so they propose this:

$$score_{Castro} = \frac{\gamma^2}{length(q_{w/s}) \times length(r_{w/s})} \tag{3}$$

The result always ranges from 0.0 to 1.0, 1.0 being the best score possible. In our previous example, given that $\gamma = 2$, $length(q_{w/s}) = 2$ and $length(r_{w/s}) = 4$, then $score_{Castro} = 2^2/(2 \times 4) = 0.5$.

Aronson. This function tries to encode more information in the result. It is a weighted average of four measures:

$$score_{Aronson} = \frac{centrality + variation + coverage \times 2 + involvement \times 2}{6} \tag{4}$$

where

$$centrality = \begin{cases} 1, & \text{if } r \text{ involves the syntactic head of } q \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$variation = \frac{4}{distance + 4} \tag{6}$$

$$coverage = \left(\frac{2}{3} \times \frac{span(r)}{length(r)}\right) + \left(\frac{1}{3} \times \frac{span(q)}{length(q)}\right) \tag{7}$$

$$involvement = \frac{\gamma}{length(q_{w/s})} \tag{8}$$

The *distance* of r is the sum of the distance values for each step taken during variant generation; the values for each step are shown in Table 11. *variation* does not contribute much in our system, because we do not generate variants, and so *distance* always has value 0.

Variant type	Distance value
spelling	0
inflectional	1
synonym or acronym/abbreviation	2
derivational	3

Table 11: Variant distances of MetaMap's scoring function (Aronson, 2001)

The *span* of either r or q is the number of words participating in the match, ignoring gaps; that is, the number of tokens from the first word participating to the last word participating, both included.

The result always ranges from 0.0 to 1.0, 1.0 being the best score possible. Let us calculate this score for the example where q = "clear cell" and q = "clear cell cystadenocarcinoma of the ovary": $centrality = 1$ because the head of q is "cell" and q contains "cell"; $variation = 4/(0+4) = 1$; $coverage = (2/3 \times 2/2) + (1/3 \times 2/7) = 0.76$; $involvemenet = 2/2 = 1$. Thus, $score_{Aronson} = (1 + 1 + 0.76 \times 2 + 1 \times 2)/6 = 0.92$.

3.3.6 Disambiguation

This module is only invoked when a span has more than one winning mapping candidate with the same score. Note that not only ambiguous terms provoke this situations — which they do, inevitably—; because of the scoring functions explained in the previous section, different terms can also receive the same score. That is, two sources of ambiguity come into play: the first is given by the Metathesaurus, when it assings different CUIs to the same term. The second is produced during runtime and depends on the scoring function used: it is possible that two different terms receive the same score. All the same, the system needs to choose only one mapping. It does so using UKB.

The algorithm behind UKB is Personalized PageRank, which has been explained before in Section 2.4.3. A possible application would be, as in Agirre et al. (2010), to first map all the non-ambiguous terms in the text and then use those as context to assign a CUI to the ambiguous one. Here we explore a somewhat different approach. Because initializing the graph is quite expensive, we want to do it as early in the processing chain as possible, and we want to do it just once. The context here consists simply of the tokens in the text (without stopwords); the system is able to provide this information as early as the NLP module is done. When the disambiguation module is put to work, it just chooses the CUI with highest activation among the mapping candidates in the PageRank vector.

4 Evaluation

No corpus of EHR texts in Spanish exists annotated manually with UMLS Metathesaurus concepts. No tool is openly available either for term normalization in the biomedical domain in Spanish. Hence, in an attempt to evaluate our preliminary system, we compare its performance indirectly to that of MetaMap's in two different English-Spanish parallel corpora. The first corpus is a subset of the Scielo corpus made available by Neves et al. (2016). The second is a small set of EHRs, some obtained by Vicomtech-IK4 in the framework of the SEMANHIS project (Gaitek 2015, IG-2015/0001027) and others collected from the Internet, all manually translated.

	Scielo		**EHR**	
	es	**en**	**es**	**en**
# documents	1,895	1,895	18	18
# words	26,490	23,374	23,311	21,093

Table 12: Description of the evaluation corpora

It must be clarified that what follows does not provide performance scores of the separate modules. That is, we have not been able to evaluate none of the modules on their own. As a preliminary work, we evaluate the entire pipeline and, by combining different configurations, we hope to gain some insight about the modules as individual processes.

4.1 Evaluation framework

The procedure for all the experiments is the same: on the one hand, the Spanish corpus is processed with our system and the English corpus with MetaMap (MM). In order to make comparable annotations, the knowledge source of MM has been reduced with the data file builder so that both systems can annotate only the same 352,075 concepts in the UMLS index. Additionally, MM's configuration is set to ignore the word order inside the spans (because our system does) and to perform WSD.

By the evaluations of MM presented earlier, we can safely assume that between a quarter and a half of the annotations made by MM are wrong, and that it misses around a quarter and a half of the entities in the text. As a consequence, MM's annotations cannot be taken as a Gold Standard, and we cannot calculate precision and recall; instead, we report the agreement between both systems by means of Cohen's Kappa coefficient (Cohen, 1960). We are aware of the fact that this evaluation is just a first attempt to measure the performance of our prototype (in Spanish) with respect to a mature system such as MM (in English). High agreement does not indicate good performance, nor does low agreement indicate bad performance. We do not report statistical significance. We cannot conclude from the results whether the prototype performs well or not. The agreement values reported should just be taken as cues or hints for the differences in

performance between the possible configurations of the modules. We provide a manual error and disagreement analysis in Section 4.4 in an attempt to elucidate these issues.

The formula for Cohen's kappa coefficient, κ, is

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{9}$$

where p_o is the proportion of units in which annotators agree (i.e., the observed agreement) and p_e is the proportion in which agreement is expected by chance (i.e., chance agreement). Alternatively, it can be formulated in frequencies as

$$\kappa = \frac{f_o - f_e}{N - f_e} \tag{10}$$

where f_o is the units in which annotators agree, f_e is the amount of agreement expected by chance, and N is the total amount of units annotated. In our framework, the units are the 352,075 concepts in the index; MM and our system agree only when both say that a given concept is present in the input document or when both say that it is not present. N, then, is 352,075 times the document amount of the corpora.

There is no universally accepted interpretation of Cohen's kappa as to what is considered high or low agreement. Landis and Koch (1977) proposed the following, which is widely cited, but has no evidential grounding:

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

The standard error of k, σ_k, is given by

$$\sigma_k = \sqrt{\frac{p_o\,(1 - p_o)}{N\,(1 - p_e)}} = \sqrt{\frac{f_o\,(N - f_o)}{N\,(N - f_e)^2}} \tag{11}$$

Cohen (1960) also formulated the significance of the difference between two independent ks, z, but it cannot be applied to our experiments because they are no independent.

$$z = \frac{k_1 - k_2}{\sqrt{\sigma_{k_1}^2 + \sigma_{k_2}^2}} \tag{12}$$

4.2 Evaluation on the Scielo Corpus

In this set of experiments, the corpus for annotation consists of parallel titles and abstracts of biomedical scientific literature. Specifically, a subset of 2,000 documents was retrieved from Spanish-English Scielo Corpus Neves et al. (2016), and manually revised,

resulting in 1,895 parallel documents. The English documents are on average 12.33 words-longs; their Spanish counterparts are 14 words-long on average.

The experimentation is incremental: first, the difference between using the rule-based boundary match as opposed to pure ngrams is observed. Then, the different re-ranking scores are tested; finally, we experiment with UKB and the other disambiguation baselines. All the experiments are performed on the same corpus; one can think of this as the development dataset, and as the test dataset the one in the next section, Evaluation on EHRs.

Experiment A: Boundary detection

In order to compare the two boundary detection methods proposed, we have annotated the corpus with the system configured in the following two ways: one uses ngram-based detection ("ngram detection") and the other our noun-phrase detector ("phrase detection"); none of the two perform re-ranking of the candidates nor discards candidates by means of a threshold (the scoring function is "lucene$_{(.0)}$"), and to perform disambiguation they use the first candidate in the list (the disambiguation method is "first"). The results of the agreement of these two systems with MM are shown in Table 13.

disambiguation method	**scoring function**	**ngram detection**	**phrase detection**
first	lucene$_{(.0)}$	**0.304 ± 0.006**	0.288 ± 0.006

Table 13: k between MetaMap and the system proposed affected by segmentation

In this experiment we obtain fair agreement with MM with any of the two boundary detection methods. Phrase-based detection yields somewhat less agreement with MM.

Experiment B: Candidate ranking

disambiguation method	**scoring function**	**ngram detection**	**phrase detection**
first	lucene$_{(.0)}$	0.304 ± 0.006	0.288 ± 0.006
	aronson$_{(.5)}$	0.309 ± 0.006	0.288 ± 0.006
	castro$_{(.7)}$	**0.375 ± 0.006**	0.352 ± 0.006

Table 14: k between MetaMap and the system proposed affected by segmentation and candidate ranking

Next, we would like to know how the different functions for candidate re-scoring and thresholds affect the results. First, for each candidate scoring function, namely *Aronson* and *Castro*, the corpus has been annotated using ngram boundary detection and "first" disambiguation, and thresholds ranging from 0.0 to 1.0 applied in each case. The same

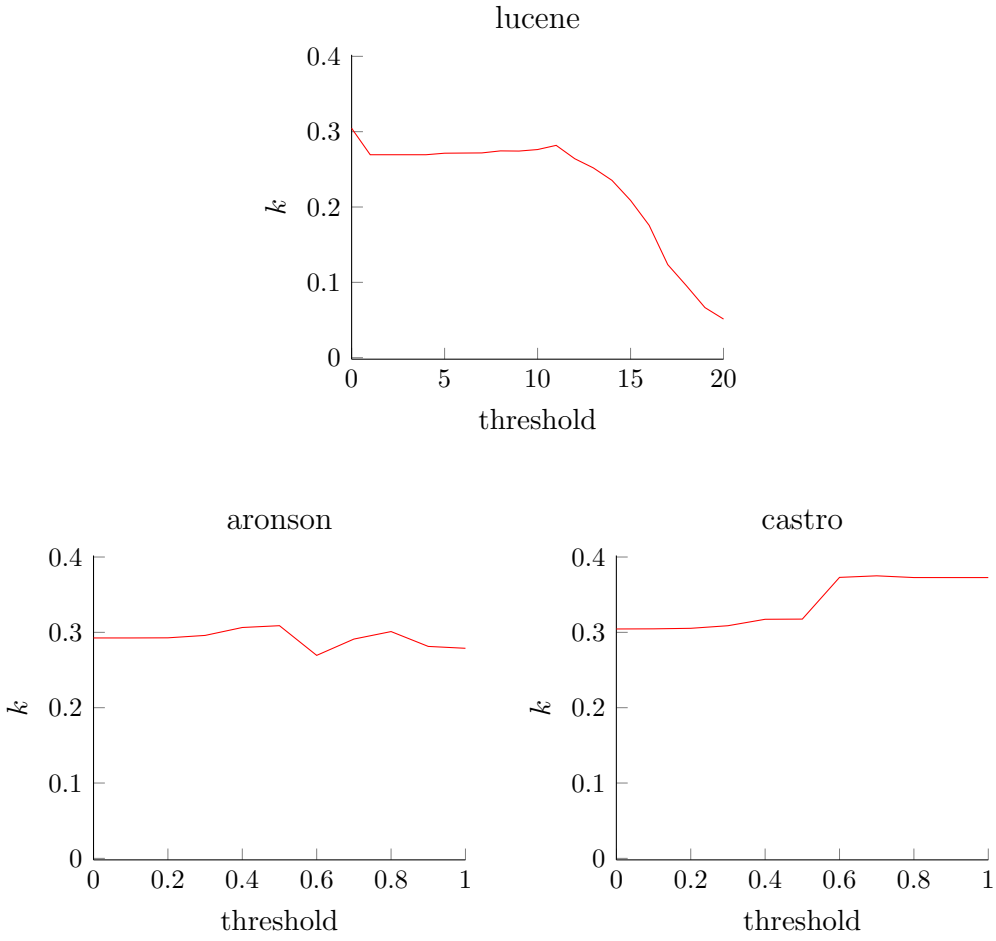


Figure 10: Threshold and k each scoring function

has been done with no re-ranking at all ("lucene"), but applying thresholds from 1.0 to 20.0 (since Lucene's results are not normalized to a range between 0 and 1). The results are shown in Figure 10.

The best agreement with MM is achieved using Castro with a threshold at 0.7: k is 0.375, that is 7 points more of agreement with respect to not re-ranking nor discarding candidates at all —still fair agreement with MM. *Aronson* does not seem to improve the agreement with MM reached by the simpler system.

The thresholds that yield the highest agreements for each re-scoring functions are: 0.0 for "lucene", 0.5 for *Aronson*, and 0.7 for *Castro*. With these winning configurations, we have performed the experiments with the noun-phrase boundary detector as well. Results are shown in Table 14. As can be seen, ngrams always yield a slightly better agreement with MM.

Experiment C: Disambiguation

In order to check whether our application of UKB contributes to the overall performance of the system, we compare it to other three simpler ways of disambiguation: choosing the first candidate of the list ("first"), choosing randomly from the list ("random"), and not choosing any candidate at all ("skip"). We have annotated the corpus with these methods, for each of the best scoring-threshold pairs in the previous experiment, and for each of the two boundary detection methods. The results are shown in Table 15.

desambiguation method	**scoring function**	**ngram detection**	**phrase detection**
skip	$\text{lucene}_{(.0)}$	0.360 ± 0.006	0.340 ± 0.006
	$\text{aronson}_{(.5)}$	0.359 ± 0.007	0.339 ± 0.007
	$\text{castro}_{(.7)}$	0.390 ± 0.006	0.366 ± 0.006
first	$\text{lucene}_{(.0)}$	0.304 ± 0.006	0.288 ± 0.006
	$\text{aronson}_{(.5)}$	0.309 ± 0.006	0.288 ± 0.006
	$\text{castro}_{(.7)}$	0.375 ± 0.006	0.352 ± 0.006
random	$\text{lucene}_{(.0)}$	0.323 ± 0.006	0.304 ± 0.006
	$\text{aronson}_{(.5)}$	0.331 ± 0.006	0.308 ± 0.006
	$\text{castro}_{(.7)}$	0.398 ± 0.006	0.372 ± 0.006
UKB	$\text{lucene}_{(.0)}$	0.343 ± 0.006	0.328 ± 0.005
	$\text{aronson}_{(.5)}$	0.349 ± 0.006	0.330 ± 0.006
	$\text{castro}_{(.7)}$	**0.412 ± 0.006**	0.387 ± 0.006

Table 15: k between MetaMap and the system proposed affected by segmentation, candidate ranking and disambiguation method

Indeed, the best agreement with MM is achieved using UKB with Castro re-ranking at threshold 0.7, and, as always, with ngram-based boundary detection. Agreement is improved by almost 11 points with respect to our first system ("first" disambiguation, re-ranking with "$\text{lucene}_{(.0)}$"), achieving moderate agreement with MM. Most interestingly, using UKB improves agreement with respect to "skip" disambiguation as well, which means that with UKB we can annotate more CUIs and improve agreement at the same time. This is illustrated in Figure 11.

The figure shows for each combination of scoring method and disambiguation the amount of CUIs annotated and the agreement reached with MM. There is a conspicuous gap between using Lucene's score or *Aronson*, and using *Castro*. The former produce around 3 annotations per document (on a total of 1895 documents) and a poorer agreement with MM. Furthermore, using UKB with Lucene or *Aronson* also yields worse agreement with MM as compared to skipping ambiguous spans. With *Castro*, on the other hand, around 2 annotations per document are produced, and the agreement improves when using UKB as compared to skipping ambiguous spans.

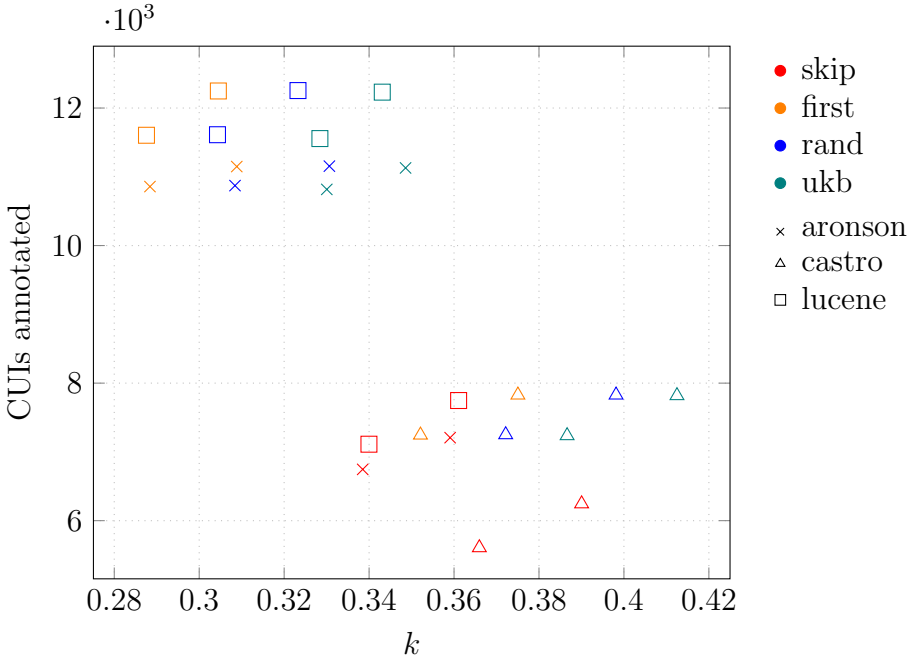


Figure 11: k and amount of CUIs annotated per scoring function and disambiguation method in the Scielo corpus

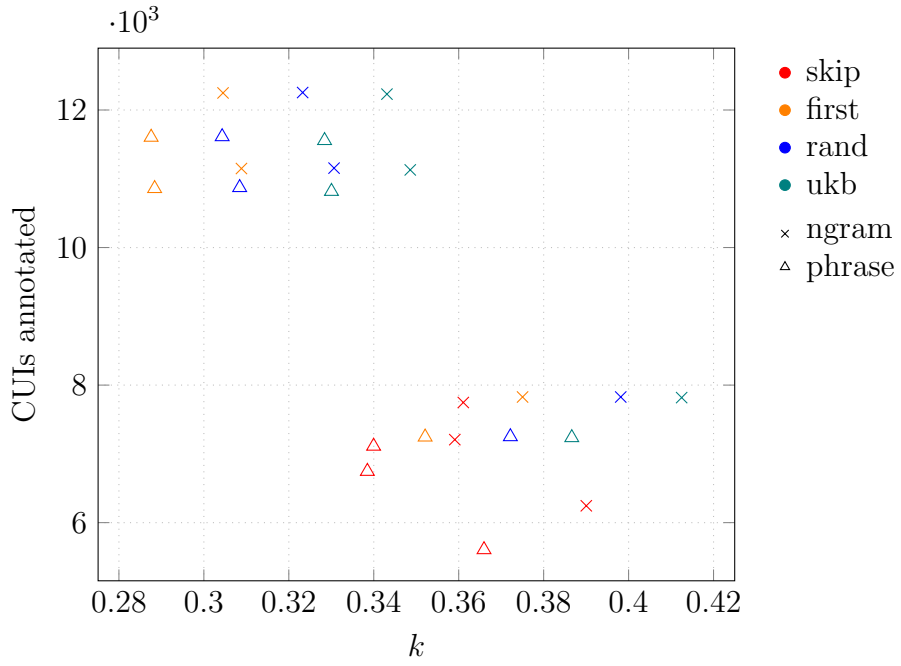


Figure 12: k and amount of CUIs annotated per scoring function and boundary detection method in the Scielo corpus

In Figure 12, results are shown for each combination of boundary detection method and disambiguation method. It is just to visualize how ngrams always yield a slightly better agreement with MM; the data points are actually the same of Figure 12.

4.3 Evaluation on EHRs

In order to evaluate the system in real EHRs, we have manually created a parallel corpus of such documents. On the one hand it contains 10 EHR excerpts, originally in Spanish, from a corpus of records obtained in the framework of the SEMANHIS Project. These records are 503 tokens-long on average. They have been translated to English manually. On the other hand, we have collected 8 health records in English from the Internet[24]. These records are 2,023 tokens-long on average and have been translated to Spanish, manually as well. In the translation process, misspellings have been corrected, so the following experiments cannot be taken to mirror a real use case scenario in this sense.

With the corpus just described, we have replicated Experiment C in the previous section. The results are shown in Table 16, and Figures 13 and 14. As can be seen, the results follow the same pattern as with the Scielo Corpus, in spite of being quite different texts, with the only difference being that the re-ranking function *Aronson* does seem to improve agreement with MM compared to not performing re-ranking. Overall, the best agreement with MM is achieved, again, with ngram-based boundary detection, the *Castro* scoring function, and using UKB to perform sense disambiguation.

desambiguation method	**scoring function**	**ngram detection**	**phrase detection**
skip	$\text{lucene}_{(.0)}$	0.306 ± 0.007	0.288 ± 0.008
	$\text{aronson}_{(.5)}$	0.336 ± 0.008	0.321 ± 0.008
	$\text{castro}_{(.7)}$	0.368 ± 0.008	0.371 ± 0.008
first	$\text{lucene}_{(.0)}$	0.261 ± 0.007	0.245 ± 0.008
	$\text{aronson}_{(.5)}$	0.299 ± 0.008	0.287 ± 0.008
	$\text{castro}_{(.7)}$	0.371 ± 0.008	0.353 ± 0.008
random	$\text{lucene}_{(.0)}$	0.286 ± 0.007	0.266 ± 0.007
	$\text{aronson}_{(.5)}$	0.330 ± 0.008	0.316 ± 0.008
	$\text{castro}_{(.7)}$	0.403 ± 0.008	0.389 ± 0.008
UKB	$\text{lucene}_{(.0)}$	0.321 ± 0.007	0.306 ± 0.007
	$\text{aronson}_{(.5)}$	0.365 ± 0.008	0.354 ± 0.008
	$\text{castro}_{(.7)}$	**0.432 ± 0.008**	0.414 ± 0.008

Table 16: k between MetaMap and the system proposed affected by segmentation, candidate ranking and disambiguation method

[24] `https://www.med.unc.edu/medclerk/`; not available any longer

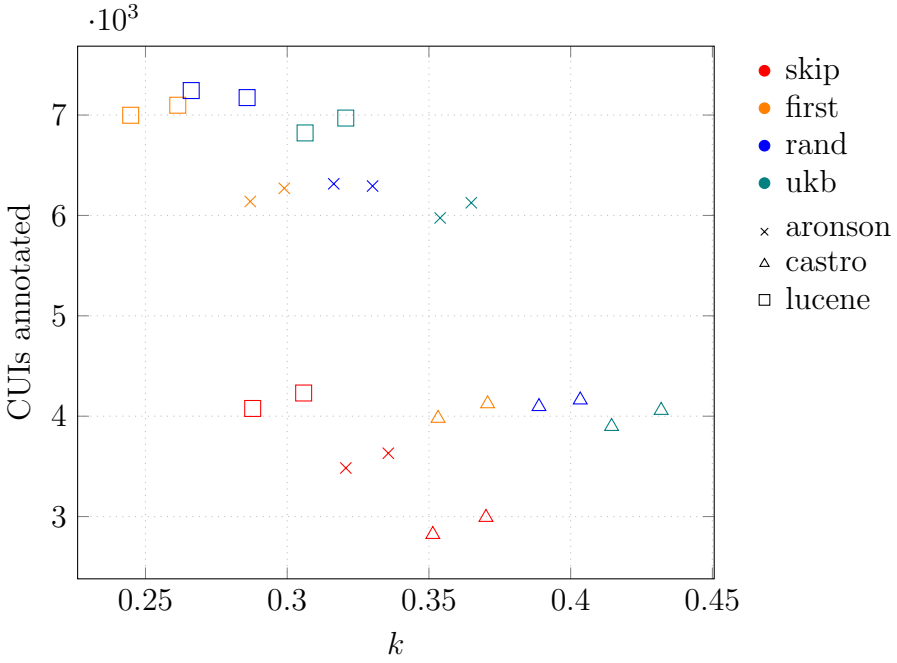


Figure 13: k and amount of CUIs annotated per scoring function and disambiguation method in the EHR corpus

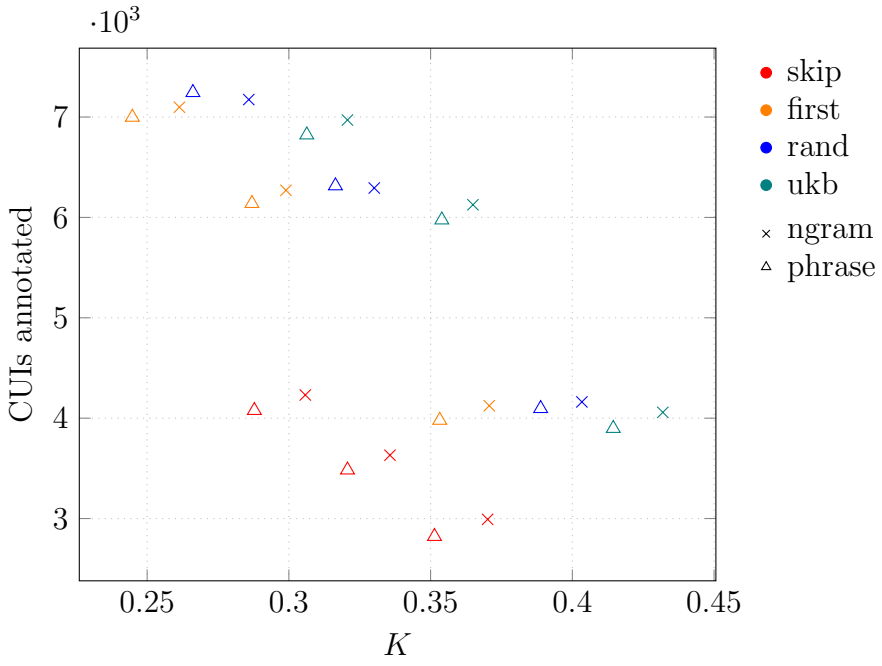


Figure 14: k and amount of CUIs annotated per scoring function and disambiguation method in the EHR corpus

4.4 Disagreement and error analysis

Because of the setup of the evaluation, we need to make a clear distinction between disagreements and errors. *Disagreements* can occur without any of the systems committing an error. That two systems agree does not mean either that they are right necessarily —although it is more likely that they are. *Errors* can be of two types: *false positives* are errors that consist of incorrect annotations; *false negatives* are correct annotations that have been missed. The latter are harder to evaluate, precisely because we do not have a Gold Standard reference nor exhaustive knowledge of the domain. In what follows, we provide a qualitative analysis of disagreement and error.

A manual study of the results has shown that there are two main factors that increase the **disagreement** between MM and our system, which are rooted in the design of the experiments themselves: *a)* not surprisingly, the fact that they are annotating translated texts, i.e., different texts, and *b)* the differences in the knowledge bases. It is possible that one of the bases has better lexical coverage of the terms that can realize a given concept; thus, the system with the richest base is more likely to recognize that concept.

Another evident source of disagreement is the differences in boundary detection: phrase-based boundary detection yields more overlapped and discontinuous spans than ngram-based detection does, which leads to not necessarily incorrect annotations but worse agreement with MM.

As for **errors**, it has been observed that many of the **false positives** in both MM and our prototype occur when input texts contain polysemous terms and the Metathesaurus does not capture the precise senses that those terms convey in the text. Thus, they are annotated with incorrect senses. That is, it is a problem of coverage in the knowledge base combined with the lack of actual understanding of the texts by the systems — annotations are done by means of pure lexical similarity. Let us illustrate the problem: the term "clavo" in Spanish has at least three meanings: *a)* clove (a spice), *b)* nail or rod (a metallic object), and *c)* corn of toe (a disease). All these senses are relevant in the medical domain: clove can cause allergic reactions; nails are used frequently in surgical treatments; foot corn is a disease. However, the term "clavo" is only related to sense *a)* and *c)* in the Metathesaurus. This is not to say that sense *b)* is not represented in the Spanish subset, but that it is not represented as "clavo". As a consequence, whenever a text contains "clavo" (and it does not form a bigger concept with its surrounding words), it will be annotated as being a disease or a spice, even if it is neither of the two.

Another source of false positives in the case of our prototype is the over-generation of spans. The ngram strategy clearly generates spans that are not meant to form syntactic units, and thus neither intended meaning units. In the text fragment "[...] arteria torácica en radiografía [...]" (*chest artery in x-ray*), the bigram [torácica, radiografía] would form a span that would, in turn, trigger mapping candidates consisting of concepts referring to chest x-ray, which is not mentioned in the example fragment. Although the phrase-based strategy was meant to overcome this problem by leveraging syntactic information, the fact that it allows for discontinuous spans produces over-generation too sometimes, especially when coordination and/or enumeration are involved.

Regarding **false negatives**, there are two main reasons for MM or our system to miss

a biomedical concept: on the one hand, it can happen that the concept is not captured in the Metathesaurus at all; on the other hand, it could be that the concept is captured but not as expressed in the text, be it because it is misspelled, abbreviated in a way that the Metathesaurus does not contemplate, or formulated in any other non-standard way. That is, false negatives are caused by a poor coverage of the Metathesaurus and the lexical variability of clinical narrative. MM relies on a powerful tool to deal with variability, the SPECIALIST Lexicon; we do not address variability but for a closed list of abbreviations. As a consequence, our system is much more likely to produce this type of false negatives, in any of its possible configurations. Additionally, phrase-based span generation can miss noun phrases in texts because it relies on linguistic annotations made with tools that are not adapted for the biomedical domain. If it misses a noun phrase and the noun phrase happens to be a relevant term, the term is not annotated.

Below, four examples of annotations are presented and described, which serve as pretexts to discuss some of the points made in more detail. Each example shows, on the one hand, the English text and MM's annotations on it, and, on the other, the corresponding text in Spanish and the annotations made by one of our prototype's possible configurations or more. CUIs highlighted in bold are taken to be correct annotations —to the extent of this work's author's knowledge on the domain, which is admittedly scarce. Example 4.4.2 has been taken from the experiments on EHRs; the rest are from the Scielo corpus. Appendix B contains the meanings of the CUIs that appear in each example.

4.4.1 Example 1

	MM
Should	
we	
rule	C1446409
out	C0439787
congenital	**C1744681**
anesplenia	
?	

	ngrams/Castro$_{(.7)}$/UKB
¿	
debemos	
descartar	**C0332196**
una	
asplenia	**C0600031**
congénita	
?	

Table 17: Error analysis: example 1

The anotations in Table 17 are a straightforward example of how parallel texts cause disagreement. "rule out" in Spanish is translated as "descartar". When MM makes —in this case incorrect— annotations for "rule" and "out", there is no way that our system will make agreeing annotations, since "descartar" does not have the meaning of any of the two words separately. We can also see that MM does not recognize the concept "congenital anesplenia". As it happens, MM's knowledge base contains "congenital asplenia" but not "congenital anesplenia"; then, MM could have only relate the two by means of its powerful variant generation mechanism, which does not seem to help in this particular example. Of course, problems like these occur in both directions. Notice that the agreement between the two systems is 0, but the annotations made by our system happen to be better than MM's (again, in this particular case).

4.4.2 Example 2

This example provides a comparison between the two systems that give the worst agreement with MM and the two that give the best, namely, on the hand, the ones that do not re-rank the candidates ("Lucene") and choose the first candidate in the list in case of ambiguity ("first") and, on the other, the ones that use *Castro* to re-score candidates and UKB to resolve ambiguities.

The text to annotate contains two main entities that should be recognized and identified, which in the English text are expressed as "positron emission tomography" and "lung malignancy". As Table 18 shows, MM identifies both of them correctly. The systems that use *Castro* make annotations with smaller spans, all of them correct. The systems that use Lucene's score, on the other hand, make annotations with bigger spans and are mostly incorrect. None of the two recognize and identify both of the main entities mentioned.

What is going on here is that Lucene, though convenient for its retrieval speed, assigns scores to the retrieved results that are not appropriate for the task of term normalization. Lucene's scoring function does not measure lexical similarity, but relevance. *Castro*, on the other hand, is more rigorous and allows for little deviations from the query.

Finally, notice that in this example we see the phenomenon involving ngram segmentation mentioned before: "pulmonar con tomografía" (*pulmonary with tomography*) is annotated as a term by the first system using ngrams, when actually it does not constitute neither a syntactical nor meaningful unit on its own.

4.4.3 Example 3

In this example (Table 19), the text to annotate contains three entities that should be recognized and identified, which in the English text are expressed as "MicroRNAs", "biomarkers", and "eye diseases". We compare the results of our system using the three scoring functions, always with ngram boundary detection and using UKB for disambiguation.

The most remarkable thing to notice is that all the results miss "MicroRNAs" and normalize correctly the other two terms. "MicroRNAs" in the Spanish text is translated

	MM
a	
positron	**C0032743**
emission	
tomography	
-	
positive	C0439178
lung	**C0242379**
malignancy	
.	

	ngrams/Lucene$_{(.0)}$/first		phrases/Lucene$_{(.0)}$/first
una			
neoplasia	**C1306459**		C0685027
maligna		C0685027	
pulmonar	C2315679		
con			
tomografía		C0040398	**C0040395**
por			
emisión			**C0032743**
de			
positrones	**C0032744**		
positiva	C1636154		C1636154
.			

	ngrams/Castro$_{(.7)}$/UKB	phrases/Castro$_{(.7)}$/UKB	
una		**C0027651**	**C0024121**
neoplasia	**C0027651**		
maligna			
pulmonar	**C2709248**		**C0024121**
con	**C0332287**		
tomografía	**C0040395**	**C0040395**	
por			
emisión			
de			
positrones	**C0032744**	**C0032744**	
positiva			
.			

Table 18: Error analysis: example 2

as "miRNAs". The UMLS Metathesaurus contains "miARNs", but not "miRNAs"; "miARNs' is not retrieved by Lucene, thus the term is not annotated.

As for the scoring functions, the example shows what is the general case: *Lucene* yields the worst agreement with MM because its scoring function is not oriented towards lexical similarity but relevance, thus producing more false positive annotations. *Castro*, in contrast, is less lax in this regard; it seems to favor precision at the expense of recall. It definitely achieves better agreement with MM in our datasets. As for *Aronson*, it does not make a difference in the agreement with MM. This might come as a surprise, because *Aronson* is the function that MM uses; however, it must be pointed out that we do not take advantage of the function in its full potential, because we do not generate span variants to query the index, information that the function can leverage to penalize variants that are too different from the original span.

	MM	Lucene$_{(.0)}$	Aronson$_{(.5)}$	Castro$_{(.7)}$
MicroRNAs	**C1101610**			
as				
potential				
biomarkers	**C0005516**			
of				
eye	**C0015397**			
diseases				
.				
Los		C0024015		
miRNAs				
como				
potenciales		C0025251	C0025251	
biomarcadores		**C0005516**	**C0005516**	**C0005516**
de				
las				
enfermedades		**C0015397**	**C0015397**	**C0015397**
oculares				

Table 19: Error analysis: example 3

4.4.4 Example 4

This example (Table 20) shows the contribution that disambiguation with UKB makes to the overall result. The text contains four terms: "hinchazón" (*swelling*), "lengua" (*tongue*), "garganta"(*throat*) and "angioedema de ACEI" (*ACEI angioedema*). Using ngram detection and *Castro* to rerank mapping candidates, one unambiguous term is found: "hinchazón". The rest of the terms are ambiguous, and using UKB we get correct senses for each of them. Notice that MM and our prototype assign different

CUIs to the term "tongue", and both are correct (and "ACEI angioedema" is not in the Metathesaurus, so "angioedema" is annotated instead). Without further analysis, one might think that UKB is doing a great job. The truth is, however, that all the senses that each of these ambiguous terms have would be correct in this context, so choosing CUIs randomly would have performed equally well.

	MM
the	
swelling	**C0013604**
is	
restricted	C0443288
only	C1720467
to	
her	
tongue	**C0040408**
and	
throat	**C0031354**
typical	C0332307
of	
ACEI	
angioedema	**C0002994**
.	

	ngrams/Castro$_{(.7)}$/skip	ngrams/Castro$_{(.7)}$/UKB
la		
hinchazón	**C0038999**	**C0038999**
se		
limita		
sólo		
a		
su		
lengua		**C1278913**
y		
garganta		**C0031354**
típico		
de		
angioedema		**C0002994**
de		
ACEI		
.		

Table 20: Error analysis: example 4

To understand the phenomenon behind this example we need to revisit how our UMLS index has been created: we have indexed 352,075 concepts that are realized by

546,309 terms, of which 2,147 are ambiguous (i.e., they are related to different CUIs). For each of the terms, we have also indexed normalized versions that do not contain stopwords nor spurious parenthetical content. Normalized terms are used for candidate mapping generation, that is, spans are queried in the UMLS index against the normalized version of the terms. In their new form, 9,973 of the terms are ambiguous. Since what we erase is spurious, we assume that the ambiguities inserted do make sense.

Let us elaborate on this idea with a real example: take the terms in the Metathesaurus "lengua" and "lengua [como un todo]" (*tongue [as a whole]*). Each is related to a different concept. As such, they are not ambiguous: two different terms related to different CUIs. Because "[como un todo]" is regarded spurious parenthetical concept, we remove it, and now "lengua" is related to two different CUIs –now it is ambiguous. The difference between the concepts captured with the terms "lengua" and "lengua [como un todo]" in the Metathesaurus are not easy to account for. We must look at their relationships: "lengua [como un todo]" does not have children, and is in turn one of the children of "lengua"; other children of "lengua" refer to parts of the tongue, such as the tongue artery, the aponeurosis of the tongue, and so on. Because "lengua" is the parent of all the concepts that refer to the parts that compose the tongue, one could think that "lengua" actually represents the concept of the tongue as a whole —as we already know, however, there exists another concept that captures precisely that, namely "lengua [como un todo]", which is, in fact, a children of "lengua". So, making "lengua" ambiguous is not completely inaccurate. If the text to annotate does not state explicitly that it is referring to the tongue as a whole, which concept ("tongue" or "tongue as a whole") would be the correct one? The first, maybe, because it respects the underspecification in the original text? Or the second, because if the text referred to a part of the tongue, it would be clearly stated? It simply depends on the task in which the term identities are to be used. We do not make any compromise in this respect, we say that both senses are correct. That is the reason why having UKB resolve these ambiguities is not praiseworthy. The ambiguities of "lengua", "garganta" and "angioedema" in the annotations shown in this example are of the type just explained.

It would be more interesting to look at annotations of ambiguous terms that have clearly distinct senses, such as "clavo". This terrm occurs twice in our evaluation dataset:

"intervenida quirúrjicamente el 5/3/15 mediante **clavo** gamma corto." → *nail*

"el pt no responde, no se retira del dolor (presión en **clavo**)" → *corn of foot*

In both cases, "clavo" is incorrectly assigned the sense of "clove", the spice. The term "menor" has two senses in the Metathesaurus: one is "underage", and the other is "lesser" or "minor". In the following fragment, "menor" has the second meaning and our prototype disambiguates it correctly:

"en **menor** medida lumbares y dorsales"

The term "derecho" is highly ambiguous in Spanish; it can mean —at least— *a)* "right" as in opposed to left, *b)* "right" as in moral or legal entitlement, *c)* "straight", or *d)* "law". The UMLS Metathesaurus associates "derecho" to senses *a)* and *c)*. In the following fragment, "derecho" has sense *a)*, and the system annotates it correctly:

"el riñón **derecho** presenta cierto retraso en la eliminación de ..."

The term "familiar" has two senses: "relative" or "member of the family", and also

"familiar", to describe something that is known or can be recognized. In the following phrase, "moribundo" has the former sense; the system identifies it correctly:

"un **familiar** moribundo"

The point we are trying to make with all these examples is that we do not know whether UKB actually helps beyond the ambiguities explained at the beginning of this section. Sometimes it does, others it does not. A much more exhaustive and rigorous evaluation would have to be made that assessed disambiguation specifically using a an annotated corpus.

4.5 Summary

It is not possible to measure precision and recall of our prototype at the moment, as is the standard method for this type of applications, simply because there is no Gold Standard to do so. There is no tool available either to compare ours with that performs biomedical term normalizations with UMLS of Spanish texts. In this context, we have proposed an indirect evaluation that consists in calculating the agreement between our prototype and MetaMap. We acknowledge that this evaluation does not measure the performance of the prototype presented, it can only tell us at best how much MetaMap and our system agree. The experiments show that

- the best agreement with MM (moderate agreement) is reached using the re-scoring function by Castro et al. (2010), and
- UKB can improve agreement when combined with this function score as compared to the random baseline.

In a qualitative disagreement and error analysis we make the following observations:

- The main source of disagreement is, of course, the fact that MM and the application presented annotate different texts; furthermore, they use different sources of knowledge, although we have attempted to make them as similar as possible by limiting MM's knowledge base to contain only the concepts that we have indexed for our system.
- Many false positive errors are produced due to the fact that the Metatheasurus does not capture all the possible meanings of the terms it contains. Because candidates are scored simply by means of lexical similarity, the system will annotate a term that is similar enough to an entry in the MM even if they denote different concepts.
- F alse positives are also produced when using the scores given by Lucene to rank candidates, and even when re-ranked using the function by Aronson (2001).
- Segmentation is another source of false positives: ngram-based and noun-phrase-based detection generate incorrect spans that get to be annotated eventually.

- As for false negatives, they mainly occur because the Metathesaurus does not capture the lexical variability in the input texts, and we do not treat this problem other than with the expansion of around 2300 abbreviations and acronyms.
- Phrase-based span detection is another source of false positives, as it can miss noun phrases due to errors in the lower-level processing of the input texts.
- We cannot draw any conclusion as to whether the way we implement UKB helps to choose the correct sense of ambiguous terms.

5 Conclusions and Future Work

We have presented a preliminary pipeline to perform biomedical term normalization in Spanish clinical texts with the UMLS Metathesaurus. Term normalization consists in linking the entities mentioned in the text to entries in knowledge bases, such as terminologies or ontologies, providing a unique identifier to the entities.

The motivation for addressing this problem is primarily that there exists extensive knowledge captured in electronic health records' narrative text, which is intractable by health care practitioners. Natural Language Processing (NLP) and specifically Information Extraction (IE) tools, such as term normalization tools, can help unearth and structure that knowledge.

The contributions of this work include:

1. A novel prototype application that performs term normalization of Spanish clinical text with the UMLS Metathesaurus.
2. An evaluation of the ixa-pipes part-of-speech tagger in clinical text.
3. The creation of two parallel Enligsh-Spanish datasets to evaluate the performance of the prototype.
4. A qualitative analysis of the errors that the prototype commits.
5. A graphical interface for the prototype.
6. An API to consult the UMLS Metathesaurus and Semantic Network.

There are many aspects of the prototype to work on. In the first place, the pipeline itself, that is, the chosen composition of modules is only a proof of concept. For instance, the mapping strategy, which favor longest match, should change to fit the task in which the annotations are to be used. Also, UKB could be activated earlier in the workflow, not only when the system needs to choose between mapping candidates that have the same score. Furthermore, each of the modules and resources needs substantial improvement —not one step of the processing is trivial, from the most basic (how to tokenize, how to treat capitalization, and so on) to the more complex (for example, how to leverage information coming from different sections of EHRs: summary, discharge, diagnosis... they are likely to contain information of different nature).

Lexical variability is not addressed in this work, and it is one of the aspects that affect most its performance. As future work, we plan to explore methods of terminology expansion to increase the coverage of the Metathesaurus. This could be done by increasing the index itself, that is, adding more terms per concept, or by generating variants of spans at runtime like MetaMap does. An interesting alternative would be to abstract from terms as discrete symbols and explore a different representation, such as word or phrase embeddings. With enough clinical text, this approach could even allow us to dispose of most low-level processing tools, which are dependent on language and domain. It would be nice, however, to have tokenizers, part-of-speech taggers and parsers adapted to process clinical narrative in Spanish, in order to promote the use of NLP in the biomedical domain in this language.

Regarding disambiguation, this work explores only one usage setup of one disambiguation tool, namely UKB. We have not looked at how different context sizes affect

the result, for instance. Another line of work is optimizing UKB's knowledge graph, which in this work contains all the concepts in the UMLS index and the relations between them. It is possible that using only a subset of concepts and/or of relations makes disambiguation more efficient.

As future work for the longer term, it is paramount to create a gold standard corpus of Spanish clinical text if this field of research is to advance in any way. We need corpora to evaluate tools in reproducible frameworks and to obtain comparable results. If substantial data is annotated, machine learning techniques could also be explored.

References

Proyecto HCDSNS Historia Clínica Digital del Sistema Nacional de Salud - Informe de Situación (Agosto 2017). Technical report, Ministerio de Sanidad, Servicios Sociales e Igualdad, 2017.

Informes SEIS: De la historia clínica a la historia de salud electrónica (5). Technical report, Sociedad Española de Informática de la Salud, 2003.

R. Agerri, J. Bermudez, and G. Rigau. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

E. Agirre and A. Soroa. Personalizing PageRank for Word Sense Disambiguation. *Proceedings of the 12th Conference of the European Chapter of the ACL*, (April):33–41, 2009.

E. Agirre, A. Soroa, and M. Stevenson. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896, nov 2010.

A. R. Aronson. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. In *Proceedings of the AMIA Symposium*, pages 17–21. 2001.

A. R. Aronson. MetaMap: Mapping Text to the UMLS Metathesaurus. 2006.

A. R. Aronson and F. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, (17):229–236, 2010.

A. R. Aronson and T. C. Rindflesch. Query expansion using the umls metathesaurus. In *Proceedings of the AMIA Annual Fall Symposium*, page 485, 1997.

A. R. Aronson, T. C. Rindflesch, and A. C. Browne. Exploiting a large thesaurus for information retrieval. In *Intelligent Multimedia Information Retrieval Systems and Management*, volume 1, pages 197–216, 1994.

N. Bhatia, N. H. Shah, D. Rubin, A. P. Chiang, and M. A. Musen. Comparing Concept Recognizers for Ontology - Based Indexing : MGREP vs . MetaMap. *AMIA Summit on Translational Bioinformatics*, 2009.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, (30):107–117, 1998.

J. R. Campbell, P. Carpenter, C. Sneiderman, S. Cohn, C. G. Chute, and J. Warren. Phase II Evaluation of Clinical Coding Schemes: Completeness, Taxonomy, Mapping, Definitions, and Clarity. *Journal of the American Medical Informatics Association*, IV(3):238–251, 1997.

X. Carreras, I. Chao, L. Padró, and M. Padró. Freeling: An Open-Source Suite of Language Analyzers. *Proceedings of the 4th Language Resources and Evaluation Conference (LREC 2004)*, IV:239–242, 2004.

F. M. Carrero, J. C. Cortizo, and J. M. Gómez. Building a Spanish MMTx by Using Automatic Translation and Biomedical Ontologies. In *Intelligent Data Engineering and Automated Learning – IDEAL 2008*, pages 346–353. Springer Berlin Heidelberg, 2008a.

F. M. Carrero, J. C. Cortizo, J. M. Gómez, and M. de Buenaga. In the development of a Spanish Metamap. In *Proceedings of the 17th ACM conference on Information and knowledge mining - CIKM '08*, pages 1465–1466. ACM Press, 2008b.

E. Castro, A. Iglesias, P. Martínez, and L. Castaño. Automatic Identification of Biomedical Concepts in Spanish Language Unstructured Clinical Texts. In *Proceedings of the 1st ACM International Health Informatics Symposium (IHI'10)*, pages 751–757. ACM, 2010.

M. F. Chiang, D. S. Casper, J. J. Cimino, and J. Starren. Representation of ophthalmology concepts by electronic systems: adequacy of controlled medical terminologies. *Ophthalmology*, 112(2):175–183, 2005.

C. G. Chute, S. P. Cohn, K. E. Campbell, D. E. Oliver, and J. R. Campbell. The Content Coverage of Clinical Classifications. *Journal of the American Medical Informatics Association*, III(3):224–233, 1996.

J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, XX(1):37–46, 1960.

M. Dai, N. H. Shah, W. Xuan, M. A. Musen, S. J. Watson, B. D. Athey, F. Meng, et al. An efficient solution for mapping free text to ontology terms. *AMIA Summit on Translational Bioinformatics*, 21, 2008.

Ministerio de Sanidad, Servicios Sociales, e Igualdad. Real Decreto 69/2015, de 6 de febrero, por el que se regula el Registro de Actividad de Atención Sanitaria Especializada. *BOE núm. 35, de 10 de febrero de 2015*, pages 10789–10809, 2015.

D. Demner-Fushman, W. W. Chapman, and C. J. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5): 760–772, oct 2009.

R. S. Dick, B. S. Elaine, and D. E. Detmer, editors. *The Computer-Based Patient Record: An Essential Technology for Health Care*. National Academy Press, 1997.

G.y Divita, T. Tse, and L. Roth. Failure analysis of metamap transfer (mmtx). In *Medinfo*, pages 763–767, 2004.

A. Fokkens, A. Soroa, Z. Beloki, G. Rigau, W. R. Van Hage, and P. Vossen. NAF: the NLP Annotation Format. Technical report, 2014. URL `http://www.newsreader-project.eu/files/2013/01/techreport.pdf`.

C. Friedman. A broad-coverage natural language processing system. In *Proceedings of the AMIA Symposium*, page 270. American Medical Informatics Association, 2000.

C. Friedman. Discovering Novel Adverse Drug Events Using Natural Language Processing and Mining of the Electronic Health Record. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 1–5. Springer Berlin Heidelberg, 2009.

C. Friedman, P. O. Alderson, J. H. M. Austin, J. J. Cimino, and S. B. Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, 1994.

C. Funk, W. Baumgartner, B. Garcia, C. Roeder, M. Bada, K. B. Cohen, L. E. Hunter, and K. Verspoor. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*, 15(1), 2014.

V. N. Garla and C. Brandt. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. *Journal of the American Medical Informatics Association*, 20(5):882–886, 2013.

F. González and D. Luna. La historia clínica elctrónica. In J. Carnicero and A. Fernández, editors, *Manual de Salud Electrónica para directivos de servicios y sistemas de salud*, chapter II, pages 75–96. Naciones Unidas, 2014.

Taher H Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526. ACM, 2002.

C. C. Huang and Z. Lu. Community challenges in biomedical text mining over 10 years: Success, failure and the future. *Briefings in Bioinformatics*, 17(1), 2016.

B. L. Humphreys, A. T. McCray, and M. L. Cheh. Evaluating the Coverage of Controlled Health Data Terminologies: Report on the Results of the NLM/AHCPR Large Scale Vocabulary Test. *Journal of the American Medical Informatics Association*, IV(6): 484–500, 1997.

L. Hunter and K. B. Cohen. Biomedical language processing: what's beyond pubmed? *Molecular Cell*, 21(5):589–594, 2006.

A. Iglesias, E. Castro, R. Pérez, L. Castaño, P. Martínez, J. M. Gómez, S. Kohler, and R. Melero. MOSTAS: Un Etiquetador Morfo-Semántico, Anonimizador y Corrector de Historiales Clínicos. *Procesamiento del Lenguaje Natural*, (41):229–300, 2008.

C. Jonquet, N. H. Shah, and M. A. Musen. The open biomedical annotator. *Summit on translational bioinformatics*, 2009, 2009.

M. Krauthammer and G. Nenadic. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, (37):512–526, 2004.

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

C. P. Langlotz and S. A. Caldwell. The completeness of existing lexicons for representing radiology report information. *Journal of digital imaging*, 15:201–205, 2002.

R. Leaman, R. Khare, and Z. Lu. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57, 2015.

B. T. McInnes. An unsupervised vector approach to biomedical term disambiguation: integrating umls and medline. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop*, pages 49–54. Association for Computational Linguistics, 2008.

G. H. Merrill. Concepts and synonymy in the UMLS Metathesaurus. *Journal of biomedical discovery and collaboration*, 4(7):7, 2009.

S. Meystre and P. J. Haug. Evaluation of Medical Problem Extraction from Electronic Clinical Documents Using MetaMap Transfer (MMTx). *Studies in health technology and informatics*, 116:823–828, 2005.

I. Montoya. *Etiquetado de Historiales Medicos Mediante SNOMED CT y CIE-10*. Master's thesis, Konputazio Ingeniaritza eta Sistema Adimentsuak Unibertsitate Masterra, Euskal Herriko Unibertsitatea (UPV/EHU), forthcoming.

R. Navigli and M. Lapata. Graph connectivity measures for unsupervised word sense disambiguation. In *IJCAI*, pages 1683–1688, 2007.

M. Neves, A. J. Yepes, and A. Névéol. The Scielo Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2942–2948, Portorož, 2016. European Language Resources Association (ELRA).

M. Oronoz, A. Casillas, K. Gojenola, and A. Pérez. Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names. In J. Ruiz-Shulcloper, editor, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, CIARP 2013*, pages 536–543. Springer Berlin Heidelberg, 2013.

S. Pakhomov. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 160–167. Association for Computational Linguistics, 2002.

Jon Patrick, Yefeng Wang, and Peter Budd. An automated system for conversion of clinical notes into snomed clinical terminology. In *Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68*, pages 219–226. Australian Computer Society, Inc., 2007.

W. Pratt and M. Yetisgen-Yildiz. A Study of Biomedical Concept Identification: MetaMap vs. People. *Journal of the American Medical Informatics Association*, pages 529–533, 2003.

J. Pustejovsky, J. Castano, B. Cochran, M. Kotecki, M. Morrell, and A. Rumshisky. Extraction and disambiguation of acronym-meaning pairs in medline. *Medinfo*, 10 (2001):371–375, 2001.

G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, (17):507–5013, 2010.

N. H. Shah, N. Bhatia, C. Jonquet, D. Rubin, A. P. Chiang, and M. A. Musen. Comparison of concept recognizers for building the open biomedical annotator. *BMC bioinformatics*, 10(9), 2009.

R. Sinha and R. Mihalcea. Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In *International Conference on Semantic Computing (ICSC 2007)*, pages 363–369. IEEE, 2007.

S. A. Stewart, M. E. Von Maltzahn, and S. S. Raza Abidi. Comparing metamap to mgrep as a tool for mapping free text to formal medical lexicons. In *Proceedings of the 1st International Workshop on Knowledge Extraction and Consolidation from Social Media (KECSM2012)*, pages 63–77, 2012.

G. Tsatsaronis, M. Vazirgiannis, and I. Androutsopoulos. Word sense disambiguation with spreading activation networks generated from thesauri. In *IJCAI*, volume 7, pages 1725–1730, 2007.

J. Yetano. Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias. 2003.

A Spurious parenthetical content in Spanish in the UMLS Metathesaurus

(trastorno)
(procedimiento)
(hallazgo)
(organismo)
(sustancia)
(estructura corporal)
(producto)
(objeto físico)
(calificador)
(entidad observable)
(anomalía morfológica)
(situación)
(ocupación)
(evento)
(espécimen)
(IF)
(medio ambiente)
(atributo)
(escala de evaluación)
(FC)
(concepto para navegación)
(célula)
(localización geográfica)
(estructura celular)
(persona)
(IB)
(sitio combinado)
(CMI)
(como un todo)
(grupo étnico)
(elemento de registro)
(estadificación tumoral)
(SMQ)
(espacio de nombres)
(metadato fundacional)
(fuerza física)
(acción)
(CMB)
(despuès de estimulación)
(propiedad)
(FISH)
(NEOM)
(general)
(Estados Unidos)
(procedimiento separado)
(polvo para reconstituir)
(EEUU)
(EE UU)
(s)
(en remisión)
(Enzima)
(RIA)
(sustancia del cuerpo)
(Planta)
(en agua)
(mama)
(Psicología)
(menos de una hora)
(manual)
(clínico)
(incluye obtencion del injerto)
(metadato del núcleo)
(región superficial)
(NADPH)
(máquina)
(polen)
(posición social)
(FIGO)
(triple calentador)
(NADH)
(melanoma cutáneo)
(contexto social)
(anote separadamente ademas del codigo para el procedimiento primario)
(NMO)
(estilo de vida)
[como un todo]
[localización]
[objeto físico]

B Meaning of CUIs in error analysis

Example 1	
C0332196	Rule out
C0439787	Out (direction)
C0600031	Congenital absence of spleen
C1446409	Positive for
C1744681	Congenital

Example 2	
C0024121	Lung neoplasm
C0027651	Neoplasm
C0032743	Positron emission tomography (PET)
C0032744	Positron
C0040395	Tomography
C0040398	Radionuclide-Computed Tomography
C0242379	Malignant Lung Neoplasm
C0332287	In addition to
C0439178	Percent positive cells
C0685027	Secondary malignant neoplasm of hilus of lung
C1306459	Cancer
C1636154	Positive dysphotopsia
C2315679	Biopsy of lung using computed tomography (CT) guidance
C2709248	Pulmonary

Example 3	
C0005516	Biological markers
C0015397	Disorder of eye region
C1101610	MicroRNAs

Example 4	
C0002994	Angioedema
C0013604	Edema
C0031354	Pharynx
C0038999	Swelling
C0040408	Tongue
C0332307	Type
C0443288	Constraint
C1278913	Tongue
C1720467	Only - dosing instruction fragment

C Demonstrator

A web-based demonstrator has been developed that allows users to introduce a text of their choosing and visualize the mappings produced by the application in an interactive user interface.

The client side of the demonstrator has been developed in Angular2. By running the system as a webservice, the demonstrator can communicate with it via HTTP. The demonstrator also communicates with an additional webservice that provides an API to query the UMLS Metathesaurus and Semantic Networks themselves, in order to enrich the demonstrator with information about the concepts that have been mapped. This webservice is the subject of the next Appendix (D).

The demonstrator consists of two pages:

In the initial page (Figure 15), users introduce their text and configure the application according to the parameters presented in Section 3.1.2 (except for the amount of results retrieved by Lucene, which is fixed to 150). Users can also choose which semantic types of the Semantic Network they are interested in; the bottom part of the page contains the whole Semantic Network in the form of a tree than can be expanded and collapsed for users to check those types they want the system to map. Notice that this is done in the client side of the demonstrator, that is, the application we have presented in this work does not offer this functionality.

The initial page leads to the result page (Figure 16) when the user submits the form. The result page is divided in three columns. The submitted text is located in the center. Annotations are marked in the text with different colors, depending on the semantic type of the concepts. On the left side, annotations, represented by the preferred names of the concepts, are shown grouped by semantic types. When the user clicks on one of the names, information about that concept appears on the right side of the page: preferred name, semantic types, a definition, and so on. Moreover, the user can also see hypernym and hyponym relations, and navigate through the concepts within this hierarchy.

Figure 15: Home page of the demonstration webpage

Figure 16: Result page of the demonstration webpage

D UMLS webservice

The U.S. National Library of Medicine has an open REST API[25] that enables users to query the UMLS for information about concepts, sources, and so on. Because this API did not cover our needs at the time of developing the demonstrator introduced in the preceding section, we decided to build our own.

It provides an interface to retrieve information from the UMLS Metathesaurus and Semantic Network (2016AA) with convenient filters, such as language and source. These knowledge sources have been loaded in MySQL databases that are consulted through an application developed in Java 8. This application consist of a Spring Boot REST API that has the following paths:

/cui
Retrieves the following information about a given CUI: atoms, preferred term, definitions and semantic types
Options: limit the search to or exclude information in a certain language, limit the search to or exclude information coming from a certain source, limit the search to a certain term type, include/exclude obsolete entries, include/exclude suppressible entries

/cui/atoms
Retrieves the atoms of a CUI
Options: limit the search to or exclude information in a certain language, limit the search to or exclude information coming from a certain source, limit the search to a certain term type, include/exclude obsolete entries, include/exclude suppressible entries

/cui/definitions
Retrieves the definitions related to a CUI
Options: limit the search to or exclude information in a certain language, limit the search to or exclude information coming from a certain source

/cui/children
Retrieves the CUIs of the children of a given CUI
Options: limit the search to or exclude information coming from a certain source, include/exclude obsolete entries, include/exclude suppressible entries

[25] `https://documentation.uts.nlm.nih.gov/rest/home.html`

/cui/parents
Retrieves the CUIs of the parents of a given CUI
Options: limit the search to or exclude information coming from a certain source, include/exclude obsolete entries, include/exclude suppressible entries

/cui/preferredTerm
Retrieves a preferred term of a CUI
Options: limit the search to or exclude information in a certain language, limit,the search to or exclude information coming from a certain source

/cui/relations
Retrieves all the CUIs related to a given CUI, optional by a given type of relation
Options: limit the search to a given relation type, limit the search to or exclude information coming from a certain source, include/exclude obsolete entries, include/exclude suppressible entries

/cui/semTypes
Retrieves the semantic types assigned to a given CUI

/semType/children
Retrieves the child semantic types of a given type in the Semantic Network

/semType/parents
Retrieves the parent semantic types of a given type in the Semantic Network

/source
Retrieves information about a source, such as the date of last modification, the language, and the full formal name