Spanish Named Entity Recognition in the Biomedical Domain

Viviana Cotik[1], Horacio Rodrguez[2], and Jorge Vivaldi[3]

[1] Department of Computer Science, FCEyN, Universidad de Buenos Aires, Argentina
vcotik@dc.uba.ar
[2] Polytechnical University of Catalonia, Barcelona, Spain
horacio@lsi.upc.edu
[3] Universitat Pompeu Fabra, Barcelona, Spain
jorge.vivaldi@upf.edu

Abstract. Named Entity Recognition in the clinical domain and in languages different from English has the difficulty of the absence of complete dictionaries, the informality of texts, the polysemy of terms, the lack of accordance in the boundaries of an entity, the scarcity of corpora and of other resources available. We present a Named Entity Recognition method for poorly resourced languages. The method was tested with Spanish radiology reports and compared with a conditional random fields system.

Keywords: Named Entity Recognition · Spanish · radiology reports · BioNLP.

1 Introduction

Named entity recognition (NER) is an information extraction task, whose goal is to identify instances of specific kind of information units in text and assign them a class. It was originally applied to carefully-written text, such as newswire. Afterwards, it began being applied to other domains, such as the biomedical, for identifying genes, proteins, drug names and diseases, among others.

The approaches to solve the NER problem include: dictionary-based, rule-based, statistical-based, machine learning (ML) and combined approaches [10,27].

The biomedical domain is specially challenging due to 1) its highly specialized terminology including a lot of often polysemous abbreviations and acronyms, 2) the use of non-standardized naming conventions and the lack of standards, even among specialists, regarding to which is the boundary of an entity, and 3) the variety of genres and author profiles, owning specific jargon and sub-languages. In addition, following situations, that highlight the challenges of NER task in the biomedical domain are described in [10,18]:

- the absence of complete dictionaries for some biological or medical named entities (NEs) and the fact that new entities are added frequently,
- abbreviations and other medical terms are often polysemous,
- there might be different ways of referring to the same entity, and

- frequently, medical terms are multi-word units, so there is a need for determining name boundaries and resolving overlap of candidate names. As mentioned in [18], it is easier for a system and for a human to determine if an entity is present or not in a text than to determine its boundaries.

Additionally, there is no standard criteria in the evaluation of biomedical NER systems. Not only the boundary of named entities, but also their class is often ambiguous, due to criteria differences among specialists. Therefore, different matching criteria have been used for Bio-NER[4] system evaluations. Furthermore, datasets are usually not published due to confidentiality issues. Accordingly, usually gold standards have to be generated. The lack of standard metrics, of publicly available datasets, and of standard annotation criteria makes the comparison of different implementations difficult. The processing of medical reports in languages other than English, such as Spanish adds a further difficulty, since there are less resources available.

In this paper we describe different approaches we have followed in order to detect anatomical entities (AEs) and clinical findings (FIs) in a set of Spanish radiology reports. The recognition of these entities is useful because: a) it enables the possibility to structure and normalize the information, b) it offers the opportunity to detect relations among findings and anatomical sites where they occurred, c) if negation is taken into account, identifying which reports contain clinical findings could allow the indexing of only relevant documents and discard those which are not relevant (do not contain clinical findings). This is as a classification task and can serve for the purposes of identifying later on, which are the specific occurrence of clinical findings in the relevant reports, and d) it could serve to notify physicians about the findings, some of which could require immediate action (alert generation). The obtention of timely information is critical in case of urgent or important findings [6]. Its automatic detection and communication is being studied [12,17].

Most of the work in biomedical NER has focused in the recognition of gene and protein names in formal texts and for English.

To detect entities, we propose and evaluate two different approaches: 1) SiMREDA, a Simple Entity Detection Algorithm for Medium Resources languages, that is based on a lookup of terms from a specialized vocabulary, on morphological knowledge and on knowledge of PoS tag patterns of AEs and FIs, and that was conceived by us as a method for BioNER in poor and medium resource languages, and 2) a ML approach, based on conditional random fields (CRF). The rest of the paper is organized as follows. Section 2 presents previous work in the NER domain. Section 3 presents the data used and the methods developed. Section 4 shows the results obtained, which are discussed after, in Section 5. Finally, Section 6 presents conclusions and future work.

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

2 Previous Work

A number of surveys have been carried out on the NER task. Various address the biological domain [3,8,28,30].

Spanish has been introduced in CoNLL-2002 and MET-1 events. Usually efforts in NER are dedicated to a specific genre and domain. To port a system to a new domain or textual genre constitutes a major challenge [22]. An overview of works dedicated to different genres and domains and also reference to previous studies can be seen in [20]. The initial approaches for NER were dictionary and rule-based. The first ones look for the appearance of terms belonging to terminologies in the texts (with exact or exact string matching). Rule-based techniques use domain knowledge or information obtained through analysis of a subset of the data. They usually have good results [31], but its construction is time consuming and often not reusable in other datasets.

Statistical methods are also used for NER. They are sometimes combined with dictionary or rule-based techniques [4]. ML methods can be supervised, for which a considerable amount of training data is needed, semi-supervised, as bootstrapping, or unsupervised. Among the supervised methods, there are classification-based and sequence-based approaches. Examples of the first are Naive Bayes (NB) and Support Vector Machines (SVM) [29]. Sequence-based approaches consider sequences of words instead of individual words or phrases considered in the classification-based approaches. Some examples include Hidden Markov Models (HMM) [26] and Conditional Random fields (CRF).[5] CRFs were the best performing systems in various challenges and have been highly ranked in others [27]. Some implementations can be seen in [25,5]. Different features used for these methods are described in [28]. See [27] for more HMM and CRF approaches descriptions. Nowadays, models developed using deep neural networks architectures provide very competitive results. As a drawback, a big amount of training data has to be available. Many semi-supervised methods for NER in the general domain are reviewed in [20]. Unsupervised learning methods are typically based on clustering. Methods are usually based on lexical resources and on large corpus of statistics taken from unannotated texts. See [20] for a review.

The impact of feature engineering in order to improve the performance of different models, such as CRF, SVM or neural networks in the clinical NER task for Spanish, English and Swedish is reported in [33].

A NER system for Spanish electronic health reports with the goal to access their factuality with a NegEx[6] implementation has been presented in [24]. Different techniques are evaluated. Their best result consists in a CRF implementation, tested with 75 electronic health reports annotated with an IAA of 90.53%. As features they use four characters prefixes and suffixes and transform terms to lower case. They consider entities that overlap as partial match. Freeling-Med [21] is used in another study as a way to automatically tag named

[5] CRF, are defined in Section 3.3.

[6] Negex is the most popular system for detecting negations and their scope

entities. They test it with 20 clinical reports looking for diseases, drugs and substances. They achieve high F1s, but use following extremely loose matching criteria: “two elements are considered to be equivalent if an element given by the system is entirely contained within an extension of a manually tagged element by six positions both to the left and to the right”. A tool similar to UMLS MetaMap Transfer (MMTx)[7] has been presented in [7] for the identification of Spanish SNOMED CT[8] terms corresponding to the *procedures* and *disruptions* hierarchies in Spanish clinical notes. The tool is tested with 100 clinical notes. An inverted index is used and a score is assigned to the retrieved terms, depending on the length of the query with respect to the retrieved terms. It is integrated with MOSTAS [13], a tool that normalizes abbreviations and acronyms, anonymizes reports and corrects spelling errors. Table 1 shows the results for Spanish NER in the medical domain.

paper	**# reports**	IAA	**P**	**R**	**F1**	**doc. types**	**ent. types**
[7]	100	66%	0.43 (0.72*) 0.35 (0.70*)	0.06 (0.09*) 0.07 (0.55*)	0.11 (0.16*) 0.06 (0.10*)	CN	DRP (SN) PR (SN)
[24]	75	90.53%	0.36 (0.70*)	0.45 (0.83*)	0.40 (0.76*)	EHR	DS
[21]	20	-	(0.97**) (1.00**) (0.84**)	(0.80**) (0.96**) (0.92**)	(0.88**) (0.98**) (0.88**)	ClR	DS DR SB

Table 1. NER results for Spanish in the medical domain. References, type of documents: ClR: clinical reports, CN: clinical notes, EHR: electronic health reports. Entity types: DS: diseases, DRP: disruptions, DR: drugs, PR: procedures, SB: substances, and SN: SNOMED CT. Other references: doc.: documents, ent: entities. First results correspond to exact matches. Results marked with (*) correspond to lenient matches and (**) to extremely loose lenient matches.

3 Material & Methods

In this section we will explain the data used for training (when it applied) and for testing purposes, the preprocessing applied to reports, and the lexicons used. SiMREDA algorithm and the CRF algorithm and its feature selection are presented next. As previously mentioned, SiMREDA was thought as a solution for cases were there are no low or medium resources available (lexicons, corpora, software tools). CRF was thought as a relatively easy to implement solution for NER when annotated datasets are available. Then, we explain the exact match and a lenient matching evaluation metric used and how they work.

[7] `https://mmtx.nlm.nih.gov/MMTx/`

[8] `https://www.snomed.org/snomed-ct`

3.1 Data

We worked with 513 radiology reports of an Argentinian hospital that were anonymized and annotated by us. Reports are short, approximately 6% of AEs and FIs are written in an abbreviated way, it contains some non-sentences and lack of punctuation signs. Furthermore, many texts lack diacritics. For the sake of uniformity we decided to remove all of them. We also normalized our reports transforming every word to lowercase. Table 2 describes the composition of the dataset and Table 3 shows the number of AEs and FIs and the number of abbreviations and acronyms found in the annotated dataset. For details about the process followed, schema and elaborated guidelines to annotate the dataset refer to [11].

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

Table 2. Composition of the dataset.

type	**total**	**different**
anatomical entities	4,398	405
finding	2,637	745
abbreviations	880	105

Table 3. Type and amount of entities, modifiers and other characteristics in the annotated reports.

Both NER algorithms are going to be evaluated with the same dataset. We split the annotated dataset into the *development dataset* (80%) and the *testing dataset* (20%). For CRF we tested different features with 5-fold cross-validation in the development dataset. Both algorithms were tested with the testing dataset.

3.2 SiMREDA algorithm

We proposed and implemented SiMREDA, a Simple Entity Detection Algorithm for Medium Resources languages. The algorithm takes as input radiology reports and gives as output the same reports with the anatomical entities and clinical findings automatically tagged. It has three modules and some variants, as shown in Figure 1. The first module is its basic module. Modules 2 and 3 are additions. Next, we describe the three modules.

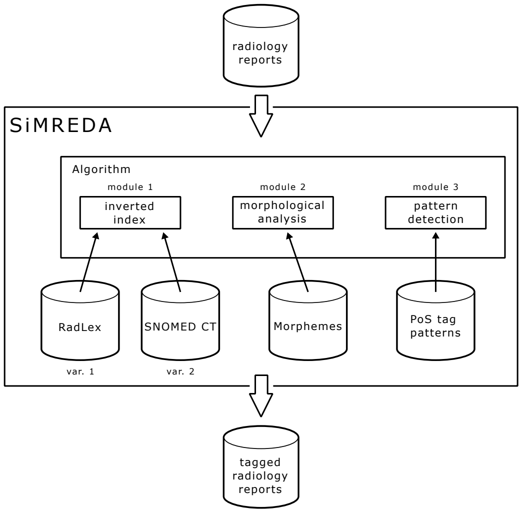

Fig. 1. Schema of SiMREDA algorithm. Its modules and variants.

Module 1: Inverted index. The module consists in a lookup of terms that come from a specialized vocabulary through the use of an inverted index. As specialized vocabulary we try two alternatives: RadLex[9] a vocabulary specific of the radiology domain, but that had to be translated into Spanish (variant 1) and SNOMED CT, that is not specific of the radiology domain, but exists in Spanish (variant 2). In variant 1, we translate to Spanish all RadLex AEs and FIs. Therefore, we use Google Translate (GT), enhanced through mappings with UMLS and Wikipedia. Also a subset of the translated terms were corrected by a physician of the radiology domain. Variant 2 takes SNOMED CORE Problem list subset as FIs and a subset of its *Body Structure* and *Substance categories* as AEs. No translation is needed. In variant 1 each word appearing in the translated terms is added to an inverted index. Stopwords are excluded. Each entry of the inverted index points to the RadLex terms where it appears and gets the most frequent class assigned (anatomical entity or clinical finding). The process with variant 2 is similar using SNOMED CT instead of RadLex. See Table 4 for an example of an inverted index of RadLex terms translated into Spanish.

Those words that appear in the reports and that also belong to the inverted index are tagged as anatomical entities or as findings, according to the class as-

[9] `https://www.rsna.org/RadLex.aspx`

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

Table 4. Example of inverted index for RadLex terms *heart*, *heart valve*, *ischemic heart disease*, *basal zone of the heart*, *aortic valve*, *mitral valve*, *mitral valve insufficiency*, *insufficiency fracture* and *heart failure* translated to Spanish. The first column has the indexed words. The second column has the RadLex terms, where the words occur, and the third column has the class assigned to the word, that depends on the class of the RadLex terms, where the word appears. The table should also have entries for the words *ischemic*, *disease*, *basal*, *zone*, *aortic*, *mitral*, *fracture* and *failure* (we do not add them because of space constraints).

signed in the inverted index. Adjacent sequence of words belonging to the same class are tagged together with their corresponding class. For example, lets assume we have following text: *se visualiza prolapso de la válvula mitral* (*a mitral valve prolapse has been noticed*). After running the algorithm that tags terms according to their presence in RadLex we would get: "se visualiza <FI>prolapso</FI>de la<AE>válvula</AE><AE>mitral </AE>" if we assume that *prolapso* appears in RadLex more times in terms referring to FIs than in terms referring to AEs and if we consider the class assigned to *válvula* in Table 4. Then, if there are contiguous words of the same class (in this case we have *válvula* and *mitral*, both tagged as anatomical entities) we tag them together with their corresponding class. In this case we would get: "Se visualiza <FI>prolapso</FI> de la <AE>válvula mitral</AE>". As a result, as the algorithm output, we have a set of radiology reports with terms referring to AEs and to FIs automatically tagged according to the translation to Spanish of RadLex anatomical and clinical finding terms.

Module 2: Morphological analysis. Graeco-Latin morphemes are used in medical terms of many languages, including Spanish. Even a small number of morphemes of Greek and Latin origin can generate a large amount of terms [15,32]. Therefore, their lookup can help discover clinical findings that do not appear in the lexicons, that are not correctly translated to Spanish or that are not well written in reports. Thus, the second module considers the appearance of those morphemes.

We implemented a simple module to detect Graeco-Latin morphemes. Therefore, we compiled a dictionary of morphemes, that includes their type -prefix or suffix- and meaning. The dictionary was built based on a reduced subset of [2]. Those words, that include morphemes corresponding to findings, in the correct position (as suffix or as prefix) are tagged as FIs replacing the tag assigned based on RadLex terms (Module 1). For example, *ascitis* -*ascites*- is not tagged as a finding based on RadLex, but our morpheme detection module detects the suffix *-itis*, so it assumes that *ascitis* is a FI and tags it as such.

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

Module 3: Pattern detection. Usually AEs and FIs satisfy certain PoS tagging patterns. For example, from 20% of our development dataset that we used to analyze the PoS tag sequences of the annotated anatomical entities and clinical findings, we discovered that many of the anatomical terms beginning with a noun continue with an adjective, that is also considered part of the AE (e.g. *testículo izquierdo* -*left testicle*- and *pared abdominal* -*abdominal wall*-, both nouns followed by adjectives). In many cases only the noun is tagged by our Module 1. We analyze the PoS tag patterns present in the previously mentioned subset of our development dataset and look for these patterns in the radiology reports in order to improve SiMREDA results, expanding the named entities to the adjectives (in this example). So, "[*testículo*](AE) *izquierdo*" is expanded to "[*tesículo izquierdo*](AE)". This constitutes module 3. Tables 5 and 6 show the most frequent PoS tagging sequences of anatomical entities and clinical findings appearing in the selected subset of the development dataset. In these tables, columns 3 and 4 show the percentage of annotated entities that have the pattern listed in column 1 and the accumulated percentage. The last column shows the probability that a sequence that has the PoS tags analyzed in the row and whose first word is tagged as an AE (Table 5) or a FI (Table 6) is an AE or a FI respectively.

3.3 Conditional Random Fields (CRF)

Conditional random fields are probabilistic models used to predict sequences of labels based on sequences of input samples. A text can be seen as a sequence of tokens. We can say that each token has an associated vector of features, such as the word's part of speech tag, the word's suffix of a given length and an

PoS tag sequence	**example**	**perc. (%)**	**acum. perc.(%)**	**prob. of being AE**
NC	bazo (spleen)	75.88	75.88	1.00
NC-AQ	músculo pilórico (pyloric muscle)	17.31	93.18	0.76
NC-NC	venas porta (portal veins)	1.66	94.84	0.63

Table 5. Detected anatomical entity patterns.

PoS tag sequence	**examples**	**perc. (%)**	**acum. percent-age (%)**	**prob. of being FI**
NC	ovariocele (ovariocele)	35.65	35.65	1.00
VMP	dilatadas (dilated)	14.57	50.22	1.00
VMI-AQ	liquido libre (free fluid)	12.61	62.83	1.00
NC-AQ	hipertrofia pilrica (pyloric stenosis)	9.57	72.4	0.81
VMP-SP-NC	aumentada de tamao (increased in size)	4.57	76.97	0.92
AQ	bfida (bifid)	2.39	79.36	1.00
NC-SP-DA-NC	incremento de la vascularizacion (increase in vascularization)	1.96	81.32	0.60

Table 6. Detected finding patterns. *liquido* is tagged as a verb, while it should be a noun (this happens because the accent is missing, *lquido* is the correct word).

indication as to whether the word is capitalized or not. The input of CRF is the sequence of tokens of the text. The features of a token and the pattern of labels assigned to previous words are used to determine the most likely label for the current token. In linear chain CRF only the label of the previous token is used. As mentioned in a previous section, CRF have been successfully used for NER and also for some other natural language processing tasks, such as PoS tagging.

We tried different set of features, some provided in different NER tasks and a set of features proposed by ourselves. We used our development dataset in order to decide the best set of features. Once we decided which to use, we used the whole development dataset as training set, and we tested the results with our testing dataset. The best set of feature is one proposed in [14] for the solution of CLEF eHealth 2015 task 1b (NER for French). Nevertheless, the relative difference among its F1 and those of our proposed feature set is very low (0.57 % relative improvement). We selected this set of features, that includes: lexical (lower case), morphological (four characters prefix and four characters suffix), reduced PoS tags, orthographic features and shape-related features (length of the token, whether the token begins with a capital letter, whether all its characters are capital letters, whether it contains only digits, only letters or letters and digits). It also takes into account context for morphology features and for PoS tags.

3.4 Evaluation of results

We will measure our algorithms with the classical exact match metrics (precision (P), recall (R) and F1) and with a lenient (or approximate) match metric, based on the MUC challenge evaluation metric and that scores partial matches (matches with wrong boundary and same entity type) as half of an exact match. Metrics are explained in detail in [9] and in the *Scoring Software User's Manual*.[10]

4 Results

In this section we present SiMREDA and CRF algorithms results, that can be seen in Table 7.

Precision, recall and F1 measure were calculated against every entity type (AE and FI) and a final overall score, that considers both entity types is also given for all the measurements. Similarly, precision, recall and F1 for partial boundary matching is calculated for every entity type (AEPM and FIPM) and a final overall score (totalPM) is calculated. In both cases we used the testing dataset composed by 20% of the annotated 513 reports (103 reports).

We are interested in a solution that retrieves a high rate of relevant entities and that the entities retrieved by the solution are actually positive (high recall and high precision). Hence, we will choose F1 metric, that balances precision and recall, as the metric in order to compare results.

Table 8 presents the Graeco-Latin morphemes related to findings, that were discovered in the testing dataset.

5 Analysis of results

Regarding SiMREDA, results are lower than CRF. Even though, they are better than results for NER detection in Spanish clinical texts presented in Table 1. SiMREDAs results are similar to best CLEF 2015 and 2016 results (for MEDLINE articles written in French). Nevertheless, results are difficult to compare since the definition of named entities differ, in some of the cases the languages are different, and also lenient match definitions differ.

[10] The Message Understanding Conference Scoring Software User's Manual.`https://www-nlpir.nist.gov/related_projects/muc/muc_sw/muc_sw_manual.html`, accessed June 2017.

	SiMREDA compared to CRF					
	SiMREDA			**CRF**		
NE	**P (%)**	**R (%)**	**F1 (%)**	**P (%)**	**R (%)**	**F1 (%)**
AE	58.74	73.25	65.20	92.09	91.56	91.82
FI	56.21	48.68	52.17	85.78	74.95	80.00
total	58.00	64.10	60.90	89.68	84.70	87.12
AEPM	65.46	77.23	70.86	95.00	92.61	93.79
FIPM	59.86	54.49	57.05	88.46	80.06	84.05
totalPM	63.73	68.9	66.21	92.42	87.45	89.87

Table 7. SiMREDA implementation compared to CRF implementation. Exact and partial match results are shown for each entity type and the overall measure (total) is also shown.

morpheme	**category**	**number of appearances (distinct)**
-itis	finding	5 (2)
-megalia	finding	18 (3)
-osis	finding	8 (4)

Table 8. Morphemes related with medical terms appearing in the test set.

Other results, not shown in Table 7 are described next. The use of translated RadLex AEs and FIs had better results than the use of SNOMED CT clinical findings and anatomical entities. This is probably because the second vocabulary has many terms that do not belong to the radiology domain, which decreases SiMREDA's precision. Furthermore, it does not contain terms specific of the radiology domain (as RadLex does), which decreases recall.

The improvement of only 10% of RadLex translations derived in a relative F1 increase of anatomical entities and findings of ∼ 7% and ∼ 4% respectively. Therefore, we conclude that it makes sense to invest effort in improving the translations.

Only 31 findings with Graeco-Latin suffixes appear in our testing dataset. Therefore, the addition of Module 2 does not improve the results in a very noticeable way (overall F1 increase of less than 1%). However, the detection of morphemes related to the medical domain helped us to detect terms that are misspelled. For example, *etenosis* for *estenosis* -*stenosis*- were found in reports and detected as findings by Module 2.

As expected, partial match results are always higher than exact match results. For example, as reported in Table 7, the overall SiMREDAs F1 is 60.90 with exact match. Partial match achieves a relative increase in F1 of 8.72%.

Also, findings show a greater increase in partial match F1s than AEs. We believe this is motivated, because it is much more complex to determine the boundaries of a FI than those of an AE. This issue was also reported during annotation. Furthermore, in our dataset findings have longer terms (in amount of words composing them) than anatomical entities, which makes its boundary detection a harder problem.

Some errors are due to following causes:

- tokenization problems: the text *(...)ascitis-* appeared in one of the reports. The tokenizer did not separate the word *ascitis* from the symbol -, so ascitis was not recognized by our algorithm as a finding
- annotation criteria (a decision was taken to annotate implants, such as *kidney implant* as an AE. The algorithm does not annotate implant as part of an anatomical entity. Also, for example, *ovarian cyst* should be annotated as [ovarian cyst](FI), while the algorithm detects [ovarian](AE) [cyst](FI)).
- annotation inconsistencies: there is a number of errors and inconsistencies in the annotations. Some of them, like the omission of annotation of entities (such as *bile duct* and *dilated*), the incorrect classification of entities (such as *gallbladder* as FI) erroneously worsens the results. The annotation of entities with wrong boundaries explains, in part, the difference of performance among the exact match and the partial match.

Our CRF results outperform others obtained with the same feature set for French [14] (the original proposal of the feature set) and for German [23].[11] Since all results are tested with different genre of data and in different languages it is not easy to draw a conclusion about the differences in the results. In Spanish and in French anatomical entities have a higher F1 than findings. That is what usually happens. It can be also noticed that results with our Spanish dataset are better in both entity types than in the original French implementation. This might have to do with the fact that our corpus is of a restricted domain -only radiology reports, while the French implementation has EMEA and MEDLINE articles-, that in our case we had two entity types, while the other case had to select among 10 entity types, and that we trained with 410 reports and tested with 103, while in the French case, 836 MEDLINE titles and 4 EMEA documents were used for training and 832 MEDLINE titles and 12 EMEA documents were used for testing. Besides, the definition of AE and FI among both systems does not necessarily coincide.

As can be seen in Table 7, as expected, CRF outperforms SiMREDA for exact as well as for partial match. Both methods require manually created resources: SiMREDA a lexicon and the elaboration of rules and CRF an annotated corpus. The CRF algorithm is much better, but SiMREDA is adequate when there are few resources available for annotation.

Concluding, the development of a dictionary-based algorithm enhanced with rules is more laborious than a ML approach such as CRF. In cases as ours, where there do not exist specific resources for the radiology domain in Spanish

[11] We consider that AE and FIs in the French dataset are *anatomy* and *disorders* hierarchies of UMLS. In the case of for German, what we consider AEs corresponds to *organs* and what we consider FI corresponds to *symptoms*, *diagnoses* and *observations*.

it is even more difficult. Nevertheless, this method has the advantage of needing few annotated data. Based on the good perspective of CRFs results, feature engineering can be carried out in order to improve results.

6 Conclusions and Future Work

In this paper we presented SiMREDA, a dictionary-based entity recognition algorithm, enhanced with morphology analysis and with a post-processing based on the analysis of PoS patterns of the entities of interest, and an algorithm based on CRF. SiMREDA approach can be used when there are no datasets annotated for implementing ML techniques and when there are no lexicons in the language of the reports. From the results obtained and the analysis carried out we can draw following conclusions.

Despite the conclusion about the coverage of SNOMED CT terms in the radiology domain obtained in [1],[12] we obtained better results with SiMREDA using a translated version of RadLex -although it is not a high-quality translation- than with SNOMED CT terms that are already in Spanish.

Based on results obtained comparing the original GT translation and a correction of a portion of it by a physician of the radiology domain, we can conclude that our algorithm is sensitive to a poor translation. The improvement of only 10% of RadLex translations improves our results. Therefore, we conclude that it makes sense to invest effort in improving the translation.

The rules added to SiMREDA in Module 3, based on the analysis of its PoS tagging patterns improved the results. It also could be noticed that the morphological processing improvement is almost imperceptible, but we can appreciate that it recognizes more AEs and FIs and that the limited increase in performance is probably due to the reduced size of the test set. We could also see that the morphological module helped in recognizing misspelled entities.

In this paper we only show the final results. But, lenient match draws better results than the exact matching for every entity type across all settings of both algorithms tested. Besides, in this use case it is more important to determine if an entity is present than to correctly determine its boundaries. Therefore, we conclude that it is important to report a precisely defined partial metric accompanying the exact match results.

We can also conclude that despite having a small annotated dataset (513 reports -see Tables 2 and 3-), we could successfully apply ML.

There are many studies than can be carried out as future work. There are some phrases, we call *prefix terms*,[13] such as "could suggest", "is visualized", that usually determine that the following noun phrase corresponds to a clinical finding. Detecting those phrases and the noun phrases that come after them, could help improve the recall of retrieved findings.

[12] The paper is not available online. Results were discussed in a personal communication.

[13] In Spanish they usually occur before the terms of interest.

The construction of abbreviation databases for Spanish radiology reports, would be probably less useful than others existing for English ([34,19]), since many of the abbreviations used in these kinds of reports do not follow naming conventions and would, therefore, be difficult to generalize to other texts. However, the subject could be studied and an abbreviation database could be constructed. Therefore, previous efforts could be studied [16].[14]

It would be interesting to detect of all the morphemes composing a word, as [32] carried out. This can help to a better understanding of the words. For instance, words that have more than one morpheme related with the medical domain (e.g. **peri**ton**itis** -peritonitis-) can be found, and their semantics can be better comprehended. Consider also *cardiopatía* -cardiopathy- and *linfoadenopatía* -lymphadenopathy-, whose decomposition into morphemes (cardio-patía and linf-o-adeno-patía) explains in which anatomical entity the findings have occurred.

There are some patterns that would also probably help to improve finding retrievals. Consider:

- AE FI, as in [ovarian](AE) [cyst](FI),
- FI AE,
- and FI (en (el |la(s?) |los|λ) |de (la(s?) |los |λ) |del) AE,[15][16] as in "[luxación](FI) de la [cadera](AE)" (hip dislocation).

With the current version of SiMREDA, these patterns are not considered as findings, but they were annotated as findings. An additional SiMREDA module that detects those patterns as entities could be constructed. It is also important to notice that detecting [ovarian](AE) [cyst](FI) as a first step, has as advantage, that it can be determined where the finding is located. If [ovarian cyst](FI) would have been detected, then this understanding would be lost.

Finally, a deep learning architecture could be implemented to improve CRFs results. Character based convolutional neural networks (CNN), recurrent neural networks (probably biLSTM), and CRF could be considered as layers. The non-existence of sufficient data to train word embeddings in this particular domain and language, might make them not very beneficial in this particular case.

References

1. Aleksovski, Z.: Testing RadLex for completeness using large database of radiology reports. In: Society for Imaging Informatics in Medicine. Annual Meeting (2014)
2. na Ambulódegui, E.S.: Manual de Terminología Médica N 2 (2012)

[14] Acronyms and abbreviations provided by the National Academy of Medicine of Colombia `http://dic.idiomamedico.net/Siglas_y_abreviaturas` and by the Spanish Ministry of Health `http://www.redsamid.net/archivos/201612/diccionario-de-siglas-medicas.pdf?0`.

[15] in |in the |from

[16] Written as a regular expression.

3. Ananiadou, S., Friedman, C., Tsujii, J.: Introduction: named entity recognition in biomedicine. Journal of Biomedical Informatics **37**(6), 393–395 (2004)
4. Basaldella, M., Furrer, L., Tasso, C., Rinaldi, F.: Entity recognition in the biomedical domain using a hybrid approach. Journal of biomedical semantics **8**(1), 51 (2017)
5. Batista-Navarro, R.T., Rak, R., Ananiadou, S.: Chemistry-specific features and heuristics for developing a CRF-based chemical named entity recogniser. In: Proceedings of the Fourth BioCreative Challenge Evaluation Workshop. vol. 2, pp. 55–59. Citeseer (2013)
6. Cascade, P.N., Berlin, L.: Malpractice issues in radiology. AJR Am J Roentgenol. **173**(6), 1439–1442 (1999)
7. Castro, E., Iglesias, A., Martínez, P., Castaño, L.: Automatic identification of biomedical concepts in Spanish-language unstructured clinical texts. In: Proceedings of the 1st ACM International Health Informatics Symposium. pp. 751–757. ACM (2010)
8. Chapman, W.W., Cohen, K.B.: Current issues in biomedical text mining and natural language processing. Journal of Biomedical Informatics **42**(5), 757–759 (2009)
9. Chinchor, N., Hirschman, L., Lewis, D.D.: Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3). Association for Computational Linguistics **19**(3)
10. Cohen, A.M., Hersh, W.R.: A survey of current work in biomedical text mining. Briefings in Bioinformatics **6**(1), 57–71 (2005)
11. Cotik, V., Filippo, D., Roller, R., Uszkoreit, H., Xu, F.: Annotation of Entities and Relations in Spanish Radiology Reports. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. pp. 177–184 (2017)
12. Do, B., Wu, A., Maley, J., S., Biswal: Automatic retrieval of bone fracture knowledge using natural language processing. J Digit Imaging **26**(4), 709–713 (2013)
13. Iglesias, A., Castro, E., Pérez, R., Castaño, L., Martínez, P., Gómez-Pérez, J.M., Kohler, S., Melero, R.: Mostas: Un etiquetador morfo-semántico, anonimizador y corrector de historiales clínicos. Procesamiento del lenguaje Natural **41** (2008)
14. Jiang, J., Guan, Y., Zhao, C.: WI-ENRE in CLEF eHealth Evaluation Lab 2015: Clinical Named Entity Recognition Based on CRF. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum. Toulouse, France (2015)
15. José María López Piñero, M.L.T.: Introducción a la terminología médica. Masson S.A. (2005)
16. Laguna, J.Y.: Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias
17. Lakhani, P., Langlotz, C.P.: Automated detection of radiology reports that document non-routine communication of critical or significant results. J Digit Imaging **23**(6), 647–57 (2009)
18. Leaman, R., Gonzalez, G.: BANNER: An executable survey of advances in biomedical named entity recognition. In: Proceedings of the Pacific Symposium on Biocomputing. vol. 13, pp. 652–663 (2008)
19. Moon, S., Pakhomov, S.V.S., Liu, N., Ryan, J.O., Melton, G.B.: A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. JAMIA **21**(2), 299–307 (2014)
20. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Linguisticae Investigationes **1**(30), 3–26 (2007). https://doi.org/10.1075/li.30.1.03nad

21. Oronoz, M., Casillas, A., Gojenola, K., Perez, A.: Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names. Lecture Notes in Computer Science, 8259. Progress in Pattern Recognition, ImageAnalysis, ComputerVision, and Applications 18th Iberoamerican Congress, CIARP 2013 pp. 36–543 (2013)
22. Poibeau, T., Kosseim, L.: Proper name extraction from non-journalistic texts. In: Computational Linguistics in the Netherlands 2000, Selected Papers from the Eleventh CLIN Meeting, Tilburg, November 3, 2000. pp. 144–157 (2000)
23. Roller, R., Rethmeier, N., Thomas, P., Hbner, M., Uszkoreit, H., Staeck, O., Budde, K., Halleck, F., Schmidt, D.: Detecting Named Entities and Relations in German Clinical Reports. In: Lecture Notes in Computer Science. vol. 10713, pp. 115–124 (2018)
24. Santiso, S., Casillas, A., Pérez, A., Oronoz, M.: Medical entity recognition and negation extraction: Assessment of NegEx on health records in Spanish. In: Bioinformatics and Biomedical Engineering - 5th International Work-Conference, IWBBIO 2017, Granada, Spain, April 26-28, 2017, Proceedings, Part I. pp. 177–188 (2017)
25. Settles, B.: Biomedical named entity recognition using conditional random fields and rich feature sets. In: Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP). COLING-04, Association for Computational Linguistics, Stroudsburg, PA, USA (2004)
26. Shen, D., Zhang, J., Zhou, G., Su, J., Tan, C.L.: Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain. In: Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13. pp. 49–56. Association for Computational Linguistics (2003)
27. Simpson, M.S., Demner-Fushman, D.: Mining Text Data, chap. 14. Biomedical Text Mining: A Survey of Recent Progress. Springer (2012)
28. Sondhi, P.: A survey on named entity extraction in the biomedical domain (2008)
29. Takeuchi, K., Collier, N.: Bio-medical entity extraction using support vector machines. Artificial Intelligence in Medicine **33**(2), 125–137 (2005)
30. Tasneem, A., B, A.: A survey on biomedical named entity extraction. Asian Journal of Engineering and Technology Innovation **4**(7), 25–28 (2016)
31. Uzuner, Ö., Solti, I., Cadag, E.: Extracting medication information from clinical text. Journal of the American Medical Informatics Association **17**(5), 514–518 (2010)
32. Vivaldi, R.E.J., Cabré, M.T.: Use of Greek and Latin forms for term detection. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2000) **78**, 855–859 (2000)
33. Weegar, R., Casillas, A., de Ilarraza, A.D., Oronoz, M., Prez, A., Gojenola, K.: The impact of simple feature engineering in multilingual medical NER. Proceedings of the Clinical Natural Language Processing Workshop pp. 1–6 (2016)
34. Xu, H., Stetson, P.D., Friedman, C.: A Study of Abbreviations in Clinical Notes. In: AMIA 2007, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 10-14, 2007 (2007)