Syntactic methods for negation detection in radiology reports in Spanish

Viviana Cotik[1] and **Vanesa Stricker**[1] and **Jorge Vivaldi**[2] and **Horacio Rodriguez**[3]
[1]Departamento de Computación, FCEyN, UBA, Argentina, {vcotik, vstricker}@dc.uba.ar
[2]Universitat Pompeu Fabra, UPF, Barcelona, Spain, jorge.vivaldi@upf.edu
[3]Universitat Politècnica de Catalunya, UPC, Barcelona, Spain, horacio@lsi.upc.edu

Abstract

Identification of the certainty of events is an important text mining problem. In particular, biomedical texts report medical conditions or findings that might be factual, hedged or negated. Identification of negation and its scope over a term of interest determines whether a finding is reported and is a challenging task. Not much work has been performed for Spanish in this domain.

In this work we introduce different algorithms developed to determine if a term of interest is under the scope of negation in radiology reports written in Spanish. The methods include syntactic techniques based in rules derived from PoS tagging patterns, constituent tree patterns and dependency tree patterns, and an adaption of NegEx, a well known rule-based negation detection algorithm (Chapman et al., 2001a). All methods outperform a simple dictionary lookup algorithm developed as baseline. NegEx and the PoS tagging pattern method obtain the best results with 0.92 F1.

1 Introduction

Text mining and natural language processing (NLP) techniques have been applied to the biomedical domain for a long time. Automatic identification of relevant terms in medical reports is a preliminary step for indexing and for search tools and it is useful for clinical, educational and research purposes.

A clinical condition mentioned in a biomedical text does not necessarily mean that a factual condition is reported, since the term or terms referring to the condition could be under the scope of negation or epistemic modality markers (hedges). For example, in *"no lymphadenopathies were detected"*, *"no ... were detected"* indicates that the medical condition (*"lymphadenopathy"*) is negated.

We refer to language constructions that denote negations as *negations* or *triggers* and to medical conditions and observations made about a particular illness in medical examinations as *findings* or *terms of interest*.

According to (Chapman et al., 2001b), many of the medical conditions described in unstructured texts in medical health records are negated. For this reason, the detection of negations in texts of the biomedical domain is an important task in the field of NLP, called BioNLP. Scope of negation has also received attention in other domains (Wiegand et al., 2010; Potts, 2011; Wor, 2010).

In this work we implement five techniques: 1) a simple approach, used as baseline, that determines if a finding is negated based on the presence of a negation term and a finding in the same sentence. The negation term is detected by dictionarylookup of negation terms; 2) an adaptation of NegEx to Spanish; the use of negation rules that were created based on 3) PoS tagging patterns, 4) constituent tree patterns, and 5) dependency tree patterns. Our goal is to decide which of the implemented methods is the best to automatically detect negations of important findings tagged in radiology reports written in Spanish.

Our methods are applied to Spanish and to a particular domain: radiology. This domain (and particularly our dataset) has the characteristic of having short reports, with usually short sentences, using informal language, containing non-standard abbreviations, and with highly noisy text. As far as we know, of our methods only NegEx has been implemented for Spanish and our implementation obtains better results. Using a Spanish dataset presents some challenges: we had to build a cor-

Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 156–165,
Berlin, Germany, August 12, 2016. ©2016 Association for Computational Linguistics

pus and annotate it, syntactic parsing tools are less developed for languages other than English, and translations needed for the development of the work incorporates errors.

Experiments were performed over a dataset prepared from a set of ultrasonography reports written in Spanish, that have been previously tagged automatically with a tool based on RadLex[1], a specific radiology lexicon. A fragment of a tagged ultrasonography report in Spanish and its translation to English can be seen below: *"Pancreas: tamano y ecoestructura normal. Retroperitoneo vascular: sin <finding>alteraciones</finding>. No se detectaron <finding>adenomegalias </finding>. (...)"("Pancreas: normal size and echotexture. Vascular retroperitoneum: without <finding> changes </ finding>. No <finding> lymphadenopathies </finding> were detected.(...) "*).

The rest of the paper is organized as follows. Section 2 presents previous work in the detection of negation terms in the medical domain. In Section 3 we present our main contributions, by explaining the methods and datasets used. Section 4 shows the results of evaluating each of the algorithms with the testing dataset. Finally, Discussions, Conclusion and Future Work are presented.

2 Previous work

The use of information retrieval techniques for automatically indexing narrative medical reports and creating terminological resources has been present at least since mid-late 90s (Aronson et al., 1994; Rindflesch and Aronson, 1994; Sundaram, 1996).

In order to determine if a finding mentioned in a narrative medical report is under the scope of negation, (Chapman et al., 2001a) developed NegEx, a simple algorithm based on regular expressions that obtained very good results for English. Several methods were built upon this simple algorithm. (Wu et al., 2011) developed a word-based radiology report search engine based in a modification of NegEx. (Harkema et al., 2009) developed ConText, based in NegEx, employing a different definition for the scope of triggers and adopting it to different type of medical reports. NegEx has been adapted to Swedish (Skeppstedt, 2011), French (Deléger and Grouin, 2012), Dutch (Afzal et al., 2014), and Spanish for clinical records written in that language (Costumero et al., 2014) and radiology reports (Stricker et al., 2015). The NegEx lexicon has been extended for Swedish, French and German (Chapman et al., 2013).

Syntactic methods have also been used. (Huang and Lowe, 2007) construct manually grammar rules using Part of Speech tagging in order to detect negations in radiology reports. (Uzuner et al., 2009) compare a NegEx extension with a machine learning technique that uses lexical and syntactic information using two corpora of discharge summaries and one of radiology reports. (Mehrabi et al., 2015) use dependency parsing to reduce NegEx False Positives. (Sohn et al., 2012) applies techniques of dependency parsing to detect negations. Therefore he compiles negation rules derived from the dependency paths.

Finally, machine learning techniques are also used for the negation detection task. (Cruz Díaz et al., 2010) compare these techniques to a regular expression-based method. (Morante and Daelemans, 2009) use them in order to establish the scope of negation in biomedical texts. (Rokach et al., 2008) perform automatic negation identification in clinical reports by means of extracting automatically regular expressions and patterns from annotated data and using them to create a learning method.

Several challenges have been performed on this topic. CoNLL 2010 Shared Task: Learning to Detect Hedges and Their Scope in Natural Language Text (Farkas et al., 2010), 2010 i2b2 NLP challenge, focused on the negation and uncertainty identification (Uzuner et al., 2011) and SEM 2012 Shared Task: Resolving the Scope and Focus of Negation(Morante and Blanco, 2012).

3 Methods

In this section we introduce the different methods developed to detect negations in radiology reports written in Spanish. The idea underlying syntactic techniques is to identify patterns of negations, manually compile negation rules, and use them to determine if a finding is under the scope of a negation or not. These methods used rules that were elaborated based on: 1) PoS tag patterns, 2) constituent tree (or shallow parsing) patterns of of the sentences and 3) dependency tree patterns (paths obtained from the dependency parsing of sentences). Rules were evaluated with the testing dataset.

Our methods only take into account the sen-

[1] http://www.radlex.org/

tence where the term of interest appears in order to determine whether it is negated or not, i.e. it does not use information of other sentences.

3.1 *Dictionary lookup* algorithm

This simple algorithm developed is based on the lookup in the text of a list of negations marked by the expert radiologist as usual negation terms used in radiology reports. Sentences containing tagged findings where a negation appears (in any order) are tagged as Negated, and those with findings and without negations are tagged as Affirmed. This algorithm will be used as baseline.

3.2 The NegEx algorithm

NegEx algorithm for negation detection takes as input medical records with tagged findings and looks for phrases (triggers) that are mostly used to denote negation, for example *"no signs of"*. It checks if the phrase is applied to negate the finding or disease using rules that take into account the distance among the finding and the negation phrase.

The set of triggers provided by the NegEx tool[2] was translated using automatic translation[3] (since translation is an expensive task and we are not experts in the domain) and revised by two non-domain experts. Those triggers that were not correctly translated were eliminated or corrected. Given that English lacks grammatical gender, while Spanish has two (male and female), additional trigger instances were generated due to inflectional properties (for example from *"no"* to *"ningún" "ninguna"*). NegEx triggers are divided into: *pseudo negation phrases*, *negation terms*, *termination terms* and *conjunction terms*. A label is used to classify each trigger in one of these groups. Triggers were classified according to their use.

This implementation differs from others (Costumero et al., 2014; Stricker et al., 2015) (and is part of our contribution) mainly in that:

- tests were performed with two different trigger sets: 1) NegEx translated triggers (described in previous paragraph). A total of 210 translated triggers were obtained. 2) triggers obtained by combining translated triggers, a set of bi and trigrams[4], and a list of triggers provided by a physician expert in the radiology domain (a total of 350 triggers),
- some end of scope triggers were added,
- coordinated negations, that were not taken into account in the English, nor in the Spanish versions were included as a trigger (*ni* - *nor*-) and NegEx algorithm was modified to include this term.

3.3 POS tagging patterns

Tags were assigned to each word of the sentence in order to determine the Part of Speech with the use of Freeling analyzer (Carreras et al., 2004). A small set of sentences were used to define negation patterns based on PoS tags. Patterns defined were:

- no +...+ verb + ...+ <*finding*>
- no +...+ <*finding*>
- sin +...+ <*finding*>
- sin +...+ <*finding*> +...+ ni +...+ <*finding*>
- no +...+ <*finding*> +...+ ni +...+ <*finding*>
- no +...+ verb +...+ <*finding*> +...+ ni +...+ <*finding*>

where "..." denotes zero or more words. The algorithm looks for these patterns in PoS tagged sentences. If a pattern occurs, the sentence is labeled as *Negated* indicating that the finding is under the scope of negation. For example: For "no se detectaron adenomegalias" we would have "RN P00CN000 VMIS3P0 FINDING", that satisfies the pattern "no +...+ verb + ...+<*finding*>".

RN represents "no". The words "sin" (without) and "ni" (nor) do not have specific negation tags (they are tagged as preposition and conjunction). That is why we look for these words directly in the text, instead of looking for some specific tag that represents them.

[2]https://code.google.com/p/negex/.

[3]Google Translate https://translate.google.com/

[4]Bi and trigrams were obtained from the 85600 report dataset (see Data subsection). Those, whose first word was *no*, were selected and the resulting were manually analyzed in order to discard those that did not correspond to triggers. 94 triggers were obtained.

3.4 Constituent tree patterns

Shallow parsing identifies the constituents of a sentence. We use this technique to manually elaborate patterns based on the phrase constituents avoiding the use of word distance to determine negation scope. Following patterns were used (patterns and phrase constituents[5] are shown):

1) no... verb... <finding>

```
    /   neg
S   -   grup-verb
    \   sn→grup-nom-mp→ w-ms→<finding>
```

2) without <finding>

```
S   -   grup-sp-> prep->sin
    \   sn→grup-nom-mp→w-mp →<finding>
```

3) no <finding>

```
S   -   neg
    \   sn→grup-nom-mp→w-mp →<finding>
```

4) no ... verb ... <finding> nor ... <finding>

```
    /   neg
    /   grup-verb
S   –   sn →grup-nom-mp→w-mp →<finding>
    \   coord
    \   sn→grup-nom-mp→w-mp →<finding>
```

5) <finding>: no

```
    /   sn →grup-nom-ms→w-ms →<finding>
S   -   no-c→:
    \   neg
```

Three steps were performed to obtain patterns from the constituent tree: 1) the finding is replaced by ”finding” and using Freeling the shallow parsing tree is obtained. 2) the tree structure is represented in an array. 3) the array is used to check whether the sentence satisfies one of the patterns previously discovered. For example, in order to check if a sentence satisfies pattern 1, it is verified if node with label S has as children a node with label *neg*, a node with label *grup-verb* and a node with label *sn* (in this order), and if node with label *sn* has as child a node with label *grup-nom-ms*, which also has as child a node with label *w-ms* and this has as child the node with content *finding*.

[5] *neg* stands for ”no”, *grup-verb* for ”verbal syntagma”, *sn* for nominal syntagma. See https://github.com/iknow/FreeLing/blob/master/doc/grammars/esCHUNKtags for further references.

3.5 Dependency tree patterns

Dependency parsing allows us to know the syntactic structure of a phrase. The method is based on syntactic context and does not take into account word distance to determine negation scope. Negation patterns are manually created based on syntactic dependency paths in the following way:

1. a small set of sentences containing all known type of negations (no, ni, sin) (no, nor, without) were parsed with a MATE dependency parser (Bohnet et al., 2013)[6]. A parse tree was obtained for each sentence (see Fig. 1),

2. negation terms were located automatically and an algorithm was developed in order to retrieve the path in the dependency tree between the negation term and the *finding* previously tagged,

3. paths were analyzed and a set of patterns that imply negation of findings was manually developed, and

4. patterns obtained in the previous step were tested with the testing dataset.

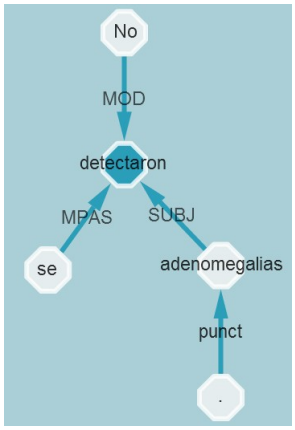

Figure 1: Example of a dependency parser tree for a sentence of the form of Pattern 1 (P1).

Patterns detected were following:

[6] The model was obtained as indicated in (Arias et al., 2014).

- P1: sentences of the form "**no** se detectaron adenomegalias" (The Spanish structure of this particular sentence corresponds to NEG (no) verb finding). The negation has a dependency relation with a word that the finding depends on.

- P2: sentences of the form "retroperitoneo vascular: **sin** alteraciones" (vascular retroperitoneum: without alterations) (anatomical part: NEG (sin) <*finding*>). The finding depends of "sin".

- P3: sentences like "via biliar **no** dilatada" (*bile duct not delated*) (anatomical part NEG <*finding*>, where NEG is "no").

- P4: sentences of the form "**No** se detectaron colecciones **ni** liquido libre" (neither collections nor free liquid has been detected) (NEG(**no**) verb <*finding*> NEG(**ni**) <*finding*>).

Data

Two datasets were used. The *analysis dataset* to infer the patterns of each of the proposed methods used and the *test dataset* to test the methods and compare their results.

Our original dataset is composed of about 85600 reports of ultrasonography studies performed in a public hospital. Reports are written in Spanish in non-structured format. They are brief (approximately five lines each) and they state what was found in the study performed on the patient. Text is noisy, characterized by frequent typos, abbreviations, sentences which are not syntactically well-formed and there is lack of punctuation in some cases.

The process to obtain both datasets was the following: An algorithm was used in order to automatically detect terms of interest (findings in the radiology domain) in the reports (Cotik et al., 2015). Then, a sentence tokenization was performed using NLTK (Loper and Bird, 2002). Only sentences with findings are selected (randomly) to create analysis and testing datasets. Finally, those sentences were annotated as containing negation with scope over the term of interest (Negated) or not (Affirmed). For the creation of the testing dataset a set of sentences were randomly selected and the following steps were performed: 1) we verified manually that sentences were neither the same (among them) nor very similar, 2) segmentation issues -e.g. different sentences that were not separated by the tokenizer- were corrected, 3) sentences with findings tagged by the algorithm and that were not considered actual findings by the annotators were eliminated and replaced by new ones. The analysis set is composed of 979 sentences and the testing set of 1000 sentences.

Findings detection

There are various inventories that serve as a basis to detect relevant terms in medical reports. Some of them are ICD10[7], a standard diagnostic terminology for epidemiology, health management and clinical purposes; and SNOMED CT[8], a clinical health terminology ontology -all of them included in UMLS (Unified Medical Language System)[9] Metathesaurus-; and RadLex[10], a lexicon centered only on radiology terms. SNOMED CT and ICD-10 are available in Spanish, RadLex is only available in English and in German. Previous implementations vary the type of inventory used to detect terms (UMLS, adaptations of ICD-10 and MeSH[11], among others). The information extraction algorithm we used to detect findings is based on the appearance of RadLex *pathological terms* in the reports. RadLex was chosen because it is the only lexicon specifically developed for the radiology domain, which is the domain under study. It has the disadvantage that no Spanish version has been developed, so it had to be translated from English. The translation is not an easy task, since, particularly, in the medical domain, there exist terms that are used differently in Spanish and in English.

Annotations

Working with languages different than English has, among others, the difficulty of the lack of data and tools. In this case we do not have a Gold Standard for validating the reliability of the new model. Annotating is an expensive task, and domain experts are not always available. The datasets build had to be annotated. The analysis dataset was annotated by two non-experts and the testing dataset

[7] http://apps.who.int/classifications/icd10/browse/2016/en

[8] http://www.ihtsdo.org/snomed-ct

[9] http://www.nlm.nih.gov/research/umls/. UMLS is a set of files and software that bring together many health and biomedical vocabularies and standards to enable interoperability between computer systems.

[10] http://rsna.org/RadLex.aspx

[11] http://www.ncbi.nlm.nih.gov/mesh

by an expert of the radiology domain and two non-experts.

All sentences (with previously tagged findings) were annotated as *Affirmed* if it is possible to infer that the finding is present in the patient, *Negated* if the finding is absent, *Probable* if it is not certain that the finding is present, but it probably is, and *Doubt* if the finding corresponds to the past or if it is not clear for the annotator if the finding is present or not. For results evaluation *Probable* annotations were considered as *Affirmed*, since physicians are interested in retrieving them, and sentences categorized as *Doubt* were replaced by other sentences (that were also annotated). In the cases where there was no agreement among annotators, usually the radiology-expert criteria was respected. In case of doubt the annotation criteria was revised by the annotators and the annotation was done according to the results of this process.

In both cases, the annotation process was performed in two stages, so that we could revise the annotation criteria. Some annotated sentences were overlapped, with the objective to calculate the Inter Rater Agreement (IRA) between annotators to measure their level of agreement. As measure for that goal we calculated Cohen's Kappa coefficient (Cohen, 1960).

Figure 2 shows the number of sentences annotated by each annotator individually and by more than one annotator in the testing dataset. Kappa coefficient (κ) was calculated for two sets: 1) 100 sentences annotated by non-expert annotator 1 and radiology domain expert (annotator 3), and 2) 100 sentences annotated by non-expert annotator 2 and annotator 3. Table 1 shows κ measure for the testing dataset. κ measure for the analysis dataset had similar results.

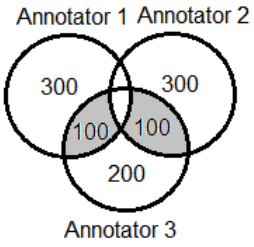

Figure 2: Number of sentences annotated by different annotators in the testing dataset.

annotators	κ
A1 and A3	0.97
A2 and A3	0.96

Table 1: IRA of expert/non-experts annotation in the testing dataset. A1 and A2 are computer science experts (not medical, nor linguistic experts), A3 is a radiology expert.

4 Results

Table 2 shows the performance of our NegEx adaptation and our syntactic methods to Spanish compared to the baseline. We show the best result of NegEx (obtained from the trigger set built from a combination of translated triggers, bi and trigrams and a list of terms suggested by the radiology expert). F1 using NegEx only with translated triggers was similar: 0.91 (81 TP, 76 FP, 144 FN and 699 TN). Results of NegEx with the original triggers (translated) and without the addition of coordinated negations (and tested with another dataset) can be seen in (Stricker et al., 2015).

Precision, Recall and F1 measure are the usual measures in the field and here are based on the interpretation of finding real negations. *F1* measure balances *precision* -how many findings identified as negated, are actually negated- and *recall* -proportion of the negated findings that were retrieved-. *Accuracy* is the rate of correctly classified sentences. True Positive (TP) refers to terms negated by the Gold Standard and correctly predicted by the methods. See Table 4 for the meaning of False Positive (FP), True Negative (TN) and False Negative (FN).

5 Discussion

All algorithms outperform *dictionary lookup*, our baseline algorithm. This makes sense, since the baseline does not take negation scope into account. For example in "*ectasia* pielica izquierda **sin** cambio de diametro postmiccional" what is negated (cambio de diametro postmiccional) is not the finding (ectasia). The baseline algorithm detects the negation (sin) and assumes wrongly that the finding is negated. This scope problem is solved in the rest of the algorithms developed.

Constituent tree patterns and dependency tree patters were tested assuming that they would perform better than PoS tagging patterns and NegEx in the detection of the negation scope, since in

Algorithm	Pattern Matching (baseline)	NegEx (adapted to Spanish)	POS Tagging Patterns	Constituent Tree Patterns	Dependency Tree Patterns
TP	201	220	219	200	194
FP	107	31	31	19	61
FN	24	5	6	25	31
TN	668	744	744	756	714
Accuracy	0.87	0.96	0.96	0.96	0.91
Precision	0.65	0.88	0.88	0.91	0.77
Recall	0.89	0.98	0.97	0.89	0.86
F1	0.75	0.92	0.92	0.90	0.81

Table 2: Performance of different algorithms with testing dataset composed by 1000 sentences.

Algorithm	NegEx (Costumero et al., 2014)	NegEx (Stricker et al., 2015)	NegEx (adapted)
F1	0.74	0.67	0.73

Table 3: Performance of different implementations of NegEx with (Costumero et al., 2014) dataset

	predicted Neg	predicted Aff
actual Neg	TP	FN
actual Aff	FP	TN

Table 4: *actual* stands for Gold Standard annotation, *predicted* for algorithms output.

these two methods we have not to consider fixed windows of words between the negation and the term of interest (as we do consider in NegEx) or each word that forms the sentence (as we do in our PoS tagging method). Nevertheless NegEx and the PoS Tagging based method have better results (not very different from constituent tree patterns). We understand that two factors influence these results: 1) the sentences of the reports are usually in our case relatively short (average: 14 words, longest: 74 words). This explains why having fixed windows of 6 words might be good enough for our data and suggests that we do not need to use more complex methods, that are independent of the length of the sentence and that do not fix word distance. That is, the linear analysis performed by PoS tagging patterns might be enough for these sentences. Dependency and constituent parsing, that perform an analysis based on the sentence structure, might be left for the most complex sentences. 2) MATE, the tool used to do the dependency parsing was trained based on a general language[12] that includes documents of the medical domain, but that is not restricted to it[13].

Regarding NegEx, another implementation was tried with a very reduced trigger set, in order to try to do it domain independent (see Table 3). F1 is similar when tested with our test set (0.91 instead of 0.92), and it is also similar (0.73) to F1 obtained by (Costumero et al., 2014) (0.74) and better than F1 obtained by (Stricker et al., 2015) (0.67) when tested with Costumero's dataset. This demonstrates that our NegEx implementation with a reduced trigger set could be used for data different that radiology reports.

Further analysis of results shows that: 1) the addition of a line of code to NegEx algorithm allows us to handle complex negations. E.g. in *"no se detectaron finding1 ni finding2"* (*"finding 1 and finding 2 were not detected"*), when *"finding2"* is the term of interest. Those kinds of negations are also handled correctly by the patterns built from our syntactic methods, but in some cases negations are much more complex and are not correctly parsed by the dependency parsing algorithm. 2) Sometimes, negations are not affecting the term of interest, but a modifier of it and the algorithm tags the term of interest as negated. For example, in *"pancreas: no visible por abundante gas"* (*"pancreas: not visible due to abundant gas"*). The

[12] https://www.iula.upf.edu/corpus/corpusuk.htm

[13] Besides, the area of documents in the medical domain is broad and the ones used differ from radiology reports.

trigger *"no"* (*"not"*) is applied to *"visible"* (*"visible"*), but the term of interest is *"gas"* (*"gas"*). 3) Constituent tree patterns method has shown to fail where there are no punctuation signs. This shows that the characteristics of the noisy text makes the success of syntactic techniques more complicated.

NegEx shows to perform better than a previous implementation for radiology reports in Spanish (Stricker et al., 2015) and similar than an implementation for general medical texts also in Spanish (Costumero et al., 2014) (see Section 5). Our Pos-Tagging results and the ones reached by (Huang and Lowe, 2007) for radiology reports in English are similar. They obtain 0.90 recall, 0.97 precision and 0.93 F1, while we obtain 0,88 recall, 0,97 precision and 0,92 F1. (Sohn et al., 2012) results for negation detection in clinical texts in English using dependency parsing are also similar to our dependency parser results. They obtain 0.74 recall, 0.97 precision and 0.84 F1, while we obtain 0.77, 0.86 and 0.81 for each of these measures. Nevertheless, it is not easy to compare results with existing papers, since languages and corpora are not the same.

6 Conclusion

Considering the different methods implemented for the detection of negations of terms of interest in radiology reports written in Spanish, NegEx has good results, but only considers partially the negation scope over the target term (since it is calculated based on a fixed-size window of words). Among the pattern methods tested, PoS tags allows us to study the ordering of words in phrases containing negations and to elaborate patterns based on them. But they are dependent on each word of the sentence. Based on a reduced dataset it is not easy to model all type of forms that sentences with negated findings may have. Constituent and dependency tree pattern methods differ from the *PoS tagging method* in that the whole structure of the sentence is used. Constituent tree method segments the sentence in syntactic related groups. These cases do not have to take so many detail into account and are easier to build. Both methods differ in that the second takes into account the dependence among each type of word in the sentence. Dependencies are modeled in a tree and each edge is labeled with the relation that exists among the words.

Detection negation in medical reports is a challenging task as it is characterized by short sentences and informal language often noisy. Furthermore, tools for Spanish in general are less developed than in other languages even more in this specific subdomain. For example, the availability of a large corpus of annotated medical reports (and specifically those in the radiology domain) would enable to have a better behavior of all language related tools (in particular POS tagging as well as constituent/dependency parsers). RadLex, is a comprehensive lexicon of radiology terms that was chosen to detect findings due to its adequacy to our domain of interest. Its translation to Spanish was made locally but unfortunately includes some errors, such as the order of resulting words and issues derived from ambiguity. All these issues made negation detection more difficult.

The high IRA obtained among the annotations performed by the specialist and two non-specialist could imply that this particular type of reports of short sentences could be annotated by non-specialists in the domain. We consider this is an important result, given the scarcity of resources.

We consider that having short sentences (ours have an average of 14 words) may contribute to the fact that NegEx and PoS tagging methods have similar results than the constituent tree method and better results than dependency tree method. An analysis should be performed with more complex sentences in order to test what happens in those cases. The effectiveness of syntactic techniques depends on the compliance of the text to the language grammatical rules. The results obtained support this asseveration.

7 Future Work

We are currently working in analyzing improvements to the dependency parser patterns and we are performing a further analysis of results, evaluating alternative methods (voting method, where the classification (Affirmed/Negated) is based on the tag received by most of the methods) and evaluating the possibility of implementing a hybrid methodology -taking the best of NegEx and syntactic methods- that reduces errors in order to obtain better F1.

We would like to extend our work for dealing with *hedges* and we plan to continue using these methods for other type of medical reports written in Spanish.

References

Zubair Afzal, Ewoud Pons, Ning Kang, Miriam Sturkenboom, Martijn Schuemie, and Jan Kors. 2014. ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus.

Blanca Arias, Núria Bel, Mercé Lorente, Montserrat Marimón, Alba Milà, Jorge Vivaldi, Muntsa Padró, Marina Fomicheva, and Imanol Larrea. 2014. Boosting the Creation of a Treebank. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 9(Suppl 11):775–781.

Alan R. Aronson, Thomas C. Rindflesch, and Browne C. Allen. 1994. Exploiting a Large Thesaurus for Information Retrieval. In *Proceedings of RIAO: Recherche d'Information Assistée par Ordinateur. Conférence*, pages 197–217, New York, USA.

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richard Farkas, Filip Ginter, and Jan Hajic. 2013. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics(TACL)*, 1(Suppl 11):415–428.

Xavier Carreras, Isaac Chao, Llus Padr, and Muntsa Padr. 2004. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001a. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001b. Evaluation of Negation Phrases in Narrative Clinical Reports. In *Proceedings of AMIA, American Medical Informatics Association Annual Symposium*, page 105, Washington, DC, USA.

Wendy W. Chapman, Dieter Hillert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E. Chapman, Mike Conway, Melissa Tharp, Danielle L. Mowery, and Louise Deléger. 2013. Extending the NegEx Lexicon for Multiple Languages. In *Proceedings of the 14th World Congress on Medical and Health Informatics*, pages 677–681, Copenhagen, Denmark.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Roberto Costumero, Federico López, Consuelo Gonzalo-Martín, Marta Millan, and Ernestina Menasalvas. 2014. An Approach to Detect Negation on Medical Documents in Spanish. In *Brain Informatics and Health*, volume 8609, pages 366–375.

Viviana Cotik, Darío Filippo, and José Castaño. 2015. An Approach for Automatic Classification of Radiology Reports in Spanish. In *Proceedings of 15th MEDINFO*, pages 634–638.

Noa P. Cruz Díaz, Manuel Jesús Maña López, and Jacinto Mata Vázquez. 2010. Aprendizaje Automático Versus Expresiones Regulares en la Detección de la Negación y la Especulación en Biomedicina [Machine Learning versus Regular Expressions in Negation and Speculation Detection in Biomedicine]. *Procesamiento del Lenguaje Natural [Natural Language Processing]*, 45:77–85.

Louise Deléger and Cyril Grouin. 2012. Detecting negation of medical problems in French clinical notes. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 697–702.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and Their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 1–12, Uppsala, Sweden.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. ConText: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports. *Journal of biomedical informatics*, 42(5):839–851, October.

Yang Huang and Henry J Lowe. 2007. A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. *Journal of the American Medical Informatics Association*, 14(3):304–311.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, volume 1, pages 63–70, Philadelphia, Pennsylvania.

Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C Max Schmidt, Hongfang Liu, et al. 2015. Deepen: A negation detection system for clinical text incorporating dependency relation into negex. *Journal of biomedical informatics*, 54:213–219.

Roser Morante and Eduardo Blanco. 2012. Sem 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 265–274, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2009. A Metalearning Approach to Processing the Scope of Negation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 21–29, Boulder, Colorado.

Christopher Potts. 2011. On The Negativity of Negation. In *Proceedings of Semantics and Linguistic Theory*, volume 20, pages 636–659, New Brunswick, New Jersey.

Thomas C. Rindflesch and Alan R. Aronson. 1994. Ambiguity Resolution While Mapping Free Text to the UMLS Metathesaurus. In *Proceedings of the 18th Annual Symposium on Computer Application in Medical Care*, pages 240–244, Washington, DC, USA.

Lior Rokach, Roni Romano, and Oded Maimon. 2008. Negation Recognition in Medical Narrative Reports. *Journal of Information Retrieval*, 11(6):1–50.

Maria Skeppstedt. 2011. Negation Detection in Swedish Clinical Text: An Adaption of NegEx to Swedish. *Journal of Biomedical Semantics*, 2(Suppl 3):S3, January.

Sunghwan Sohn, Stephen Wu, and Christopher G Chute. 2012. Dependency parser-based negation detection in clinical narratives. *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2012:1–8.

Vanesa Stricker, Ignacio Iacobacci, and Viviana Cotik. 2015. Negated Findings Detection in Radiology Reports in Spanish: an Adaptation of NegEx to Spanish. In *IJCAI - Workshop on Replicability and Reproducibility in Natural Language Processing: adaptative methods, resources and software*, Buenos Aires, Argentina.

Anita Sundaram. 1996. Information Retrieval: A Health Care Perspective. *Bulletin of the Medical Library Association*, 84(4):591–593.

Özlem Uzuner, Xiaoran Zhang, and Tawanda Sibanda. 2009. Machine Learning and Rule-based Approaches to Assertion Classification. *Journal of the American Medical Informatics Association : JAMIA*, 16(1):109–115.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–556.

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP'10, pages 60–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

2010. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrew S. Wu, Bao H. Do, Jinsuh Kim, and Daniel L. Rubin. 2011. Evaluation of Negation and Uncertainty Detection and Its Impact on Precision and Recall in Search. *Journal of digital imaging*, 24(2):234–242, April.