Negated Findings Detection in Radiology Reports in Spanish: an Adaptation of NegEx to Spanish

Vanesa Stricker[1], **Ignacio Iacobacci**[2] **and Viviana Cotik**[1]
[1]Departamento de Computación, FCEyN, UBA, Argentina, {vstricker,vcotik}@dc.uba.ar
[2]Department of Computer Science, Sapienza University of Rome, iacobacci@di.uniroma1.it

Abstract

Entity recognition in biomedical texts is an important step in the path towards automatizing clinical text analysis. In order to understand which conditions are present and which are absent, negation detection has to be performed. Most of the available work in this domain has been carried out in the English language.

In this article we present SpRadNeg, which is an adaptation of NegEx to the Spanish language. NegEx is an English rule-based negation detection algorithm. We have tested SpRadNeg with radiology reports, obtaining a precision of 0.87 and a recall of 0.49. We also propose a method to automatize text annotation based on Machine Learning techniques with 0.91 precision and 0.89 recall.

1 Introduction

Automatic identification of relevant terms in medical reports is useful for clinical, educational and research purposes.

A clinical condition identified in the text is not necessarily present, since the term that represents it could be negated or have an uncertain condition associated to it. For example, in *"no se detectaron dilataciones ventriculares"* for *"no ventricular dilatation were detected"*, *"no se detectaron"* (*"were not detected"*) indicates that the medical condition (*"ventricular dilatation"*) is negated. There are many language constructions, that in some contexts denote negations such as *"no se puede ver"* (*"it cannot be seen"*), *"libre de"* (*"free of"*) and *"sin ninguna evidencia de"* (*"no evidence of"*) among others. We are going to call *negations* or *triggers* to these language constructions. We call *findings* or *terms of interest* to medical conditions and observations made about a particular illness in tests and medical examinations. For instance, *"dilataciones ventriculares"* (*"ventricular dilatations"*), *"herida"* (*"wound"*) are called findings. Texts can also contain *hedges*, which indicate the uncertainty of conditions, for instance *"sugestivo de"* (*"suggesting"*).

According to [Chapman *et al.*, 2001b], approximately half of the medical conditions described in unstructured texts of the medical domain are negated, that is: they are described implicitly or explicitly as non existent in a patient. For example, the *"hipertension"* (*"hypertension"*) clinical condition appears negated in the following medical text, *"no se observan signos de hipertension portal"* (*"no signs of portal hypertension are observed"*).

For this reason, the detection of negations in the text of the biomedical domain is an area of study of the field of Natural Language Processing (NLP) called BioNLP. Negation detection is also studied in other domains [Potts, 2011].

The goal of the work is to take a set of medical records of the radiology domain (ultrasonography reports) written in Spanish with findings tagged automatically with a tool based on a specific radiology corpus [Cotik *et al.*, 2015] and to develop an algorithm so as to determine if the findings are negated or not. In order to achieve this goal, we have developed SpRadNeg. SpRadNeg consists of an adaptation and improvement of NegEx algorithm [Chapman *et al.*, 2001a] to Spanish for the radiology domain. NegEx is an algorithm that uses a list of terms or triggers to determine whether clinical conditions are negated or not in a sentence in medical records. The algorithm requires to have manually annotated texts. Since manual annotation is a time-consuming task, a machine learning technique has been tested in order to evaluate the possibility of increasing automatically the size of the annotated texts.

The NegEx adaptation from English to Spanish requires the comprehension of the algorithm, the remodeling of the triggers in Spanish (i.e. translation and further treatment) and the obtention of a corpus in Spanish with tagged findings and that is manually annotated in order to determine if findings are negated or not.

An example of a tagged ultrasonography report in Spanish and its translation to English can be seen below:
"384 —15y 3m—20090412—A423517 Higado: lobulo caudado <FINDING>aumentado</FINDING> de tamano, tamano y ecoestructura normal. Via biliar intra y extrahepatica: no <FINDING>dilatada</FINDING>. Paredes y contenido normal. Pancreas: tamano y ecoestructura normal. Retroperitoneo vascular: sin <FINDING>alteraciones</FINDING>. No se detectaron <FINDING>adenomegalias</FINDING>. Ambos rinones de caracteristicas normales. (...)"(*"Liver: <FINDING>enlarged</FINDING> caudate lobe, size and echostructure normal. Intra and extrahepatic bile duct: not <FINDING>dilated</ FINDING>. Wall and content appear normal. Pancreas: normal size and echo-*

texture. Vascular retroperitoneum: without <FINDING> changes </FINDING>. No <FINDING> lymphadenopathy </FINDING> was detected. Both kidneys of normal characteristics. (...) ")

Each sentence with findings is used as input for NegEx. Previously an annotation has to be performed in order to know whether the finding is Affirmed or Negated. The following sentences illustrate the format required by NegEx. Each line contains the number of the report, the finding (in our case tagged automatically), the sentence where it appears and, finally, *Affirmed* if the finding is not negated or *Negated* if it is denied.
-384 dilatada Via biliar intra y extrahepatica: no dilatada Negated (*384 dilated Intra and extra hepatic bile duct: not dilated Negated*)
-384 aumentado Higado: lobulo caudado aumentado de tamano Affirmed (*384 enlarged Liver: enlarged caudate lobe Affirmed*)

Diverse methods and resources of the NLP area will be used in order to achieve our goal. Some of the challenges of our proposed solution are:

- to provide an adaptation of the original NegEx triggers to Spanish.
- the adaptation of the algorithm to Spanish, a language with limited resources and tools (such as annotated corpora and NLP tools).
- to have an automatic finding detector. The accuracy of our present approach relies in the correct identification of terms of interest.

The rest of the paper is organized as follows. Section 2 presents previous work in the detection of negation terms in the medical domain, including the original NegEx approach and its adaptations to other languages different than English. Section 3 presents our main contribution, by explaining the methods, materials and data sets used. Section 4 shows the results of testing SpRadNeg with different data sets, compares the results with another implementation of NegEx to Spanish which has been applied to a different kind of input data and shows the results of an attempt to improve SpRadNeg (with improvable results) and an attempt to automatize annotation (with satisfactory results). Finally, Discussions, Conclusion and Future Work are presented.

2 Previous work

The use of information retrieval techniques for automatically indexing narrative medical reports is present since late 1994, [Aronson *et al.*, 1994; Rindflesch and Aronson, 1994; Sundaram, 1996]. The need to determine not only if a finding is mentioned on narrative medical reports but also if such finding is present or absent inspired the work of Chapman *et al.* [2001a]. They developed an algorithm based on regular expressions called NegEx, which implements several phrases indicating negation and limits the scope of the negation phrases. The promising results of this simple approach have motivated the development of other works based on it. Wu *et al.* [2011] developed a modified version of NegEx (negation phrases of the radiology domain and hedge identification were added) and used it in a word-based radiology report search engine. Harkema *et al.* [2009] developed ConText, a NegEx-based tool, that employs a different definition for the scope of triggers. It also expands the detection of negation in findings with three new categories: hypothetical, historical, and experienced. Finally, this approach is adapted for six different types of medical reports (including radiology). Other works centered their efforts to adapt NegEx to different languages, as Skeppstedt [2011] for Swedish, Chapman *et al.* [2013] for French, German, and Swedish and Costumero *et al.* [2014] for Spanish. There exist other approaches for the negation detection task, for instance the combination of pattern matching and machine learning techniques performed by Cruz Díaz *et al.* [2010]. Morante and Daelemans [2009] applied machine learning to establish where the scope of a single negation ends. Rokach *et al.* [2008] extracted automatically several regular expressions and patterns from annotated data and used them to train a decision tree. Uzuner *et al.* [2009] trained a SVM based not only in words but also in several features as *eye color* with their corresponding values (e.g. green, brown). Several challenges have been performed on this [Uzuner *et al.*, 2011; Kim *et al.*, 2009] and other domains [Farkas *et al.*, 2010].

3 Methods

In this section we explain the NegEx algorithm, the approach followed to adapt the algorithm to Spanish, its use to detect negations in radiology reports and the problems encountered in the process. We present a comparison with the results of the approach taken by Costumero *et al.* [2014] to adapt NegEx for the detection of negations in clinical records written in Spanish.

3.1 The NegEx algorithm

NegEx is an algorithm for negation detection in medical reports that is used to determine whether a finding or disease is absent or present in a patient according to the medical record description. The algorithm 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. If the algorithm determines that a finding or disease is negated we say that it is absent. E.g. *"no sings of infection"* would return that the finding *"infection"* is negated.

In order to determine the accuracy of the algorithm, NegEx uses a Gold Standard (GS) that consists of a set of sentences with tagged findings and an annotation telling whether the identified terms are negated or not. Usually the annotation is performed manually.

In the NegEx original version [Chapman *et al.*, 2001a] 35 negation phrases are identified and divided into two groups. The first group is composed by *pseudo negation phrases*: phrases that indicate double negatives (not ruled out). In this case the presence of the negation trigger does not indicate the absence of the clinical condition. The second group consists

of phrases used to deny findings. The phrases are represented by regular expressions and could be in one of two groups: preceding the findings or following it. A NegEx extension [Chapman *et al.*, 2013] adds two new groups: *termination terms*, that indicate the end of the scope of the negation trigger (e.g.: "but"), and *conjunction terms*, that indicate the possibility of the presence of a negation (e.g.: "except"). A label is used to classify each trigger in one of these groups.

NegEx only takes into account the sentence where the term of interest appears in order to determine whether it is negated or not, i.e. it does not use information of other sentences.

The output of NegEx for one of the input sentences shown in Section 1 is *"384 dilatada Via biliar intra y extrahepatica: no dilatada Negated Via biliar intra y extrahepatica: PREN no PREN dilatada Negated."* (*"384 dilated Intra and extra hepatic bile duct: not dilated Negated Intra and extra hepatic bile duct: PREN not PREN dilated Negated"*). The sentence corresponds to report number 384. *"dilated"* is the term of interest and it was manually tagged as Negated. The output of NegEx tells it is negated and shows in which position of the sentence the trigger appears. Finally, the PREN label indicates that the trigger precedes the finding.

3.2 The NegEx adaptation

The process to obtain SpRadNeg is as follows: a set of medical reports is chosen. An algorithm is applied in order to automatically detect findings. Then, a sentence tokenization is performed using NLTK [Loper and Bird, 2002]. We also obtain the Spanish triggers and build the Gold Standard (GS). After that, NegEx is applied to our data set, using the triggers obtained beforehand to detect if the findings tagged are negated or not. Negex uses the GS to get quantitative and qualitative results. Finally, we analyze NegEx results in order to propose improvements to the algorithm. Next sections describe these processes.

Data

We used two data sets. With the first (our data set of radiology reports) we tested our algorithm. The other was obtained through personal communication with Costumero *et al.* [2014]. This was the data set used to test their algorithm, and we also used it to compare our results with their results.

Our data set consists of about 85600 reports of ultrasonography studies performed in a public hospital. Reports are written in Spanish in non-structured format (the first part is semi-structured, see Section 1). They are brief (approximately 5 lines each) and they state what was found in the study performed on the patient. From our data set a smaller set of reports was selected. The idea is to have a corpus of data with similar characteristics as the NegEx corpus in order to make a reasonable comparison. Therefore, the size of the corpora used to test previous works was analyzed. Generally 500 sentences with negated findings and 500 with positive findings were used.

We selected 10 radiology reports of our data set. They were composed by 66 sentences. Only 31 sentences had findings or diseases. The remaining 35 were discarded. 35 findings were found in the sentences. Manually, 20 negation phrases were detected in the sentences containing findings. Summarizing, 10 reports contain about 20 sentences with negation phrases, and 10 sentences without negation phrases. Therefore, 500 reports were selected from the whole data set to find about 500 sentences containing negation phrases and 500 without negation phrases. These 500 reports compose our corpus.

The other data set, used originally by Costumero *et al.* [2014], was extracted by the authors from SciELO [Packer, 1999] using the sections entitled *"Reporte de caso"* ("*Case report*") *"A propósito de un caso"* ("*About a case*") and *"Caso clínico"* ("*Clinical case*"), among others.

Findings detection

There are various inventories that serve as a basis to detect relevant terms in medical reports. The International Classification of Diseases (ICD[1]) is a standard diagnostic tool for epidemiology, health management and clinical purposes. The most widespread version is known as ICD10, the 10th revision. SNOMED CT[2] is a clinical health terminology ontology, owned and distributed by The International Health Terminology Standards Development Organization (IHTSDO[3]). UMLS[4] (Unified Medical Language System) is a set of files and software that bring together many health and biomedical vocabularies and standards to enable interoperability between computer systems. Finally, RadLex[5] is a lexicon centered only on radiology terms. SNOMED CT, UMLS and ICD-10 are available in Spanish, RadLex is only available in English and in German.

In the original NegEx implementation UMLS is used to detect terms. The Swedish adaptation [Skeppstedt, 2011] uses UMLS, KSH97-P -a Swedish adaptation of ICD-10- and MeSH [6].

We used an information extraction algorithm [Cotik *et al.*, 2015] based on the appearance of RadLex *pathological terms* in the reports in order to tag the findings in the 500 reports. RadLex was chosen because it is a lexicon specifically developed for the radiology domain. It has the disadvantage that it does not exist a Spanish version, 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. In the pathological term detection used by SpRadNeg, the algorithm matched the longest possible string among eligible matches.

Implementation details

In the resulting corpus there are sentences that contain more than one finding. For example, for the sentence *"no se detectaron colecciones ni liquido libre"* (*"collections or free liquid were not detected"*), *"colecciones"* (*"collections"*) and *"liquido libre"* (*"free liquid"*) are two different findings. In these cases, the sentence is repeated as many times as findings it has. In the example, the sentence from the pre-processed corpus gives us two sentences, in the resulting corpus, one

[1] http://www.who.int/classifications/icd/en/

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

[3] http://www.ihtsdo.org/

[4] http://www.nlm.nih.gov/research/umls/

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

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

for *"colecciones"* (*"collections"*) and the other for *"liquido libre"* (*"free liquid"*).

Triggers

NegEx triggers were translated into Spanish using Google Translate[7]. We decided to do automatic translation, since translation is an expensive task and we are not experts in the domain. Translations were revised by a non-expert and those that were not correct were eliminated or corrected.

We performed a work similar to the performed by Skeppstedt [2011]. English lacks grammatical gender, while Spanish has two (male and female). Adjectives have gender agreement. For some of those cases we generated inflections of adjectives and we expanded the English negative quantifier (for example from *"no"* to *"ningún" "ninguno" "ninguna"*). We obtained 340 translated triggers.

Annotations

An important issue in the adaptation of an existing system to another language is the lack of a Gold Standard for validating the reliability of the new model. Annotating is an expensive task, and domain experts are not always available. In this case we decided to do the annotation by non-experts. Therefore, a set of reports was automatically tagged for findings, then all the sentences with findings were annotated by two non-specialist annotators as *Affirmed* if it is possible to infer that the finding is present in the patient, or *Negated* if the finding is absent. This constituted the corpus to test SpRadNeg.

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, for N items classified into C mutually exclusive categories, we calculated the Cohen's Kappa coefficient. The equation is as follows:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (1)$$

where Pr(a) is the relative observed agreement among raters, and Pr(e) is the hypothetical probability of chance agreement. Using the observed data it is possible to calculate the probabilities of each annotator randomly choosing each category.

Table 1 shows the number of sentences annotated by each annotator individually, the number of shared sentences (annotated by both) and the κ measure. Annotation criteria was revised and adjusted after the first annotation.

number of sentences	# shared sentences	#individual sentences	κ
160	40	60	0.91
1000	250	375	0.95

Table 1: Size of Annotation set (column 1), and Inter Rater Agreement. # denotes *number of*. Row 1 has the values of the first annotation and row 2 of the second.

[7]https://translate.google.com/

3.3 Automatic Classification

Classification is the task of assigning one or more classes to a single element. For this purpose, we need to have a set of annotated data in order to learn the implicit relations in the class assignment. The mark which is used to identify all assignable classes is called *tag*. In the context of negation detection, our elements are the analyzed medical reports, and the tag is the indication of the negation of the finding.

We decided to do two tests. Due to the high cost of creating a manually annotated corpus as an input to NegEx, we decided to try Test 1: the use a machine learning algorithm in order to create automatic annotations. We also created Test 2 to use SpRadNeg output as input of a Machine Learning (ML) model in order to improve SpRadNeg classification. In Test Number 2 we fed the model with the output of SpRadNeg. This output was composed of the sentences with their findings -both were input to the algorithm-, and their corresponding annotation (positive or negative). This dataset was used as input for a Naive Bayes (NB) classifier. Since we were looking for the feasibility of this approach we chose an algorithm which could be seen as a baseline for this kind of approach. Alternative models could improve our results.

We chose a ML toolkit for NLP tasks called MALLET[8] (MAchine Learning for LanguagE Toolkit). MALLET uses the *bag-of-words* model to represent the sentences. Bag-of-words defines a dictionary, containing the whole vocabulary included in the training set, in which each word is mapped to a unique position in a vector. The sentences are represented as vectors with the length of the dictionary. Each position, commonly known as feature value, has the amount of occurrences of the word in the sentence.

4 Results

Table 2 shows the performance of SpRadNeg, our NegEx adaptation to Spanish with our radiology data and the results of applying SpRadNeg to Costumero *et al.* [2014] data. The performance of Costumero *et al.* [2014] algorithm and data is also shown. Sentences reported as missing were not taken into account, because of three reasons: either 1) there were doubts in the annotation process, 2) they corresponded to hedges or 3) the two annotators annotated them differently.

Tables 3 and 4 show the results of applying ML techniques to improve SpRadNeg results (see Section 3.3 Test 2) and to automate manual annotations of positive/negative findings (Section 3.3 Test 1). The number of sentences correspond to the test set size.

For measuring accuracy, *10-fold cross validation* was performed for both Test 1 and Test 2. This task splits an annotated corpus in *n* parts or *folds*, trains the classifiers with *n-1 folds* and uses the remaining *fold* for testing. Afterwards, this process is repeated *n* times, using each time a different *fold*. The results of each run are averaged.

F1 is a measure that balances precision -from the identified as negated, how many really are negated- and recall -proportion of the negated findings that were retrieved-. The accuracy is the rate of correctly classified sentences. Equations for these measures are:

[8]http://mallet.cs.umass.edu/

Algorithm	Costumero *et al.* [2014]	SpRadNeg	SpRadNeg
Data Set	SciELO	SciELO	Radiology
Sentences	500	500	1000
Missing	46	46	21
TP	61	63	200
FP	25	45	30
FN	18	16	208
TN	350	330	540
Accuracy	0.82	0.87	0.76
Precision	0.71	0.58	0.87
Recall	0.77	0.80	0.49
F1	0.74	0.67	0.63

Table 2: Performance of the SpRadNeg algorithm: first column corresponds to the performance of Costumero et al. algorithm using their data (obtained through personal communication), second column corresponds to the performance of SpRadNeg with Costumero et al. data, third column to SpRadNeg implementation with our radiology data set.

Score	SpRadNeg	ML to SpRadNeg
Sentences	196	196
TP	32	25
FP	5	12
FN	34	41
TN	125	118
Accuracy	0.80	0.73
Precision	0.86	0.68
Recall	0.48	0.38
F1	0.62	0.49

Table 3: Performance of Naive Bayes to improve SpRadNeg results (Test 2). Column *SpRadNeg* shows the results of SpRadNeg. Column *ML toSpRadNeg* shows the result of applying NB to the results of SpRadNeg.

$$\text{accuracy}= \frac{\#TP+\#TN}{total}$$
$$\text{precision}= \frac{\#TP}{\#TP+\#FP}$$
$$\text{recall}= \frac{\#TP}{\#TP+\#FN}$$
$$\text{F1}= 2 * \frac{prec*recall}{prec+recall}$$

where # denotes *amount of*, and TP (True Positive) denotes a Negated tag predicted by the algorithm that is Negated in the annotation (Goldstandard, GS), FP (False Positive) is the case when the algorithm tags as Negated but the GS determines it is Affirmed. FN (False Negative) stands for Affirmed tag done by SpRadNeg, and Negated in the GS, and TN: both, SpRadNed and GS determine that it is Affirmed.

4.1 Discussion

SpRadNeg with SciELO data has higher accuracy and recall than the Algorithm of Costumero *et al.* with their own data. Precision and F1 are lower.

Comparing SpRadNeg with SciELO data and SpRadNeg with radiology reports, the first has higher accuracy, recall and F1 than the second, but radiology data has much higher precision (i.e. of all the findings that SpRadNeg determines that are negated, how many of them really are negated). This result is better with SpRadNeg (both with SciELO and with radiology data).

Score	ML with Annotations
Sentences	196
TP	59
FP	6
FN	7
TN	124
Accuracy	0.93
Precision	0.91
Recall	0.89
F1	0.90

Table 4: Performance of Naive Bayes to automatize positive/negative findings annotations process (Test 1).

If we analyze both NegEx adaptations with SciELO data, Costumero *et al.* adaptation has better results. This might be because their triggers are more adapted to their domain than ours.

The goal we would like to achieve is to have a method that tags those findings present in the radiology reports that are negated -to discard them in order to get a final set of reports with findings present in patients-. If many of the terms of interest are falsely identified as negated, the findings will not be discarded. So, the final set of reports would contain reports with findings that are not present in patients. We would like to minimize that error. It is important to know that there is generally a trade-off between precision and recall. While high recall is certainly beneficial when the radiologist is searching for uncommon entities, it can be highly problematic when searching for common diseases and findings [Wu *et al.*, 2011]. Instead, high precision provides best results for those cases. We would like to minimize that error, and therefore maximize precision. SpRadNeg has the best precision.

Results of Test 2 (use of Naive Bayes in order to improve SpRadNeg) were not satisfactory. Table 3 shows that the accuracy of the results when applying ML is lower than the results when not applying it. Other ML models, such as MaxEnt or SVM, which could have improved the results were not tested. However, from the previous results we consider that the application of a ML method to the output of SpRadNeg -in the way we are doing it- is not an useful approach to take to improve NegEx. Results of Test 1 (automatization of annotation process, see Table 4) reach an accuracy of 0.93. This is a promising result, since it implies that this time-consuming process could be automatized.

A first analysis of the problems found in the classification of the findings (as positive or negative) was performed. The main problems encountered are the following:

- in some cases the trigger is affecting not to the term of interest, but to 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 trigger *"no"* (*"not"*) is ap-

plied to *"visible"* (*"visible"*), but the term of interest is *"gas"* (*"gas"*).

- some terms, for example *"libre"* (*"free"*) appear as a negation, but from the knowledge of our data set we imagine that this term is probably part of a larger term called *"líquido libre"* (*"*free liquid"). We plan to analyze bigrams and trigrams of a reduced test set in order to know which words appear near the term *"libre"*(*"free"*).
- in some cases findings were incorrectly tagged, because of not being lemmatized.
- complex negations, not always have good results. E.g. in *"no se detectaron finding1 ni finding2"* (*"finding 1 and finding 2 were not detected"*), when *"finding2"* is the term of interest results could be incorrect.

5 Conclusion

We present an ongoing work to adapt NegEx, an algorithm for the detection of negations of terms in medical texts, to Spanish texts in the radiology domain. Our approach, called SpRadNeg, differs from others because of various reasons. It is applied to Spanish and it differs from a previous adaptation to this language because of 1) the domain used -which is specifically radiology- and 2) the length of text is concise (5 lines vs. about 20 lines). Finally it differs from others because of the method and lexicon used to detect findings. Both, negation and hedge detection are crucially important for information extraction, since an event or relationship has to be distinguished by its factual information.

In order to test results, an annotated dataset is needed (Gold Standard). Gold Standards are difficult to obtain and as far as we know a public GS of negation detection for radiology reports or general medical reports in Spanish is not available. The best solution would be to have an annotation provided by a specialist, but we have decided to take an alternative and less *expensive* approach: to manually annotate the corpus by non-specialized people. Annotations were overlapped in order to be able to test IRA, with satisfactory results. In order to obtain a less *expensive* way of getting annotations, we used Naive Bayes to automatically annotate reports based on previous manual annotation. The accuracy of 0.93 of this test (see Table 4) indicates that if we had a relatively reduced set of sentences annotated by an expert our ML algorithm could be applied to expand this set and use it as GS to test our algorithm. We did not use the results of the automatic annotation as input to SpRadNeg.

The goal we would like to accomplish is to have a method that tags those findings present in radiology reports that are not negated. If many of the terms of interest are falsely identified as negated, the findings will not be tagged in the report. We would like to minimize this error, and therefore maximize precision. SpRadNeg has better precision than Costumero *et al.* algorithm.

A method to improve SpRadNeg has been tested and reported. The improvement of SpRadNeg through Naive Bayes had poor results. Although other methods (such as SVM or MaxEnt), which would have probably given better results have not yet been tested, we feel that the application of ML techniques to SpRadNeg results, in the way we tried it, is not a useful method to improve SpRadNeg results. Nevertheless, in order to affirm this, those methods should be previously tested.

5.1 Future Work

We are currently working on improving our results and in incorporating *hedges*.

The improvement of our results includes 1) working on the trigger set. This includes analyzing the frequency of occurrence of each trigger, improving the generation of new triggers in cases were inflections exist, and trigger classification into classes. 2) analyzing in detail the sentences incorrectly tagged in order to determine if different techniques should be applied to some particular negation terms, 3) lemmatizing sentences. SVM and MaxEnt could also be tested to try to improve SpRadNeg (Test 2).

In addition, we will work on analyzing the scope of negations. Fixed rules, dependency parsing and machine learning techniques can be used as done previously in the work of Morante and Daelemans [2009]. Finally, we are considering to test our improvements on BioScope [Vincze *et al.*, 2008] -a freely annotated corpus on handling negation and uncertainty in biomedical texts- and in SciELO radiological texts (there are only 66 available) and to obtain a Gold Standard developed by a specialist in the radiology domain.

References

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

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

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

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

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

[Cotik *et al.*, 2015] Viviana Cotik, Darío Filippo, and José Castaño. An Approach for Automatic Classification of Radiology Reports in Spanish. In *Proceedings of MEDINFO. To Appear*, 2015.

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

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

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

[Kim *et al.*, 2009] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9, Boulder, Colorado, 2009.

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

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

[Packer, 1999] Abel Laerte Packer. SciELO-An Electronic Publishing Model for Developing Countries. *ELPUB*, pages 268–279, 1999.

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

[Rindflesch and Aronson, 1994] Thomas C. Rindflesch and Alan R. Aronson. 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, 1994.

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

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

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

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

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

[Vincze *et al.*, 2008] Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. The BioScope corpus: Biomedical Texts Annotated for Uncertainty, Negation and Their Scopes. *BMC bioinformatics*, 9(Suppl 11):S9, 2008.

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