An Approach to Detect Negation on Medical Documents in Spanish

Roberto Costumero[1], Federico Lopez[2], Consuelo Gonzalo-Martín[1], Marta Millan[2], and Ernestina Menasalvas[1]

[1] Universidad Politécnica de Madrid - Centro de Tecnología Biomédica, Madrid, Spain
{roberto.costumero,consuelo.gonzalo,ernestina.menasalvas}@upm.es
[2] Universidad del Valle, Colombia
{federico.lopez,marta.millan}@correounivalle.edu.co

Abstract. The adoption of hospital EHR technology is significantly growing and expected to grow. Digitalized information is the basis for health analytics. In particular, patient medical records contain valuable clinical information written in narrative form that can only be extracted after it has been previously preprocessed with Natural Language Processing techniques. An important challenge in clinical narrative text is that concepts commonly appear negated. Though worldwide there are nearly 500 million Spanish speakers, there seems to be no algorithm for negation detection in medical texts written in that language.

Thus this paper presents an approach to adapt the NegEx algorithm to be applied to detect negation regarding clinical conditions in Spanish written medical documents. Our algorithm has been trained with 500 texts where 422 different sentences and 267 unique clinical conditions were identified. It has been tested for negated terms showing an accuracy obtained is of 83,37%. As in the detection of definite affirmed conditions, the results show an accuracy of 84,78%.

Keywords: Natural Language Processing, Negation Detection, Medical texts.

1 Introduction

Health data analytics highly relies on the availability of Electronic Health Records (EHRs). The adoption of hospital EHR technology is significantly growing and is expected to continue growing. While one third of the US hospitals (35%) had already implemented some kind of EHR technology in 2011, it is expected that by 2016 nearly every US hospital (95%) will use EHR technology.

EHR contains both structured and not structured information such as: the patient's medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results. This information is being collected and managed by a health care provider or organization. Analysis of this information can make it possible to develop evidence-based decision making tools that providers can use to improve patient's care.

D. Ślęzak et al. (Eds.): BIH 2014, LNAI 8609, pp. 366–375, 2014.

However technology for health care data analysis is not mature enough due to, among other reasons, a lack of standards, interoperable data schemas and natural text and image processing tools. In particular in this paper we focus on natural text processing.

Patient medical records contain valuable clinical information written in narrative form. Thus, in order to find relevant information it is often necessary to extract it from free-texts in order to support clinical and research processes. An important feature of the clinical narrative text is that it commonly encloses negation concepts. According to Chapman et al. [1], around half of all clinical conditions in narrative reports are negated.

Several systems and methods have been proposed for negation detection [1], [2], [3], [4], [5] [6], [7], [8], [9], [10], [11], [12].

NegEx is introduced in [2] to identify negation terms. In order to identify negated finding and diseases, NegEx algorithm uses regular expressions to determine the scope of trigger terms. A clinical condition is marked as negated whether within the scope of a trigger term. NegEx was applied to selected sentences from discharge summaries with a 94.5% specificity, 84.5% precision and 78% sensivity.

In [1], NegEx' performance is analyzed taking into account different kinds of clinical reports. The study uses three types of negation phrases: pre-UMLS, post-UMLS and pseudo-negation. Negation phrases were extracted from previous analysis of NegEx on discharge summaries, a system called SymText and also negation phrases were added by the authors. NegEx obtained an average precision of 97%. However, precision ranged from 84% to 19% depending on the section of the pathology report.

Recently, challenges in translation of negation triggers from English negation lexicon (NegEx) to Swedish, French and German have been analyzed in [3]. OWL and RDF multilingual lexicon version were developed based on an extended NegEx version proposed in [4].

Elkin et al. [5] introduce a method based on NegEx to identify the scope of the negation trigger. The method is applied to small segments generated by isolating sentences from health records. Precision and recall values of 91.2% and 97.2%, were respectively obtained.

The extended NegEx version, called ConText [4], is based on regular expressions and helps to determine whether a clinical condition not only is negated, but also if it is hypothetical, historical or experienced by someone other than the patient. To do this, several contextual properties (hypothetical, historical or the experiencer) are taken into account to modify clinical conditions.

In [6] Negfinder, a program to identify negated patterns present in medical documents, is described. Documents are preprocessed replacing a concept by its UMLS concept ID. Then, negations are distinguished and grammar rules are used to associate them with single or multiples concepts preceding or succeeding them. The sensivity and specificity obtained from evaluation when Negfinder was applied to medical documents to detect negations were between 91% and 96%.

A hybrid approach was proposed in [7]. The approach combines regular expression matching with grammatical parsing to automatically detect negations in radiology reports. Negated phrases were identified with sensitivity and precision values of 92.6% and 98.6%, respectively.

Skeppstedt [8], [9], [13] adapted the English rule-based negation detection system, NegEx to Swedish. The results [8] showed lower precision and recall values than results obtained for the English version. In [9], NegEx system adapted to Swedish is analyzed to a subset of free-text entries from the Stockholm EPR corpus. Particularly, the study has been centered in SNOMED CT terms having the semantic categories 'finding' or 'disorder'. Recently, in [14] Swedish health records are studied emphasizing the analysis on four entities: Disorder, Finding, Pharmaceutical Drug and Body Structure. The study investigated how well named entity recognition methods work on clinical texts written in Swedish and whether to divide the Medical Problem category into more specific entities could be meaningful.

On the other hand, several methods based on machine learning techniques have been proposed [10], [11], [12]. A machine learning system, consisting of two classifiers to determine the scope of negation in biomedical texts, is introduced in [10]. The classifiers determine if the tokens in a sentence are negation signals and find the full scope of these negation phrases. An error reduction of 32.07% w.r.t similar systems was obtained [11] in different text types.

In [12] a pattern learning method is proposed in order to automatically identify negations in medical narrative texts. According to the authors, the accuracy is improved with respect to other methods based on a machine learning approach. Four steps integrate the method: corpus preparation, regular expression pattern learning, patterns selection and classifier training.

Though there are 500 million Spanish speakers worldwide (According to [15]) as far as our knowledge there is no learning method for negation detection in medical text written in Spanish.

Thus in this paper we present an approach to adapt the NegEx algorithm to be applied to detect negation regarding diseases in Spanish written medical documents. The paper contributions are as follows: i) a list of terms for Spanish is compiled from clinical reports containing those terms that identify negation; ii) the frequency of the terms has been calculated on a corpus to compare the values to the corresponding ones in English for which the analysis has already been performed; iii) an implementation of the NegEx algorithm for Spanish and iv) the evaluation of the algorithm with Spanish texts.

The rest of the paper has been organized as follows: in section 2 we present the adaptation of the method proposed by [3] to work with Spanish texts. In this section we also present the list of terms that have been identified in Spanish as triggers for negation. In section 3 we present the results of applying the implementation of our approach over 500 medical texts, containing 422 different sentences and 267 unique clinical conditions. To end with, section 4 presents the main conclusions obtained so far as well as the future lines of development.

2 Method

In this section we present the approach we have followed to adapt NegEx algorithm for Spanish texts. The approach we present is similar to the one presented in [3].

In order to do so the processes depicted in figure 1 have been performed:

1. **Medical texts annotation.** Several medical texts have been manually annotated to detect situations in which terms appear to be negated.
2. **Labelling of terms.** The terms from NegEx are extracted and translated into Spanish. The list of terms is enriched with those that have been previously detected in the annotated medical texts and known synonyms.
3. **Frequency calculation.** Using a corpus the frequency of negated terms is calculated to categorize them according to their appearance (frequency). This makes it possible to compare the most frequently used terms in Spanish in comparison to those in English.
4. **Evaluation.** The NegEx algorithm is adapted to Spanish using the list of terms previously categorized and evaluated with a set of real clininal texts.

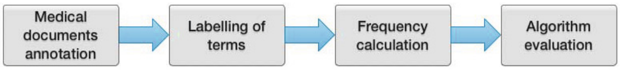


Fig. 1. Steps of the approach

In the following sections we detail these processes.

2.1 Medical Documents Annotation

In order to be able to detect negation, one of the requirements of the algorithm is to have a Gold Standard or a corpus of Spanish expressions. This is the first problem to be overcome as there is not such standard for Spanish texts. Consequently, the first step is to build such a corpus. The result of this process will be a set of annotated sentences to act as Gold Standard to test the performance of the negation detection algorithm.

2.2 Labelling of Terms

Another required process is to generate the translation to Spanish of the concepts found in the original NegEx lexicon as negation terms (see an example in table 1). Next, the lexicon has to be enriched with synonyms and clinical related common phrases in Spanish. Translation was done by a team composed by Computer Science Researchers, Physicians and Computational Linguists experts. It is interesting to note that not only terms used in [2] were translated but also the terms of the lexicon in the implementation of ConText [4] were translated.

Table 1. Example of translated terms

Term in English	Term in Spanish
can be ruled out can rule him out can rule out	se puede descartar
no, not	no
no evidence	sin evidencia no evidencia
no new	sin novedad
no support for	no hay soporte para
no suspicion of	ninguna sospecha de

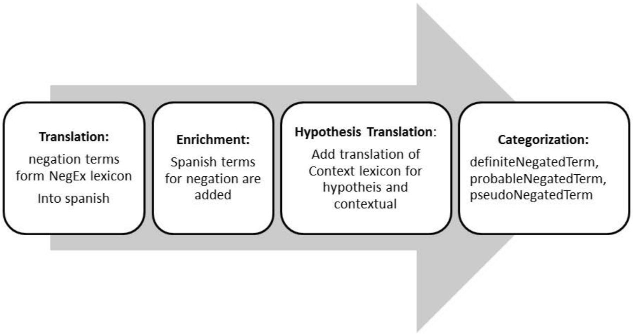


Fig. 2. Phases of the labelling process

According to [3], each translated term is assigned a certain category: i) Definite Negated Term, ii) Probable Negated Term, iii) Pseudo Negated Term.

This process of terms labelling is depicted in figure 2 and includes the depicted steps.

2.3 Frequency Calculation

Once all the sentences of the corpus are categorized, the frequency of each term of the lexicon on the corpus is calculated. The median value of the frequency for each term is also calculated and then the terms are categorized as follows:

- **No appearance:** For those terms with frequency equal to zero.
- **Infrequent:** Used to categorize those terms which frequency is greater than zero but lesser than the median value.
- **Frequent:** Those terms that appear with frequencies higher than the median value.

2.4 Evaluation

In order to validate the Spanish approach for negation detection a test dataset composed by pieces of previously (manually) annotated text written in Spanish is required. The testing corpus was generated by spliting into sentences the texts in the corpus previously described. This sentences were indexed using Apache Lucene [16]. For the indexing process the ICD-10 standard [17] is used to identify the sentences with an underlying clinical condition. After this process was completed, the set of sentences selected is manually annotated. Then the performance of the algorithm will be tested on the set of testing examples and results are shown. It is important noting that the NegEx algorithm has not been changed in its essence and only the terms used as triggers have been changed. Our aim is to show that due to the different structure of spanish language in comparison to that of english the performance of the algorithm should not be too high and consequently extensions of NegEx to deal with the morpho-sintatic struture of spanish is required to improve results. In what follows we present the results of applying the process in a set of selected texts.

3 Results

The Spanish approach has been programmed in Java, and it has been tested using 500 reports obtained from SciELO [18] as input corpus. In the extraction process the entitled sections: "Reporte de caso" (Case report), "A proposito de un caso" (About a case), "Caso clinico" (Clinical case) and similar ones were used. The resulting corpus is composed of 1,164,712 words, 65,605 out of which are different words.

In order to validate the approach the calculation of the frequency of terms for negation in Spanish texts has been performed to compare the obtained frequencies with those reported in the literature for English. In what follows we present the results obtained, in detail.

3.1 Frequency of Terms in Spanish

The frequencies shown in this section were calculated using the algorithm presented in Section 2.3. The results of the application of this algorithm are presented in the following section.

3.2 Lexical Analysis

Tables 2, 3 and 4 show the relative frequencies for terms both in Spanish and English which demonstrate Definite Negated Existence, Probable Negated Existence and Pseudo Negated Existence, respectively.

Note that the frequencies for frequent Definite and Probable Negated terms are higher in Spanish, whether Pseudo Negated terms are more frequent in English. Generally speaking, infrequent terms perform similarly in both languages,

Table 2. Definite Negated Terms frequencies in Spanish and English

Spanish (58)			English (60)		
Frequent	Infrequent	No appearance	Frequent	Infrequent	No appearance
14	20	23	4	24	32

Table 3. Probable Negated Terms frequencies in Spanish and English

Spanish (50)			English (78)		
Frequent	Infrequent	No appearance	Frequent	Infrequent	No appearance
4	17	29	3	37	38

Table 4. Pseudo Negated Terms frequencies in Spanish and English

Spanish (18)			English (16)		
Frequent	Infrequent	No appearance	Frequent	Infrequent	No appearance
1	13	4	2	13	1

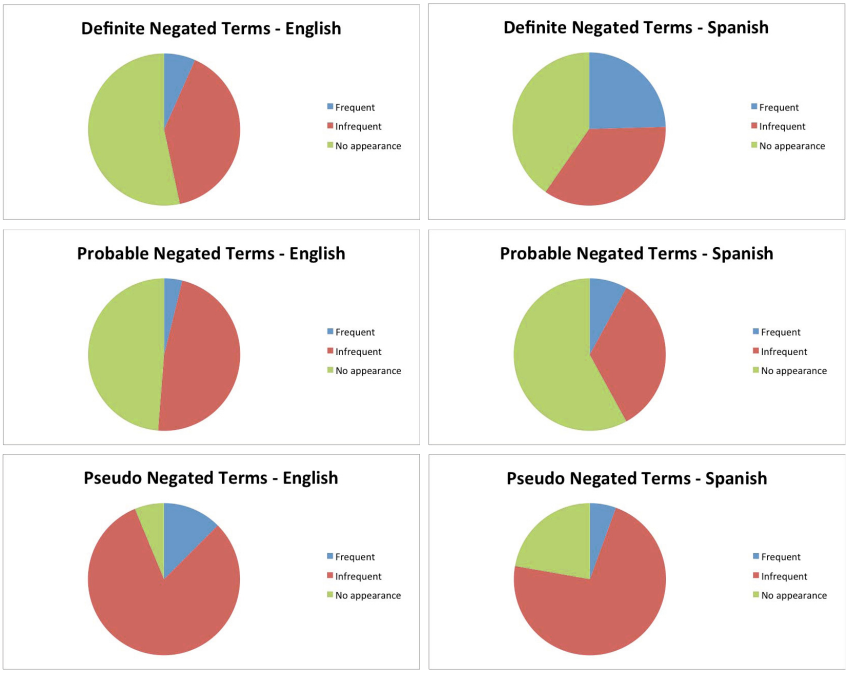
though Probable and Pseudo Negated Terms appear more frequently in English (see table 5). It is important to note that some of the terms that were obtained in the labelling process do not appear in the corpus that has been used for testing and evaluating the algorithm which may affect the quality of the results.

3.3 Validation of NegEx implementation for Spanish

The Spanish version has been tested with 500 reports where we have identified 422 different sentences and 267 unique clinical conditions. Our experiments show the following performance values: A precision of 49.47%, a recall of 55.70% and a F-Measure of 52.38% were obtained when using the Definite Negated terms as the positive set in the classification task. This process also showed an accuracy of 83.37%. When the Definite Existence terms were used as the positive set, a precision of 86.86%, a recall of 95.2% and a F-Measure of 90.84% were obtained. In this case, the accuracy is similar to the previous one with a value of 84.78%.

We observed that the number of False Positives is high and that makes the precision of the algorithm to go very low for the cases specially of negation detection. The reason behind this result was expected as we firstly showed as the structure of Spanish differs from that of English and the rules implemented to calculate the scope of negation in English not always agrees with those to analyze negation in Spanish. Consequently, future research will be done in order to adapt Negex not only by translating triggering terms but also by adding rules after a deeper analysis of the negation process in Spanish grammar.

Table 5. Comparison of frequencies in Spanish and English



4 Conclusions

In this paper we have presented an adaptation of the NegEx algorithm to be used for clinical texts written in Spanish. First, a list of terms has been identified both from the translation of those identified previoulsy in NegEx and later enriched with synonyms and terms from manual annotation of medical texts in Spanish. Second, the frequency of terms in Spanish has been calculated and compared to that of the terms in English. The differences in frequencies of the terms in both languages suggests that the corpus can be biased an should be enlarged to contain appearances of those terms that do not appear. Finally, an implementation of NegEx algorithm adapted for Spanish has been evaluated and values of accurary and recall sugest that the results can yet be improved if the scope is properly adapted. On the other hand some of the terms that were identified as negated terms did not appear in the corpus that has been used suggesting also that the corpus has to be enlarged to contain all the possible negated terms. Thus as future work we propose to generate an improved version of the Spanish corpus and to annotate more sentences to be used for evaluation with a new implementation of the algorithm. Besides as it has also been mentioned, improvement of NegEx with rules that govern the negation process in spanish should improve results obtained so far.

References

1. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G., Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: Evaluation of negation phrases in narrative clinical reports (2002)
2. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. J. Biomed. Inform. 2001, 34–301 (2001)
3. Chapman, W.W., Hillert, D., Velupillai, S., Kvist, M., Skeppstedt, M., Chapman, B.E., Conway, M., Tharp, M., Mowery, D., Deleger, L.: Extending the negex lexicon for multiple languages. In: Studies in Health Technology and Informatics, vol. 192, pp. 677–681. IOS Press (2013)
4. Harkema, H., Dowling, J.N., Thornblade, T., Chapman, W.W.: Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics 42(5), 839–851 (2009)
5. Elkin, P.L., Brown, S.H., Bauer, B.A., Husser, C.S., Carruth, W., Bergstrom, L., Wahner-Roedler, D.: A controlled trial of automated classification of negation from clinical notes. BMC Med. Inf. and Decision Making 5 (2005)
6. Mutalik, P., Deshpande, A.M., Nadkarni, P.M.: Research paper: Use of general-purpose negation detection to augment concept indexing of medical documents: A quantitative study using the umls. JAMIA 8(6), 598–609 (2001)
7. Huang, Y., Lowe, H.J.: A novel hybrid approach to automated negation detection in clinical radiology reports. Journal of the American Medical Informatics Association 14(3), 304–311 (2007)
8. Skeppstedt, M.: Negation detection in swedish clinical text: An adaption of negex to swedish. J. Biomedical Semantics 2(S-3), S3 (2011)
9. Skeppstedt, M., Dalianis, H., Nilsson, G.H.: Retrieving disorders and findings: Results using snomed ct and negex adapted for swedish. In: LOUHI 2011 Health Document Text Mining and Information Analysis 2011: Proceedings of LOUHI 2011 Third International Workshop on Health Document Text Mining and Information Analysis Bled, Slovenia, vol. (744), pp. 11–17 (2011)
10. Morante, R., Liekens, A., Daelemans, W.: Learning the scope of negation in biomedical texts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, pp. 715–724 (2008)
11. Morante, R., Daelemans, W.: A metalearning approach to processing the scope of negation. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL 2009, pp. 21–29. Association for Computational Linguistics (2009)
12. Rokach, L., Romano, R., Maimon, O.: Negation recognition in medical narrative reports. Inf. Retr. 11(6), 499–538 (2008)
13. Skeppstedt, M.: Negation detection in swedish clinical text. In: Proceedings NAACL HLT Second Louhi Workshop on Text and Data Mining of Health Documents, pp. 53–60 (2010)
14. Nilsson, G.H., Dalianis, H., Skeppstedt, M., Kvist, M.: Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. Journal of Biomedical Informatics (in press, 2014)

15. Instituto Cervantes. Electronic references (2013)
16. The Apache Foundation. Lucene. Programa de Computador (March 2000)
17. World Health Organization. International Statistical Classification of Diseases and Related Health Problems, 10th Revision. Version for 2006, ICD-10 (2006)
18. Packer, A.L.: Scielo - an electronic publishing model for developing countries. In: Smith, J., Ardö, A., Linde, P. (eds.) ELPUB. ICCC Press, Washington, DC (1999)