Annotation of negation in the IULA Spanish Clinical Record Corpus

Montserrat Marimon, Jorge Vivaldi, Núria Bel
Universitat Pompeu Fabra
Roc Boronat 138
08018 Barcelona
{montserrat.marimon|jorge.vivaldi|nuria.bel}@upf.edu

Abstract

This paper presents the IULA Spanish Clinical Record Corpus, a corpus of 3,194 sentences extracted from anonymized clinical records and manually annotated with negation markers and their scope. The corpus was conceived as a resource to support clinical text-mining systems, but it is also a useful resource for other Natural Language Processing systems handling clinical texts: automatic encoding of clinical records, diagnosis support, term extraction, among others, as well as for the study of clinical texts. The corpus is publicly available with a CC-BY-SA 3.0 license.

1 Introduction

With the deployment of Electronic Health Records (EHR), much effort is being devoted to the development of text-mining tools that assist in converting information described in texts into structured data for applications that range from assisting in medical diagnosis to the coding of clinical findings and procedures to bill insurance companies. The ultimate objective of these tools is the extraction of factual knowledge from textual data. Therefore, they are mainly interested in developing special components that identify those facts that do not hold true, as in *patient without nodules*. The availability of annotated texts makes the use of supervised machine learning methods possible, and it also allows for a fair comparison and evaluation of different methods, thus contributing to the improvement of the technology.

In what follows, we describe the IULA Spanish Clinical Record Corpus (IULA-SCRC), a corpus of 3,194 sentences extracted from anonymized clinical records and manually annotated with negation markers and their scope, and the corresponding annotation guidelines.[1] To the best of our knowledge, this is the first corpus of medical Spanish texts manually annotated for negation, although two test-sets of about 500 and 1000 sentences for evaluating particular negation detection systems already exist, as described later in the Related Work section.

Because no standard negation annotation schema still exists, our annotation schema has taken into account the currently existing English corpora annotated for negation, trying to be comprehensive with current practices (Mutalik et al., 2001; Szarvas et al., 2008; Morante and Daelemans, 2012).

After this introductory section, in Section 2 we briefly describe negation structures in Spanish, in Section 3 we describe the corpus design, in Section 4 we present the guidelines we have followed to identify and classify negation information and in Section 5 we provide details of tags and statistics of the resulting annotated corpus, then, in Section 6 we review existing related corpora on which we have designed our annotation schema and, finally, in Section 7 we conclude.

2 Negation in Spanish

The most prominent negation marker in sentential negation in Spanish is the pre-verbal adverb *no* (1).

(1) *Juan* ***no*** *come carne.* (Juan does not eat meat.)

Scope is the part of the sentence that is affected by a preceding negation marker that syntactically dominates it. Most frequently, sentential negation is expressed with a negation marker that scopes

[1] The corpus described in this paper has been made publicly available for research purposes and it is freely downloadable from: http://eines.iula.upf.edu/brat/\#/NegationOnCR_IULA

Proceedings of the Workshop Computational Semantics Beyond Events and Roles (SemBEaR), pages 43–52,
Valencia, Spain, April 4, 2017. ©2017 Association for Computational Linguistics

over the verb phrase. However, scope may also correspond to non-verbal phrases, as in (2), where the negation marker scopes over the adverb *siempre*.

(2) *Juan* ***no*** *siempre come carne.* (Juan does not always eat meat.)

In addition to the adverb *no*, there is a fairly heterogeneous group of pre-verbal words which also express sentential negation (3). These negation markers are: the pronouns *nada* (nothing) and *nadie* (nobody); the determinant *ninguno* (none); the adverbs *nunca, jamás* (never), *tampoco* (neither) and *nada* (nothing); and the phrases introduced by the coordination particle *ni* (nor).

(3) ***Nadie*** *ha venido.* (Nobody has come.)

Examples in (4) show a second pattern where these negation words follow the verb. In this position, they require a negation preceding the verb.

(4) (a) ***No*** *ha venido* ***nadie***. (Nobody has come.)
(b) **Ha venido* ***nadie***. (Has come nobody.)

In this structure we distinguish two groups of elements: a negative inducer and a negative polarity item. The first one allows the presence of the second one in post-verbal position.

Negative polarity items (NPIs) include: post-verbal negation words, indefinite NPs (5.a), and aspectual and scalar NPIs (5.b). Negative inducers (NIs) include: rhetorical interrogatives; comparative and superlative constructions (5.c); adverbial and nominal quantifiers (5.d); negative adverbs; negative verbs, nouns, and adjectives expressing doubt, opposition, deprivation or absence, or emotive factives (5.e); the conjunction *ni* (neiter); and the preposition *sin* (without).

(5) (a) *Juan* ***apenas*** *lee libro* ***alguno***. (Juan hardly reads any books.)
(b) *Esto* ***no*** *vale* ***ni un pimiento***. (This is not worth a light.)
(c) *Juan es* ***más listo*** *que* ***nadie***. (Juan is smarter than anyone.)
(d) *Este examen es* ***demasido*** *difícil* ***para*** *que lo apruebe* ***nadie***. (This test is too difficult for anyone to approve.)
(e) *Es* ***improbable*** *que haya estado* ***nunca*** *en mi casa.* (It's unlikely she/he's ever been in my house.)

In addition to sentences and phrases, in Spanish single words can also be denied with the adverb *no* and by prefixation. In word negation, prefixes that express absence, opposition, falsehood, reversal, deprivation or removal, such as *a-, anti-* and *des-*, as in *amoral* (amoral), *anticapitalista* (anticapitalist), and *desleal* (disloyal) are used. Other negative prefixes in Spanish are: *in-, sin-*, and *contra-*.

Finally, coordination and enumeration of negated words or phrases is also possible. In these structures, the first element follows the rules we have just presented, and the following coordinated elements can be preceded or not by the conjunction *ni*, but the last element must include the negative conjunction.

3 Corpus description

The basic material for compiling this resource was obtained from a set of 300 clinical reports from several services of one of the main hospitals in Barcelona (Spain). These reports were delivered to us already anonymized. After a first examination of these reports, it was observed that there was a set of 17 sections (e.g. ”Physical Examination”, ”Diagnostic”, ”Procedures”, ”Reasons for consultation”,...) that appeared in most of these reports. To compile the corpus only the five sections with more data were considered. In Table 1 we show the final number of sentences chosen from each section. Up to 3,000 sentences from these sections were separately collected and shuffled in order to make sure that no traceability of personal data was possible.

It is normal practice for automatic processing of clinical records to work with correct texts (Lai et al., 2015), thus, a simple set of regular expressions was used to correct most common misspellings. Remaining misspellings were manually corrected. Before annotating these reports, they were preprocessed for sentence identification.

Section	Sent.	%	Chosen
Physical exploration	5,193	34.61	1,090
Evolution	5,463	36.41	1,147
Radiology	1,751	11.67	367
Current process	980	6.53	205
Comp. explorations	1,619	10.79	339

Table 1: Statistics about corpus composition.

4 Annotation guidelines

In this section, we first introduce the underlying general annotation criteria. Second, we describe the guidelines we have followed to identify negation cues and their scopes.[2] Finally, we present the different classes of medical terms we have identified.

4.1 Underlying criteria

Our approach for annotating negation aims at supporting automatic processing for information extraction, which is usually supported by a dictionary coming either from a medical database or from a Named Entities recognition system. Information extraction systems are usually designed for extracting relations among entities. Ultimately, they are used to extract ”facts”. The presence of a negation marker might change the status of what a fact is.

Accordingly, in our annotation, first, negation markers are lexically defined: they are a list of words that change the factual status of what follows them, i.e. the scope. Second, we encode negation scope on syntactic terms: it is the maximal syntactic unit that is affected by the negation marker. However, as we will describe below, there are linguistic phenomena that escape from these general statements.

We annotate as negation markers only those negation words that affect the assertion made by other words in the sentence, because they change its factual status. This is the case, for instance, of the adverb *no* and some negative predicates, such as *ausencia de* (absence of).

However, we do not consider as negation markers those predicates that bear more information than bare negation. We discard verbs like *desaparecer* (to disappear), which indeed contains the information of a change of state. Other examples of predicates which are not considered negation markers are the verbs *retirar, suspenderse,* and *erradicar* (to remove, to call off, and to eradicate), and the noun *retirada* (removal). Also note that we do not consider the verb *negar* (to deny), as in *el paciente niega síntomas de abstinencia* (the patient denies withdrawal symptoms), a negation cue either, since, following clinician expert's advice, this communication verb is considered, in factual terms, an statement of what someone says.

As for terms like *asintomático* (asymptomatic), which shows morphological prefixation (*a*-, *des*-, *dis*-), we decided to follow the current practice in medical text annotation for automatic processing (see Table 5) and not to annotate them as negation markers. Besides the fact that it is normal practice, we have considered the following motivations.[3] First, negative prefixed terms in Spanish medical domain are mostly lexicalized and most of them can easily be found in existing medical term databases. Second, most of them, in particular nouns, are coined terms, as they have a different specialized meaning from that of the non-prefixed counterpart and a different meaning, too, from the bare negation of the positive term, for instance *deshitratación* (dehydration) and *no hitratación* (no hydration) or *degeneración* (degeneration) vs. *no generación* (no generation). Third, not all prefixed words can be compositionally analyzed, as the non-prefixed counterpart does not exist (Dzuganova, 2006), *a-febril* (afebrile) vs. *a-morfo* (amorphous) or *ex-cluir* (exclude), for instance. Finally, prefixed words, as full words, can be in the scope of another negation marker. The interpretation of a double negation in these cases is uncertain, consider, for instance, *non-atypical hyperplasia* or *no mitral valve insufficiency*.

4.2 Negation cues

In our corpus, negation cues are words that express negation: adverbs, negative predicates, and the preposition *sin*. Examples in (6) show negations expressed by the preposition.

(6) (a) ***Sin*** *soplos audibles* (Without audible X); ***sin*** *signos de TVP* (without signs of X).

(b) ***Sin*** *que se observen claros defectos de ventilación* (With no clear X observed).

The most frequent negative adverb in the corpus is the adverb *no*. This adverb negates verbal forms (7.a), nouns (7.b), and adjectives (7.c).

(7) (a) ***No*** *ausculto soplos* (I don't auscultate X); ***no*** *se palpan masas* (X are not palpated).

(b) ***No*** *edemas en extremidades inferiores* (No X in lower extremities).

[2] In the examples we provide, cues are marked in bold and their scopes are underlined; in the next section, we present the actual tags we have used in the corpus. Also note that in the translated examples, medical terms are not translated, but they are replaced by ”X”.

[3] Note that in Spanish there are no negative suffixes like the English *less*.

(c) *Temblor discal* ***no*** *continuo en mano izquierda* (No continuous X in left hand).

Another negative adverb that we find in the corpus is *tampoco*. This adverb only negates verbal forms, as in (8).

(8) ***Tampoco*** *objetiva focos sépticos* (Neither objectify X).

We also mark as negation cues the following predicates: the verb *descartar* (to rule out) (9.a), the noun *ausencia de* (absence of) (9.b),[4] and the adjective *incapaz de* (unable to) (9.c).

(9) (a) *Se* ***descarta*** *enolismo* (X is ruled out).
(b) ***Ausencia de*** *edemas* (Absence of X).
(c) ***Incapaz de*** *levantarse de la silla* (Unable to get up from the chair).

The adjective *negativo* (negative), which is very frequent in medical texts, expresses negation in different ways. It may deny a sign, indicating on physical examination that a finding is not present (10.a); or it may deny a laboratory test, indicating that a substance or a reaction is not present (10.b). Sometimes, even though it clearly expresses negation, the specific bacteria or organism the cultures are negative for is not explicitly said in the sentence (10.c). We even have some examples where the negated test or sign is not expressed in the sentence: *negativo* is neither followed nor preceded by the noun it modifies. Thus, the adjective *negativo* is always marked as cue, even when its scope is not present in the sentence.

(10) (a) *Murphy* ***negativo*** (X negative).
(b) *Serologías VHB y VHC* ***negativos*** (X negative).
(c) *Hemocultivos de control* ***negativos*** (X negative).

Negative polarity items (11) (cf. Section 2) are also annotated as such. Note that the most frequent case is coordination.

(11) (a) ***No*** *objetivando* ***ninguna*** *focalidad neurológica mayor inmediata* (Not objectifying any inmediate main X).
(b) ***No*** *masas* ***ni*** *megalias* (Neither X or Y); ***sin*** *soplos* ***ni*** *roces* (without X or Y).

Double negation sentences (12.a), in which two negatives yield affirmative, are not marked. Note that example (12.b) is not a false negative, since *desaparecer* (to disappear) is not considered a negation marker.

(12) (a) ***No*** *se puede* ***descartar*** *la etiología epiléptica de los episodios* (X can not be ruled out).
(b) ***Sin*** *llegar a desaparecer del todo* (Without disappearing altogether).

4.3 Scope

Traditionally, scope is the part of the sentence that is being negated. The scope is determined on the basis of syntax: the maximal syntactic phrase that is affected by the marker. In our corpus, the negation cue is not included in its own scope.

As we show in (14), the scope of negated nouns extends to their complements and/or modifiers that follow them (14.a); the scope of negated adjectives extends to their complements, but the modified noun that in Spanish precedes the adjective is not annotated as scope (14.b); and the scope of negated verbs includes every verb dependent that follows it, and, we show in (14.c), constituents that precede the verb are not annotated as scope. This decision affects, in particular, verb subjects, which are however annotated in the scope when they are located after the verb (as in Bioscope). The only exception to this rule is when there is an unaccusative verb, for which we also annotate the subject, as we will see in example (19.d) below.

(14) (a) ***No*** *edemas en extremidades inferiores* (No X in lower extremities).
(b) *Temblor discal* ***no*** *continuo en mano izquierda* (No continuous X in left hand).
(c) *El estudio realizado de forma ambulatoria hasta el momento* ***no*** *mostró alteraciones significativas* (The study performed on an outpatient basis so far showed no significant alterations).

The preposition *sin* has a scope over the following noun phrase (15.a) and verb phrase (15.b) and, as before, all modifiers and complements of the nominal and verbal heads are included.

[4] Note that the cue in (9.b) includes both the noun and the preposition, and that the cue in (9.c) includes the adjective and the preposition.

(15) (a) ***Sin*** *signos de TVP* (Without signs of X); ***sin*** *contraindicaciones para el procedimiento* (Without contraindications to the procedure).
(b) ***Sin*** *objetivar trombosis* (Without objectifying X), ***sin*** *que se observen claros defectos de ventilación* (With no clear X observed).

Negative predicate cues scope over their complements (16).

(16) (a) *Se* ***descarta*** *enolismo* (X is ruled out).
(b) ***Ausencia de*** *edemas* (Absence of X))
(c) ***Incapaz de*** *levantarse de la silla* (Unable to get up from the chair).

Because of its special characteristics already explained in section 4.2, the adjective *negativo* scopes over its modified noun, which precedes it (17.a) or over the PP that includes the denied test or sign (17.b). When this adjective functions as an attribute or a predicative complement (17.c), the scope is the subject. Finally, when the subject is a relative pronoun, we annotate as the scope its antecedent (17.d).

(17) (a) *Focalidad* ***negativa*** (X negative).
(b) ***Negativo*** *para FOP* (Negative for X).
(c) *Las serologías para VIH, VHB y VHC resultaron* ***negativas*** (X were negative). *El urocultivo es* ***negativo*** (X is negative).
(d) *Se tomó urocultivo, que resultó* ***negativo*** (It was taken X, which was negative).

In coordination, the cue scopes over all coordinated elements (18).

(18) (a) ***No*** *masas* ***ni*** *megalias* (Neither X nor Y).
(b) ***Sin*** *soplos* ***ni*** *roces* (Without X or Y).
(c) ***No*** *refiere síndrome miccional, cambios en el ritmo deposicional* ***ni*** *otra sintomatología acompanante* (Does nor refer X, or Y, or Z).

Discontinuous scopes are also annotated. These are examples like (19.a), where the adjective appears between the noun and its modifier, elliptical constructions, such as (19.b), relative clauses, where the antecedent of the relative pronoun is also annotated as discontinuous scope (19.c), and unaccusative verbs, whose subject is also included in the scope of the negation cue, even though it precedes the verb (19.d).

(19) (a) *Hemocultivos* ***negativos*** *de control* (Control negative X).
(b) *Parcialmente orientado (sí en tiempo y persona,* ***no*** *en espacio)* (Partially oriented (yes in time and person, not in space)).
(c) *Se trató con antibiótico que* ***no*** *recuerda* (S/he was treated with antibiotics which s/he does not remember).
(d) *El dolor* ***no*** *mejorado con nolotil* (Pain has not improved with nolotil).

4.4 Medical term classes

Most of the cues that are present in the corpus scope over medical named entities. Table 2 shows the classes we have distinguished among these entities. In the next section we will present the actual tags we have used for manually annotating them in the corpus.

Class	Used for
Body structure	- Anatomical structure - Body part - Organ or organ component - Deformity - Tissue, ...
Substance & Pharmacological/ biological product	- Pharmacological substance - Biological substance - Enzyme - Body substance - Diagnostic substance, ...
Clinical finding	- Disease or syndrome - Finding - Sign/Symptom - Abnormality - Clinical state, ...
Procedure	- Diagnostic procedure - Laboratory procedure - Therapeutic Procedure - Administration of medicine - Health care activity, ...

Table 2: Medical term classes.

This classification was taken from the SNOMED Clinical Terms (SNOMED CT), a multilingual clinical healthcare terminology used in clinical documentation.[5] This resource defines 19 top level hierarchies (or classes), we have chosen five of them which are the most frequent classes found in this type of reports. For operative reasons we collapse SNOMED classes ”Substance” and ”Pharmacological/biological product” in a single medical term class.

5 Corpus Annotation

The annotation was made with Brat, a web-based tool for text annotation.[6] In this section, we present the actual tags we have used in the annotation. We also discuss the annotation agreement and provide some statistics of the corpus.

5.1 Tags

In Table 3 we show the list of tags that are used to mark negation cues and the text spans that function as scope. In addition, tagged links are used to describe the relationships between them: we use the tag Scope to link scopes to negation cues (Figure 1 and Figure 2) and the tag DiscScope to annotate discontinuous scope phenomena explicitly (Figure 3).

Tag	Entity
Negmarker	*no, tampoco, sin*
NegPredMarker	negative verbs, nouns and adjectives
NegPolItem	*ni, ninguno,...*
BODY	body structure
SUBS	substance...
DISO	clinical finding
PROC	procedure
Phrase	nonmedical text spans

Table 3: Tags for entities.

Negation cues are marked by two different tags. We use the tag NegMarker for basic negation markers (Figure 1.a-c): the adverbs *no* and *tampoco*, and the preposition *sin*. We use the tag NegPredMarker for negative verbs, nouns, and adjectives (Figure 1.d-f). In addition, the tag NegPolItem (Figure 1.b-c) is used for NPIs.

We have used four tags for the medical named entities (see Section 4.4) that are in the scope of a negation marker: BODY for body structures, SUBS for substances and pharmacological/biological products, DISO for clinical findings (Figure 1.c-e), and PROC for medical procedures (Figure 1.b). In addition, we use the tag Phrase for:

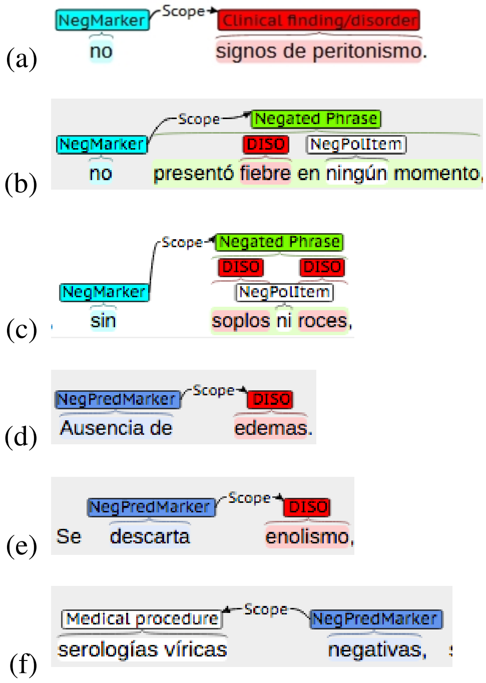

Figure 1: Annotation examples: tags for negation cues.

- Negated phrases that are not of the medical domain (Figure 2.a).
- Text spans that are not headed by an entity belonging to one of the medical classes we have considered (Figure 2.b).
- Complete coordinated phrases (Figure 2.c-e).

5.2 Agreement analysis

In order to evaluate the guidelines, 500 sentences were annotated by three computational linguists advised by a clinician. Disagreements were discussed after three different annotation rounds until reaching a consensus. Annotation guidelines were updated accordingly. Then, we measured the consistency of the annotations for the negation markers and their scope, but not of the entities annotations which were validated using SNOMED. The inter-annotator agreement Kappa rates were 0.85 between annotators 1 and 2, and 1 and 3; 0.88 between annotators 2 and 3.

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

[6] http://brat.nlplab.org/

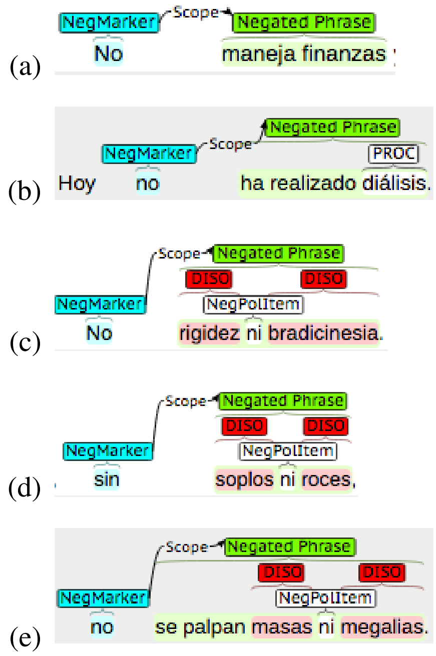

Figure 2: Annotation examples: tags for scope entities.

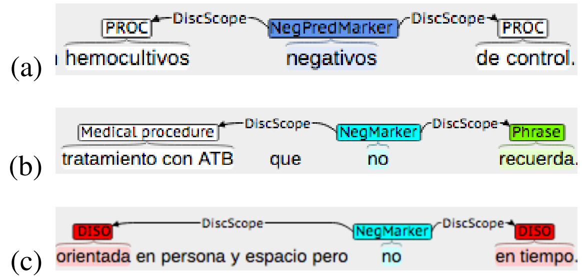

Figure 3: Annotation examples: discontinuous scopes.

5.3 Corpus statistics

Final annotated corpus details are in Table 4. The most frequent tag for cues is `Negmarker`, which appears 1,007 times (519 marking the adverb *no* and 488 marking the preposition *sin*). The most frequent NPI is *ni*, which appears 109 times, whereas the most frequent negative predicate is *negativo*, which appears 63 times.

6 Related work

Most of existing corpora in the biomedical domain annotated with negation have been developed as test sets of systems to detect negated expressions. Most of these resources show a common set of annotations (see Table 5). All annotate negation markers and their scope. Negative predicates are annotated by most of them, but each one considers a different list of predicates. None annotates morphological-related negation phenomena (prefixes or suffixes). In general, discontinuous scope is not taken into account. Finally, no one annotates the actual negation marker within the scope. Now, we briefly describe the most salient characteristics of each system and resulting annotation.

Number of sentences	3,194
Number of annotated sentences	1,093
Number of `Negmarker` entities	1,007
Number of `NegPredMarker` entities	86
Number of `NegPolItem` entities	114
Number of `BODY` entities	7
Number of `SUBS` entities	14
Number of `DISO` entities	1,064
Number of `PROC` entities	93
Number of `Phrase` entities	278

Table 4: Corpus statistics.

Negfinder by (Mutalik et al., 2001) uses a lexical scanner with regular expressions to identify negation and a context-free grammar parser to associate negation markers to their scope. In the testset only bare negative words are annotated, while words (medical terms) whose meaning is change of state, e.g. *stopping* or *discontinuing a drug*, are not annotated, nor are medical terms having a negative prefix (*akinesia*).

Chapman et al. (2001) developed NegEx, a simple regular expression-based algorithm to determine whether a finding or disease mentioned within medical reports was present or absent. NegEx implements (up to 35) negative and pseudo-negative phrases, limits their scope and rules out sentences having double negation. There are different versions of NegEx (South et al., 2007; Harkema et al., 2009), and it has been adapted to Swedish (Skeppstedt, 2011), French (Deléger and Grouin, 2012), Dutch (Afzal et al., 2014), and Spanish (Costumero et al., 2014). In addition, the systems developed by Sohn et al. (2012) (DepNeg) and Mehrabi et al. (2015) (DEEPEN) are based on or use NegEx complemented with a dependency-based parser to improve scope detection. And, in another line of research, Goldin and Chapman (2003) use Naive Bayes and Decision Trees to increase the NegEx's precision of negation with only the word ”not”. In these NegEx-based systems, negative predicates such as *denies,*

Corpus/System	Language	Technique	Uncertainty	Basic Negation	Morph. Neg.	Negative Pred.	Disc. Scope	Focus
Negfinder	EN	lexical scanner + CFG	no	yes	no	yes	no	no
NegEx	EN,ES GE,SW	Regular Expression pattern matching	no	yes	no	yes	no	no
DepNeg & DEEPEN	EN	dependency parsing	no	yes	no	yes	no	no
Goldin & Chapman's	EN	machine learning NB & DT	no	yes	no	no	no	no
Cotik et al.'s	ES	rules PoS-tag NegEx & ST	no	yes	no	yes	no	no
NegHunter	EN	rules based on grammatical info	no	yes	no	yes	no	no
Elkin et al.'s	EN	negation ontology	yes	yes	no	yes	no	no
BioScope	EN	manual	yes	yes	no	yes	no	no
BioInfer	EN	manual	no	yes	no	no	no	no
IULA-SCRC	ES	manual	no	yes	no	yes	yes	no

Table 5: Comparison of different proposals to negation annotation in the biomedical domain.

declines and *no complaints of* are annotated.

Cotik et al. (2016) developed syntactic techniques based in rules derived from PoS tagging patterns, constituent tree patterns and dependency tree patterns, and an adaptation of NegEx, to determine if a medical term is under the scope of negation in radiology reports written in Spanish. Since they translate the Negative predicates provided by the NEgEx tool, these are included in the test-set.

Another rule-based negation algorithm is NegHunter, developed by Gindl et al. (2008), which uses grammatic information such as tense and part-of-speech to detect negation in clinical practice guidelines lexically marked by adverbs, prepositions and a few predicates (*absence, freedom, deny, decline* and *lack*).

Finally, Elkin et al. (2005) developed a mechanism for automated annotation of negation of clinical concepts invoking an ontology. Negative predicates are annotated, including the verb *to deny*.

As for other annotated biomedical corpora, the following resources have been developed with explicit aim of somehow annotating negation. In general, they annotate more cases of negation than the test-sets just reviewed. In what follows we review their most salient characteristics.[7]

The BioScope corpus (Szarvas et al., 2008) gathers medical and biological texts (20,879 sentences) annotated for negation cues, speculation and their linguistic scope. The minimal unit that expresses negation is marked as cue and its scope is extended to the largest syntactic phrase. The scope includes the negation cue, and leaves the subject out, but only in active sentences.

The BioInfer corpus (Pyysalo et al., 2007) contains 1,100 sentences from abstracts of research articles where biomedical relations are annotated for negation.

The 2010 i2b2/VA NLP Challenge Corpus (information extracted from (Wu et al., 2014)) contains 871 de-identified reports from different hospitals and medical centers. Negation as such is not annotated, but each medical term is associated with different tags, one of these being ”absent” which seems to match with what others consider negated expressions. This annotation includes also what we have called morphology-related and inherently negated terms such as *afebrile*.

Finally, the BioNLP Genia Event Extraction Corpus (Kim et al., 2008) is frequently mentioned

[7] Some of these corpus are only found in the literature and are not publicly available.

in the related literature. However, although a negation attribute is mentioned at event level, cues and their scope are not annotated.

7 Conclusions

In this article, we have introduced the annotation guidelines of the IULA Spanish Clinical Record Corpus annotated for negation. We have described the underlying criteria and we have motivated the choice of a syntactically-based general criterion, as well as the relation of our annotation schema with other negation-annotated corpora already available, although all but one are for English. The corpus currently contains about 3000 sentences and it is licensed with Creative Commons 3.0 CC-BY-SA license. This resource has been developed for supporting text-mining systems either to serve as a test set for rule based systems or as training data for machine learning based systems. Nevertheless, it is also a good resource for the study of clinical texts.

Acknowledgments

We want to acknowledge the support of Dra. Pilar Bel-Rafecas, clinician, and the comments and suggestions of the two anonymous reviewers that have contributed to improve the final version of this paper. This work was partially supported by the project TUNER (TIN2015-65308-C5-1-R, MINECO/FEDER)

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. *BMC Bioinformatics*, 15(373):1–12.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

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. Springer International Publishing.

Viviana Cotik, Vanesa Stricker, Jorge Vivaldi, and Horacio Rodríguez. 2016. Syntactic methods for negation detection in radiology reports in spanish. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing (BIONLP 16)*, pages 156–166. Association for Computational Linguistics, Berlin, Germany.

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.

Bozena Dzuganova. 2006. *Bratislavske Lekarske Listy*, 107(8):332–335.

Peter L. Elkin, Steven H. Brown, Brent A. Bauer, Casey S. Husser, William Carruth, Larry R. Bergstrom, and Dietlind L. Wahner-Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC Medical Informatics and Decision Making*, 5(13).

Stefan Gindl, Katharina Kaiser, and Silvia Miksch. 2008. Syntactical Negation Detection in Clinical Practice Guidelines. *Studies in Health Technology and Informatics*, 136:187–192.

Ilya M. Goldin and Wendy W. Chapman. 2003. Learning to detect negation with 'not' in medical texts. In *Proceedings of the Workshop on Text Analysis and Search for Bioinformatics at the 26th Annual International ACM SIGIR Conference (SIGIR-2003).*

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.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(10).

Kenneth H. Lai, Maxim Topaz, Foster R. Goss, and Li Zhou. 2015. Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics*, 55(C):188–195.

Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M. Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul R. Dexter, C. Max Schmidt, Hongfang Liu, and Mathew J. Palakal. 2015. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *Journal of Biomedical Informatics*, 54:213–219.

Roser Morante and Walter Daelemans. 2012. Conan doyleneg: Annotation of negation in conan doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 1563–1568, Istanbul, Turkey.

Pradeep G. Mutalik, Aniruddha Deshpande, and Prakash M. Nadkarni. 2001. Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents: A Quantitative Study Using the UMLS. *Journal of the American Medical Informatics Association*, 8(6):598–609.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Bjorne, Jorma Boberg, Jouni Jarvinen, and Tapio Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(50).

Maria Skeppstedt. 2011. Negation Detection in Swedish Clinical Text: An Adaption of NegEx to Swedish. *Journal of Biomedical Semantics*, 2(3):1–12.

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

Brett R. South, Shobha Phansalkar, Ashwin Deepak Swaminathan, Sylvain Delisle, Trish Perl, and Matthew H. Samore. 2007. Adaptation of the NegEx algorithm to Veterans Affairs electronic text notes for detection of influenza-like illness (ILI). In *AMIA Symposium*, pages 11–18. American Medical Informatics Association.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The BioScope Corpus: Annotation for Negation, Uncertainty and Their Scope in Biomedical Texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45, Stroudsburg, PA, USA.

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, and David Carrell. 2014. Negation's Not Solved: Generalizability Versus Optimizability in Clinical Natural Language Processing. *PLoS ONE*, 9(11).