A Machine-Learning Approach to Negation and Speculation Detection in Clinical Texts

Noa P. Cruz Díaz
Department of Information Technology, University of Huelva, Huelva, Spain. E-mail: noa.cruz@dti.uhu.es

Manuel J. Maña López
Department of Information Technology, University of Huelva, Huelva, Spain. E-mail: manuel.mana@dti.uhu.es

Jacinto Mata Vázquez
Department of Information Technology, University of Huelva, Huelva, Spain. E-mail: jacinto.mata@dti.uhu.es

Victoria Pachón Álvarez
Department of Information Technology, University of Huelva, Huelva, Spain. E-mail: victoria.pachon@dti.uhu.es

Detecting negative and speculative information is essential in most biomedical text-mining tasks where these language forms are used to express impressions, hypotheses, or explanations of experimental results. Our research is focused on developing a system based on machine-learning techniques that identifies negation and speculation signals and their scope in clinical texts. The proposed system works in two consecutive phases: first, a classifier decides whether each token in a sentence is a negation/speculation signal or not. Then another classifier determines, at sentence level, the tokens which are affected by the signals previously identified. The system was trained and evaluated on the clinical texts of the BioScope corpus, a freely available resource consisting of medical and biological texts: full-length articles, scientific abstracts, and clinical reports. The results obtained by our system were compared with those of two different systems, one based on regular expressions and the other based on machine learning. Our system's results outperformed the results obtained by these two systems. In the signal detection task, the *F*-score value was 97.3% in negation and 94.9% in speculation. In the scope-finding task, a token was correctly classified if it had been properly identified as being inside or outside the scope of all the negation signals present in the sentence. Our proposal showed an *F* score of 93.2% in negation and 80.9% in speculation. Additionally, the percentage of correct scopes (those with all their tokens correctly classified) was evaluated obtaining *F* scores of 90.9% in negation and 71.9% in speculation.

Received March 2, 2011; revised February 23, 2012; accepted February 23, 2012

 • Published online 31 May 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.22679

Introduction

Medical practitioners are increasingly incorporating results and findings from clinical studies into their work. The availability of vast databases of scientific articles allows access to this material, although the huge volume also makes it difficult to locate relevant material.

Furthermore, some hospitals have electronic records of their patients' medical background and many others are proceeding to digitize records. This enables physicians to carry out clinical studies which allow progress in evidence-based medicine. However, as in the case of access to scientific information, physicians need to have efficient tools to access this information.

The advanced information access or text-mining tools referenced herein cannot be based on a simple approach such as a bag of words. It is necessary to analyze the text in greater depth. This analysis should include negation and speculation detection. This is especially necessary in the biomedical domain, where, depending on the type of documents, the number of speculative or negated sentences varies from 13–20% (Vince, Szarvas, Farkas, Móra, & Csirik, 2008).

Negation transforms an affirmative sentence into a negative one, such as "Neither pneumothorax nor fractures were appreciated." Speculation is used to express that some fact is not known with certainty, e.g., "may correspond to an incipient pneumonic process."

Both negation and speculation detection can be divided into two phases. In the first, the expressions that indicate negation/speculation are identified. In the second, the scope of the expressions are determined, i.e., which words are

affected by an expression of negation or speculation. In this article, we focus on both phases in a collection of clinical documents.

The rest of this article is organized as follows. First, we present the most relevant work related to the detection of negation/speculation, whether based on regular expressions or machine learning. Then we describe the architecture of the system implemented to solve the task of detecting negation and speculation. We proceed to examine the collection of Bioscope documents, focusing on the clinical texts used for training and evaluation of the proposed system, and describe the set of features used to represent each word of the text. The evaluation framework is detailed next, and our results presented and discussed. We conclude and note our future work.

Background

Negation detection has gained much attention in recent years, especially in the medical domain. Process negation can be useful for several natural language processing (NLP) applications such as information extraction, opinion mining, sentiment analysis, and paraphrasing and recognizing textual entailment. For example, many authors have studied the role of negation in the sentiment analysis task, which consists of identifying the opinions expressed in a text and classifying texts accordingly. Councill, McDonald, and Velikovich (2010) defined a system that can identify exactly the scope of negation in free text. Their system achieved an 80.0% *F* score; the authors concluded that, as they expected, performance was improved dramatically by introducing negation scope detection. In a more recent work, Dadvar, Hauff, and de Jong (2011) investigated the problem of determining the polarity of sentiments in movie reviews when negation words, such as "not" and "hardly," occur in the sentences. The authors observed significant improvements in the classification of the documents after applying negation detection. These works confirm how important it is to apply negation detection in different biomedical domains.

In the biomedical domain, which is the focus of our work, there are many examples that show how negation detection is essential to treat NLP, text-mining, and information retrieval problems. For example, Averbuch, Karson, Ben-Ami, Maimon, and Rokach (2004) included negation detection in the task of context-sensitive medical information retrieval. The authors explained that the context of negation, a negative finding, is of special relevance because many of the most frequently described findings are those denied by the patient or subsequently "ruled out." Hence, if negation is not taken into account in this task, many of the retrieved documents will be irrelevant. Denny, Miller, Waitman, Arrieta, and Peterson (2008) identified QT prolongation from electrocardiogram (ECG) impressions using a general-purpose natural language processor. In this work, the authors applied a modified version of the NegEx algorithm to identify the negation. The authors asserted that NLP with negation detection can extract concepts from ECG impressions with high accuracy. Most recently, Denny et al. (2012) investigated how NLP improves identification of colorectal cancer (CRC) testing in an electronic medical record. As part of its natural language processing, they identified unified medical language system (UMLS) concepts found in each sentence along with information of its relevant context as well as information about whether or not the concept is negated. Also, an algorithm identified negated phrases as well as common verbs and other modifiers that change the status of CRC-related testing (e.g., refused, declined). The results showed that applying NLP to an electronic health record detected more CRC tests than either manual chart review or billing records review alone.

Previous studies on the problem of negation detection have been based on two different approaches: regular expressions and machine-learning techniques.

Among the investigations based on regular expressions, that carried out by Chapman, Bridewell, Hanbury, Cooper, and Buchanan (2001) is worth noting. Their algorithm, NegEx, determined whether a finding or disease mentioned within narrative medical reports was present or absent, as well as the scope of the sentence that was affected by the negation. The identification of the scope was done at the UMLS concept level. The algorithm was defined by the authors themselves as simple, but it has proven to be powerful in negation detection in discharge summaries. Because of this potential, we compared our results with those obtained by NegEx on the same document collection. Although the results were not directly comparable, the evaluation demonstrated the superiority of our system. The reported results of NegEx showed a positive predictive value (PPV or precision) of 84.5%, sensitivity (or recall) of 77.8%, and a specificity of 94.5%. However, when NegEx was applied to a set of documents from a different domain than that for which it was conceived, the overall precision was lower (Mitchell et al., 2004). Perhaps that is one of the explanations as to why NegEx obtained worse results than our system.

Other interesting research work based on regular expressions is that of Mutalik, Deshpande, and Nadkarni (2001) and Elkin et al. (2005). Huang and Lowe (2007) were aware that negated terms may be difficult to identify if negation signals are more than a few words away from them. To address this limitation in automatically detecting negations in clinical radiology reports, they proposed a novel hybrid approach, combining regular expression with grammatical parsing. The sensitivity of negation detection was 92.6%, the PPV was 98.6%, and the specificity was 99.8%.

However, the most recent works in the field of negation detection are based on a machine-learning approach. Examples of detecting negated concepts in medical narrative using machine-learning techniques are the research by Averbuch et al. (2004) and Goldin and Chapman (2003).

Here, we describe the research conducted by Morante, Liekens, and Daelemans (2008) because it is among the most recent and, as in our case, they used the Bioscope

DOI: 10.1002/asi

corpus. Their machine-learning system consisted of two classifiers. The first one decided if the tokens in a sentence were negation signals. The second determined which words in the sentence were affected by the negation. They applied postprocessing to increase the number of fully correct scopes. With this approach, the algorithm showed an *F* score of 80.99% and 50.05% of scopes correctly identified. The system architecture used by the authors was the same as in our system; however, with a different classifier and set of features, we improved on their results.

An improvement on this system was presented by the authors in 2009 (Morante & Daelemans, 2009a). In this case, they used four classifiers instead of one to find the full scope of the negation signals. Three classifiers predicted whether a token was the first token, the last, or neither in the scope sequence. A fourth classifier used these predictions to determine the scope classes. The test file was preprocessed using a list of 17 negation signals extracted from the training data set. Instances with these negation signals are directly assigned to their class, so the classifiers only predicted the class of the rest of the tokens.

The set of documents used for experimentation was wider because the whole BioScope corpus was used. The previous system used only abstracts.

The third difference between these two systems is that in this case a more refined set of attributes was used. For clinical documents, the *F* score of negation detection was 84.2% and 70.75% of scopes were correctly identified. For complete papers, the *F* score was 70.94% and 41% of scopes were correctly identified. In the case of abstracts, the *F* score was 82.60% and the percent of scopes correctly classified was 66.07%.

Morante and Daelemans (2009b) extended their research to include hedge detection and scope hedge detection. They showed that the same scope-finding approach can be applied to both negation and hedging. The *F* score of hedging detection for clinical documents was 38.16%; 26.21% of scopes were correctly identified. For complete papers, the *F* score was 59.66%, and 35.92% of scopes were correctly identified. The *F* score for abstracts was 78.54% and the percentage of scopes correctly classified was 65.55%. We compared the results obtained by this approach with those obtained by our approach. Our system showed better performance in general, especially in the case of speculation.

Another recent work is that developed by Agarwal and Yu (2010). In this work, the authors detected negation cue phrases and their scope in clinical notes and biological literature from the BioScope corpus using conditional random fields (CRF) as a machine-learning algorithm.

The authors selected all negation and speculation sentences from the three subcorpora and an equal number of nonnegation or speculation sentences randomly chosen. These new subcorpora are divided into two groups; the first one is used for training and the other for testing.

The best CRF-based model achieved an *F*1 score of 98% and 95% on detecting negation cue phrases and their scope in clinical notes, and an *F*1 score of 97% and 85% on detecting negation cue phrases and their scope in biological literature. In the case of speculation, the results were 88% and 86% in detecting speculation cue phrases and their scope in biological literature and 93% and 90% in clinical notes. Comparison of our system with that system was not possible because the evaluation process performed in both cases was different.

Method

System Architecture

To solve the problem of negation and speculation detection, our system was modeled as two consecutive classification tasks. These were implemented using supervised machine-learning methods trained on the annotated clinical documents from the BioScope corpus. As shown in Figure 1, in the training phase the data set was preprocessed to obtain a valid representation for the classification algorithm, both in the signal detection and the scope detection phase. In this representation and for the training phase, each instance was a token of the clinical subcollection which has a number of associated features. In the test phase, an instance was a signal–token pair from the sentence.

Finally, the classification models for each of the two tasks were generated using different classification algorithms.

In the test phase, the classification models obtained in the training phase were used to assess system performance.

When the signals were detected, a classifier decided if the tokens in a sentence were at the beginning of a negation or speculation signal, inside or outside. This enabled the system to find complex negation signals formed by more than one word. When the scope was detected, another classifier determined at sentence level the tokens affected by the signals previously identified. This means that, for every sentence that had negation or speculation signals, the classifier decided if the other words in the sentence were inside or outside the signal. The process was repeated as many times as signals appeared in the sentence. Both phases are shown in Figure 2. We also tested the scope-finding system using the gold-standard negation and speculation signals as shown in Figure 3.

This is an unbalanced class problem, so it was considered that applying sampling techniques to the data could help solve the problem and improve the system performance.

In this type of problem, the classification algorithms tend toward the majority class. Sampling techniques can solve this problem through an *oversampling* of the minority class and an *undersampling* of the majority class, using a random strategy. In our case, a supervised resample technique was used in the scope detection phase. This technique produces a random sample data set using sampling with replacement. A 0 value in the class distribution parameter leaves the class distribution as it is, whereas a value of 1 ensures the class distribution is uniform in the output data. After experimenting with different class distribution parameter

DOI: 10.1002/asi

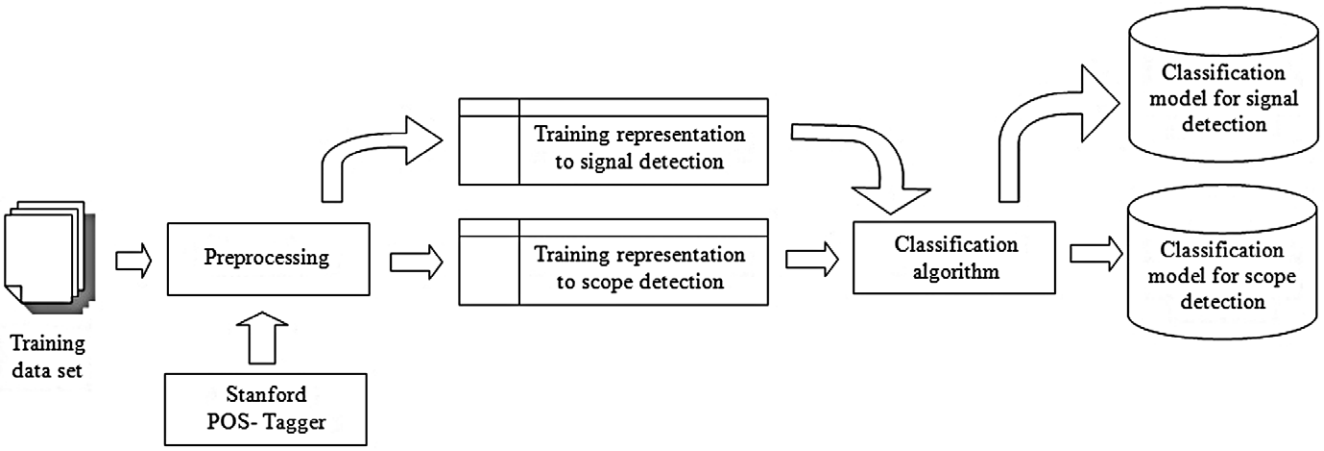

FIG. 1. Training system architecture.

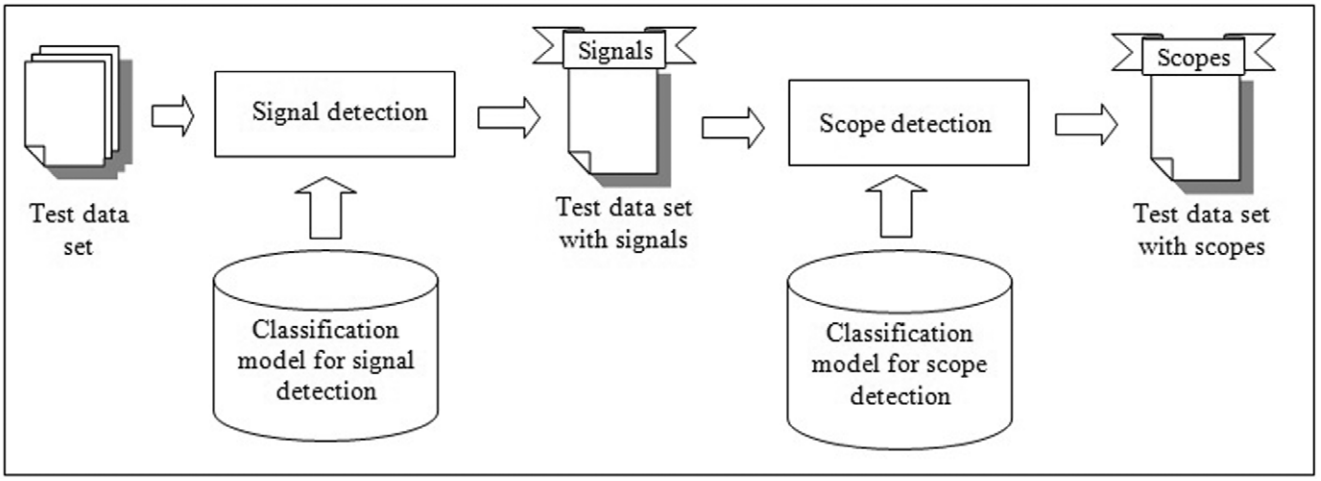

FIG. 2. Whole system testing.

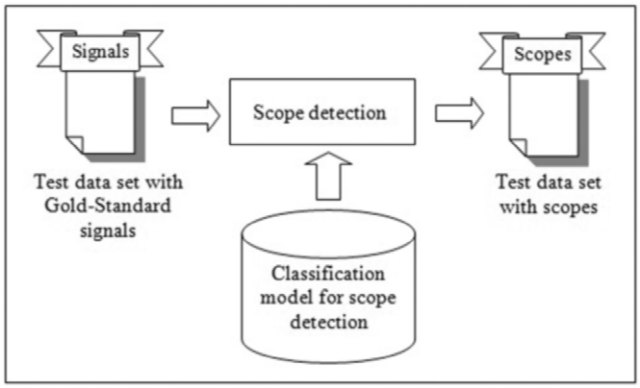

FIG. 3. Testing the scope detection system.

values, we decided to use the value 0.3 because it achieved the best results. We also experimented with resample techniques in the signal detection phase, but this was not effective.

Naïve Bayes and C4.5 algorithms implemented in Weka (version 3.6) were used. Weka (Witten & Frank, 2005) is a popular machine-learning software suite that supports several standard data-mining algorithms. C4.5 is an algorithm for learning classification tasks that builds decision trees from a set of training data in the same way as ID3 (Quinlan, 1986), using the concept of information entropy (Quinlan, 1993). Decision trees are robust; they admit discrete and numerical values, and the splitting criterion (information gain) is fairly well established and accepted as good. Moreover, this method allows us to obtain the rules that explain the different ways of negation and speculation. Garcia, Fernandez, and Herrera (2009) have showed how the approach of using sampling techniques with a C4.5 decision tree is highly competitive in terms of accuracy and is suitable for imbalanced problems.

Another classifier with which we experimented was a support vector machine (SVM) as implemented in LIBSVM by Chang and Lin (2001). We chose this classifier over others because it has proven very powerful in text classification

DOI: 10.1002/asi

TABLE 1. Statistics on the three subcollections in the BioScope corpus.

	Clinical	Papers	Abstracts
# Documents	1,954	9	1,273
# Sentences	6,383	2,670	11,871
% Negation sentences	13.55	12.69	13.45
% Speculation sentences	13.39	19.43	17.69

TABLE 2. Statistics on the subcollection of clinical documents in the BioScope corpus.

# Documents	1,954
# Sentences	6,383
# Words	42,495
# Negation signals	872
# Speculation signals	1,137
# Words in the scope of any negation signal	3,364
# Words in the scope of any speculation signal	5,336

tasks that often achieve the best performance, as described by Sebastiani (2002). We experimented with linear, polynomial, radial basis function (RBF), and sigmoid kernels; we also used LIBSVM to optimize the parameters "g" (gamma) and "c" (cost). As values to assess, we used those recommended by Hsu, Chang, and Lin (2003): $c = 2^{-5}; 2^{-3}; \ldots; 2^{15}$ and $g = 2^{-15}; {}^{-13}; \ldots; 2^{3}$.

The results obtained by each classifier are detailed in the Results section.

Text Collection

The document collection used in our study was part of the BioScope corpus, the first freely available resource with an annotation of negative/speculative signals and their scope. It consists of the following:

1. Clinical documents: This represents the major part of the corpus and was used for the clinical coding challenge (Pestian et al., 2007) organized by the Computational Medicine Center in Cincinnati, Ohio, in 2007.
2. Full articles: Five articles from FlyBase and four articles from the open access BMC Bioinformatics Web site.
3. Abstracts: 1,273 scientific abstracts obtained from the Genia corpus (Collier et al., 1999). These types of documents are the main targets for various text-mining applications, such as protein interaction mining, because of their public accessibility.

Table 1 summarizes the chief characteristics of the three subcollections. The third and fourth rows of the table show the percentage of negative and speculative sentences that appears in each subcollection. In articles and abstracts, the number of speculative sentences is greater than the number of negative sentences. This is because ambiguity is a characteristic of these types of documents.

The subcollection used in our experiments consisted of clinical documents. It contains 1,954 documents, each having a clinical history and an impression section in which the radiologist describes the conclusion or diagnosis obtained from the radiographies. Moreover, this subcollection represents the major portion of the corpus and is the densest in negative and speculative signals. Specifically, 4.78% of the words in the subcollection of clinical documents are negation or speculation signals. In the subcollection of articles, the percentage is 1.73%, whereas in the abstracts only 1.57% of the words are signals. As shown in Table 2, 6,383 sentences were used, which contained 872 negation signals and 1,137 speculation signals. The most frequent negative signals are *no* (77.0%), *without* (11.1%), and *not* (6.8%). In the case of speculative signals, the most common are *or* (22.5%), *may* (9.4%), and *evaluate for* (7.2%). Likewise, 6.15% of the words belong to the scope of any signal.

Attributes

The set of documents used in our experiments is organized into sentences. A sentence is a sequence of tokens each of which starts on a new line. For this reason, we decided to work at sentence level. To obtain the part of speech (POS) of tokens, we used the Stanford POS tagger, which is a Java implementation of the tagger based on maximum entropy originally written by Kristina Toutanova (Toutanova & Manning, 2000).

All tokens that appear in the subcollection of clinical documents used in our research were represented by a set of features that were different in each of the two phases into which the task was divided. In both phases, we started with a large pool of selected features based on experience and previous works. These features encoded information about the signal, the paired token, their contexts, and the tokens in between. The final feature set was obtained using the information gained and chi-squared feature selection techniques implemented in Weka on the initial set of attributes, starting with all the features and eliminating the least informative features.

In the task of identifying negation and speculation signals, the instances were represented by the 13 features shown in Table 3. The features are divided into token level and token context level. All of the token context-level features are used in the case of the token to the left and the right of the token in focus.

In that case, feature selection experiments showed that the most informative feature was the lemma of the token, followed by the POS of the token and the lemmas and the POS of the tokens in context. Features such as a prefix that indicates if the token starts with "in" or "un" were irrelevant and therefore were eliminated in the final set of attributes.

When the scope of the negative and speculative signals in the subcollection was detected, we experimented with the following combination features:

DOI: 10.1002/asi

TABLE 3. Features in the task of identifying negation and speculation signals.

Token level	Token context level
Lemma	Lemma
POS	POS
Binary feature that indicates if a token was at the beginning of a sentence or not	Binary feature that indicates if a token was at the beginning of a sentence or not
Binary feature that indicates if a token was at the end of a sentence or not	Binary feature that indicates if a token was at the end of a sentence or not
Binary feature that indicates if a token was at the beginning of a document or not	Binary feature that indicates if a token was at the beginning of a document or not
Binary feature that indicates if a token was at the end of a document or not	Binary feature that indicates if a token was at the end of a document or not
Tag which takes the values BN, IN, BE, IE, O as if the token was at the beginning of a negation signal, inside a negation signal, at the beginning of a speculation signal, inside a speculation signal or outside	

Note. POS = part of speech.

- Of the signal: Lemma, POS, and a tag that takes the values NEG if the signal was a negation signal or ESP if the signal was a speculation signal
- Of the paired token: Lemma, POS, and a tag that indicated if the paired token is inside or outside of the scope of the negation or speculation signal and which takes the values ISN, ISE, OS
- Of the tokens between the signal and the token in focus: The distance in number of tokens and chain of POS types
- Others: A tag that indicated the location of the token relative to the signal and that takes the values PRE, INS, POST. The lemma, POS, and type of the first token is to the left and the first token to the right. The chain of types corresponding to the tokens is between the signal and the token in focus. The signal number is divided by the total number of tokens in the sentence. The token number is divided by the total number of tokens in the sentence.

The most informative features in scope detection, according to the feature selection experiments, were the information about the signal (the lemma of the signal, followed by the tag that indicated if the signal was a negation or speculation and the POS of the signal). In that case, information about whether the signal or the paired token was at the beginning or at the end of a document or sentence was the least informative. Besides removing less significant features from the final set of features, we tested different combinations of attributes that have shown worse performance than that of the combination explained above.

Evaluation

To obtain the system performance, two different tests were carried out: token level evaluation and signal level evaluation. In both cases, in the negation and speculation detection task, a token was correctly identified if it had been classified as being at the beginning, inside, or outside the negation or speculation signal. Precision and recall and their harmonic mean *F* score (Van Rijsbergen, 1979) were used as measures.

$$\text{Precision (P)} = \frac{\#\text{tokens correctly negated by the system}}{\#\text{tokens negated by the system}}$$

$$\text{Recall (R)} = \frac{\#\text{tokens correctly negated by the system}}{\#\text{tokens negated in the test collection}}$$

$$F\ \text{score} = \frac{2\text{PR}}{\text{P}+\text{R}}$$

In the token level evaluation, within the task of identifying the scope, a token was correctly classified if it was properly identified as being inside or outside the scope of all negation or speculation signals that appear in the sentence. This meant that if there was more than one negation or speculation signal in the sentence, the token was correctly assigned a class for each of these signals. The evaluation takes the token as a unit. As measures, we used the same ones as in the negation and speculation detection task. In this case,

$$\text{Precision (P)} = \frac{\#\text{tokens belonging to some scope correctly identified by the system}}{\#\text{tokens belonging to some scope identified by the system}}$$

$$\text{Recall (R)} = \frac{\#\text{tokens belonging to some scope correctly identified by the system}}{\#\text{tokens belonging to some scope in the test collection}}$$

The *F* score was calculated using the same expression as in the negation and speculation detection task. In the scope identification task, the percentage of scopes correctly classified was also evaluated. This is a signal level evaluation and therefore takes the signal as a unit. In this case, the scope associated with a signal was correct when all the tokens of a sentence had been correctly classified as inside or outside the scope of the signal.

Finally, note that in both evaluations, negation and speculation were evaluated separately.

Results

The aim of our system was to identify negation and speculation signals and their scope in clinical documents. The results were obtained by training and evaluating the system with the subcorpus of clinical reports of the BioScope corpus. Specifically, the subcollection was randomly divided into three parts; two thirds were used to train and one third to evaluate. Our results were compared in different ways.

As we detailed in the System Architecture section, we experimented with the C4.5 and SVM classifiers. With the

DOI: 10.1002/asi

TABLE 4. Performance of negation and speculation signal detection of the C4.5 classifier and the SVM classifier, in terms of precision, recall, and *F* score (%).

	Precision	Negation recall	*F* score	Precision	Speculation recall	*F* score
C4.5 classifier	96.5	98.0	97.3	92.1 (92.8[a])	92.1 (93.4[a])	92.4 (93.1[a])
SVM classifier Linear $c = 2$, $g = 2^{-11}$	96.5	97.4	96.9	94.3 (94.5[a])	93.2 (93.6[a])	93.7 (94.1[a])
SVM classifier Polynomial $c = 2^{-3}$, $g = 2^{-1}$	97.3	93.9	95.6	95.5	80.6	87.45
SVM classifier RBF $C = 2^{5}$, $g = 2^{-5}$	96.8	97.1	96.9	95.9	93.2	94.9
SVM classifier Sigmoid $c = 2^{7}$, $g = 2^{-7}$	96.8	97.1	96.9	95.4 (95.4[a])	93.2 (93.4[a])	94.3 (94.4[a])

Note. SVM = support vector machine.
[a]Result obtained after applying the postprocessing algorithm.

latter, we carried out experiments with the main types of kernels, optimizing in each case the parameters "c" and "g." Table 4 shows the results obtained by these classifiers.

The *F* score shown by our system for some speculation cases was obtained by applying postprocessing. For reasons of space and clarity in the tables, the results obtained after applying the postprocessing algorithm are shown only in cases where this process improves the initial results. The pseudocode of that algorithm is shown in Figure 4. When the signal was formed by a single token, the algorithm changed the class of tokens classified as inside a signal for the start of a signal. When the signal was formed by more than one token and the different types of tokens in the signal did not match, i.e., some had been classified as speculation and others as negation, we consulted the ratings given to that expression so far and the class was replaced by that with the highest appearance frequency. If it had not yet been classified, the class was replaced as speculation (because the number of speculation signals is greater than the negation signals in both data sets).

Although the results obtained in the negation signal detection task are slightly higher than those obtained in speculation, all the algorithms in general achieve a great performance value. The best *F* score in negation, as shown in the third column in Table 4, was obtained by the C4.5 classifier (97.3%). However, the difference between the results obtained by the best model of SVM is not significant (97.3% vs. 96.9%). In the speculation detection task, the SVM classifier with RBF kernel obtained an *F* score of 94.9%. That is the best result although, as we mentioned above, in general all the classifiers presented a good performance value. Only the SVM classifier when the polynomial kernel is used obtained a lower *F* score value than the best.

In the negation and speculation signal detection task, four different systems were used to compare the performance of our system. Two baseline algorithms were used. The first baseline was created by tagging the two most frequent expressions of negation and speculation in the training data set as negation and speculation signals. In the second baseline, the eight most frequent expressions were used.

Likewise, in the case of negation we compared the results achieved by our system with the results obtained by NegEx for the same test data set. The comparison for speculation detection with NegEx could not be performed because this system was not designed to detect these types of signals.

We also compared our system's results with the system developed by Morante and Daelemans (2009a, 2009b). This system is very efficient both in negation detection and in speculation detection. The results shown by this system were obtained by the authors by training on the full abstract subcollection and testing on the clinical subcollection, both from the BioScope corpus.

Table 5 shows the results for these two classifiers and the four baseline systems.

In the case of negation detection, as shown in the third column, the first baseline already obtained a reasonably good performance value. This is because the two most frequent expressions of negation represent 88.1% of all expressions of negation present in the training data set. This does not happen in the case of speculation, where, as shown in the last column, the performance was lower. In this case, the two most frequent expressions of speculation represent only 31.9% of the total.

The second baseline system shows how, with a more comprehensive list of signals, it is possible to improve the performance values obtained by the first baseline. NegEx, for its part, returned the worst results in negation detection. This may be because the system is not specially designed to work with documents from the radiology domain. As shown in the third column, Morante and Daelemans's (2009a) system achieved the best *F* score value (99.0%) in negation detection. However, the difference in the *F* score value obtained by the second baseline (97.4%) and by the C4.5 (97.3%) or SVM (96.9%) classifiers is minor. In all cases, the result would be comparable to those obtained by a rater performing the same task.

Speculation detection, as shown in the *F*-score values obtained by the baseline algorithms, is more complicated because the most frequent signals are not concentrated in a small number of expressions, as in negation detection. In this case, as shown in the last column, the difference

DOI: 10.1002/asi

```
IF length(signal)=1
      IF signal.type="in" THEN
            signal.type="bn"
      ENDIF
      IF signal.type="ie" THEN
            signal.type="be"
      ENDIF
      IF signal doesn't appears in test dictionary THEN
            dictionaryTest.add(signal)
      ENDIF
      IF signal appears in test dictionary THEN
            increase frequency
      ENDIF
ENDIF
IF NOT
      IF all words inside the signal don't have the same
      type THEN
            Find signal in train dictionary classified as
            negation
            Find signal in train dictionary classified as
            speculation
            IF NOT appears THEN
                  Find signal in test dictionary classified
                  as negation
                  Find signal in test dictionary classified
                  as speculation
                  IF appears THEN
                        Compare frequencies and change the
                        type of the signal for those with
                        higher frequency
                        Increase frequency of those with
                        higher frequency in test dictionary
                  ENDIF
                  IF NOT appears THEN
                        Change the type of the signal for
                        speculation
                        dictionaryTest.add(signal)
                  ENDIFNOT
            ENDIFNOT
            IF appears THEN
                        Compare frequencies and change the
                        type of the signal for those with
                        higher frequency
                        Increase frequency of those with
                        higher frequency in train dictionary
            ENDIF
      ENDIF
ENDIFNOT
```

FIG. 4. Signal detection postprocessing algorithm pseudocode.

TABLE 5. Performance of negation and speculation signal detection of our classifier, baseline algorithms, NegEx, and the system developed by Morante and Daelemans (2009a, 2009b) in terms of precision, recall, and *F* score (%). Our system uses a C4.5 classifier for the negation signal detection and a SVM RBF classifier for the speculation signal detection.

	Precision	Negation recall	*F* score	Precision	Speculation recall	*F* score
Baseline 1	98.1	85.9	91.6	97.2	33.6	50.0
Baseline 2	96.5	98.4	97.4	94.9	70.5	80.9
NegEx	63.9	67.4	65.6	—	—	—
Morante	100	98.0	**99.0**	71.2	52.3	60.3
Our system	96.5	98.0	97.3	95.9	93.2	**94.9**

Note. SVM = support vector machine.

DOI: 10.1002/asi

TABLE 6. Performance of scope detection of a C4.5 classifier and a SVM classifier in terms of precision, recall, *F* score, and PCS with gold standard signals (%).

Gold standard signals								
	Precision	Negation recall	*F* score	PCS	Precision	Speculation recall	*F* score	PCS
C4.5 classifier	91.7	88.7	90.2	91.3	80.1 (81.8[a])	70.9 (70.0[a])	75.3 (75.5[a])	55.7 (58.6[a])
SVM classifier Linear $c = 2^{-1}$, $g = 2^{-1}$	93.3	90.7	92.0	89.4	89.6	68.1	77.4	56.4
SVM classifier Polynomial $c = 2^{3}$, $g = 2^{-3}$	94.1	91.8	92.7	87.8	87.8	75.0	80.9	68.4
SVM classifier RBF $c = 2^{15}$, $g = 2^{-5}$	93.8	92.7	93.2	89.4	89.9	72.3	80.1	67.4
SVM classifier Sigmoid $c = 2^{11}$, $g = 2^{-13}$	93.0	91.5	92.2	90.3	89.9	67.8	77.3	58.6

Note. SVM = support vector machine; PCS = percentage of correct scopes.
[a]Result obtained after applying the postprocessing algorithm.

TABLE 7. Performance of scope detection of a C4.5 classifier and a SVM classifier in terms of precision, recall, *F* score, and PCS with predicted signals (%).

Predicted signals								
	Precision	Negation Recall	F-score	PCS	Precision	Speculation recall	*F* score	PCS
C4.5 classifier	89.6	85.8	87.7	89.2	73.2	58.9	65.3 (65.4[a])	49.5 (51.9[a])
SVM classifier Linear $c = 2^{-1}$, $g = 2^{-1}$	91.8	88.3	90.0	87.2	83.2	57.7	68.1	50.9
SVM classifier Polynomial $c = 2^{3}$, $g = 2^{-3}$	92.8	88.0	90.3	86.9	84.5	54.0	65.9	62.1
SVM classifier RBF $c = 2^{15}$, $g = 2^{-5}$	92.1	89.7	90.9	87.8	84.8	62.5	71.9	62.9
SVM classifier Sigmoid $c = 2^{11}$, $g = 2^{-13}$	91.6	86.6	89.1	87.6	83.9	49.6	62.3	52.8

Note. SVM = support vector machine; PCS = percentage of correct scopes.
[a]Result obtained after applying the postprocessing algorithm.

between all systems in terms of precision, recall, and *F* score is relevant. The SVM RBF classifier provides the highest performance. The *F*-score value obtained by this algorithm is 94.9% compared to 60.3% obtained by the system developed by Morante and Daelemans (2009b). This system presented a low-value *F* score, which is even lower than that reached by the second baseline (80.9%). The difference between the C4.5 and the SVM RBF classifier is not relevant (94.9% vs. 93.1%), as shown in Table 4.

Therefore, in the case of negation detection, all systems except NegEx achieved high performance values. In terms of speculation detection, the SVM RBF classifier obtained the best results, and there was a significant difference compared with the Morante–Daelemans (2009b) system. We also experimented with C4.5 and SVM classifiers. This latter was proven with different kernel types and with optimized values of "c" and "g" parameters. We report the results obtained by these classifiers in Tables 6 and 7.

Also, as in the case of signal detection, some of the results were obtained by applying a simple postprocessing algorithm on the output of the classifier. This algorithm removed the scope consisting of a single token. If a token was classified as belonging to the scope of a signal, but the word on the left and right was outside the scope, the algorithm changed the type of the token as not belonging to the signal. Figure 5 shows the pseudocode of the postprocessing algorithm.

```
i<-2
WHILE i<length(words) DO
      IF words(i-1).type="os" AND words(i+1).type="os" AND
      words(i).type<>"os" THEN
            words(i).type<-"os"
      ENDIF
      i<-i+1
ENDWHILE
```

FIG. 5. Scope detection postprocessing algorithm pseudocode.

As shown in Table 6, both classifiers had similar performance in the case of gold standard signals. In the negation detection task, the results obtained by the systems were almost the same in terms of *F* score (93.2% and 90.2%

DOI: 10.1002/asi

TABLE 8. Performance of scope detection of our classifier and the system developed by Morante and Daelemans (2009a, 2009b) in terms of precision, recall, *F* score, and PCS with gold standard signals and predicted signals (%).

	Gold standard signals				Predicted signals			
	Negation		Speculation		Negation		Speculation	
	Morante	Our system	Morante	Our system	Morante	Our system	Morante	Our system
Precision	91.6	93.8	79.1	87.8	86.3	92.1	68.2	84.8
Recall	92.5	92.7	78.1	75.0	82.1	89.7	26.4	62.5
F score	92.0	**93.2**	78.6	**80.9**	84.2	**90.9**	38.1	**71.9**
PCS	87.2	**89.4**	60.5	**68.4**	70.7	**87.8**	26.2	**62.9**

Note. SVM = support vector machine; PCS = percentage of correct scopes. With gold standard signals, our system consists of a SVM RBF classifier in negation detection and a SVM polynomial classifier in speculation detection. With predicted signals our system consists of a SVM RBF classifier both in negation and in speculation.

obtained by the SVM RBF and C4.5 classifiers, respectively). However, the C4.5 classifier identified correctly a slightly higher number of scopes (91.3%) than did the SVM classifier (90.3% in the best of the models), as shown in the last row of the first two columns of Table 6.

As occurred when we compared the results obtained in the signal detection experiments, and due to the complexity of the speculation detection task, the results when we are dealing with speculation are worse than when we are dealing with negation. In that case, the SVM classifier outperforms the results obtained by the C4.5 classifier.

The *F*-score value obtained by the SVM polynomial classifier was 80.9% versus the 75.5% obtained by the C4.5 classifier. In terms of PCS (percentage of correct scopes), the difference between both classifiers was important because the result obtained by the C4.5 classifier was 58.6%, whereas that obtained by the SVM polynomial classifier was 68.4%. These values are shown in the third and fourth columns of Table 6. The results are competitive in negation but improvable in speculation, especially in terms of PCS.

With predicted signals as shown in Table 7, in negation detection, the differences between the classifiers were not significant both in terms of the *F* score and in terms of the PCS measure. The C4.5 classifier identified 89.2% of the full scopes correctly, while the SVM RBF classifier correctly identified 87.8%.

In speculation detection using predicted signals, the results obtained by the SVM RBF classifier were higher than those of the C4.5 classifier. In terms of the *F* score, the results obtained by the C4.5 classifier were 65.4% whereas those obtained by the SVM RBF classifier were 71.9%. The difference in PCS measure was 11%; the SVM RBF classifier obtained 62.9% against 51.9%. This difference is due to errors that the C4.5 classifier accumulated in scope detection where its *F* score was significantly lower than the *F* score obtained by the SVM classifier.

For this task, we again compared our results with the results obtained by Morante and Daelemans (2009a, 2009b) as shown in Table 8. The evaluation of the system carried out by these authors is the same as in the signal detection task, i.e., training the system on the whole abstract subcollection and testing it on the clinical subcollection. The comparison was done in two ways: using as signals those that appear directly in the documents (gold standard signals) and using the signals that the system has identified in the previous phase (predicted signals).

With gold standard signals, both systems had similar performance measures. In the case of negation detection, in terms of precision, recall, *F* score, and PCS measure, the systems are efficient. These results are shown in the first two columns of Table 4. Our system obtained a higher *F* score (93.2%) than the system developed by Morante and Daelemans (2009a; 92.0%). Also, in terms of PCS, our system correctly identified more full scopes (89.4%) than those identified by Morante and Daelemans (87.2%).

As in signal detection, the results for speculation are worse. The third and fourth columns show the results obtained by both systems. In this case, the results can be improved, especially in the PCS measure where we obtained a value of 68.4% and Morante and Daelemans (2009b) obtained a value of 60.5%.

With predicted signals, in negation detection, the differences between both systems were slightly significant in terms of PCS measure. Our system identified 84.2% of the full scopes correctly, whereas the Morante–Daelemans system (2009a) correctly identified 70.7%. The performance of our classifier was also higher than that of Morante and Daelemans in terms of the *F* score, but in this case the difference was not relevant.

In speculation detection using predicted signals, the results obtained by our system were considerably higher than those of Morante and Daelemans (2009b). Both in terms of *F* score and PCS measures, our results doubled the values of theirs. We achieved a value of 71.9% in *F* score while Morante and Daelemans obtained a value of 38.1%. The difference in PCS measure was greater, specifically 62.9% against the 26.2% obtained by Morante and Daelemans.

These results show that our system is comparable with a competitive system such as Morante and Daelemans's and in some cases may even surpass the results obtained by those authors. This is especially important when evaluating the

DOI: 10.1002/asi

whole system; our system correctly identifies around twice as many scopes associated with speculation signals as that of the Morante and Daelemans system.

Error Analysis

In analyzing the cases in which the system did not detect the signals correctly, we found that the errors could be classified into two types: (a) false-positive errors in which the system identified as signals words that are not marked as signals in the subcollection and (b) false-negative errors in which the system did not identify as signals words that are marked as signals in the subcollection.

The first category of errors, in negation signal detection, was observed in 11 cases out of the 313 negation signals presented in the test subcollection. This represents an error rate of 3.51%. Most of these errors occur because speculation signals that include the words "not" or "no" appear in the subcollection. For example, the signal "no evidence" is always marked as a speculation signal in the subcollection. Each time this signal appears, our system identifies the word "no" as a negation signal. This is because the word "no" appears 433 times in the training subcollection identified as a signal compared to five occurrences of "no evidence."

In speculation signal detection, there were 17 cases of error. The total number of speculation signals in the test subcollection is 428; therefore, the error rate is 3.97%. Errors in this case arise mainly because our system identifies as signals some words that appear in the training collection mostly classified as such. However, in the test collection, in some cases, these words are not signal words. For example, the signal "could" is marked in the subcollection 38 times as a speculation signal, but twice it is not marked. In these two cases, our system classified the words as a speculation signal.

The second type of error, in negation signal detection, occurred in 6 cases out of the 313 negation signals presented in the test subcollection. The error rate is 1.91%, slightly lower than in the case of false-positive errors. These errors occur because some signals are always marked as speculation signals except in one case, in which it is marked as a negation signal. For example, this occurred with the signals "may" or "rule out." The first is marked as a speculation signal 66 times and the second one 37 times in the train subcollection. Obviously, in the case where these signals are marked as negation signals, our system classifies them as speculation signals and therefore the system fails.

When we are dealing with speculation, the errors occur 26 times. The error rate is 6.07%, higher than in the other type of errors. In this case, errors are mostly of two types. One consists of signals that included the words "no/not/cannot" and our system classified only the words "no/not/cannot" as a negation signal (this type of error is the same as the false-positive errors in the negation detection described above). The other type of error occurs because our system does not identify as speculation signals those expressions that have infrequent occurrences. An example is "maybe," which only appears twice in the train subcollection and each time is marked as a speculation signal. Here, our system does not detect these signals as such.

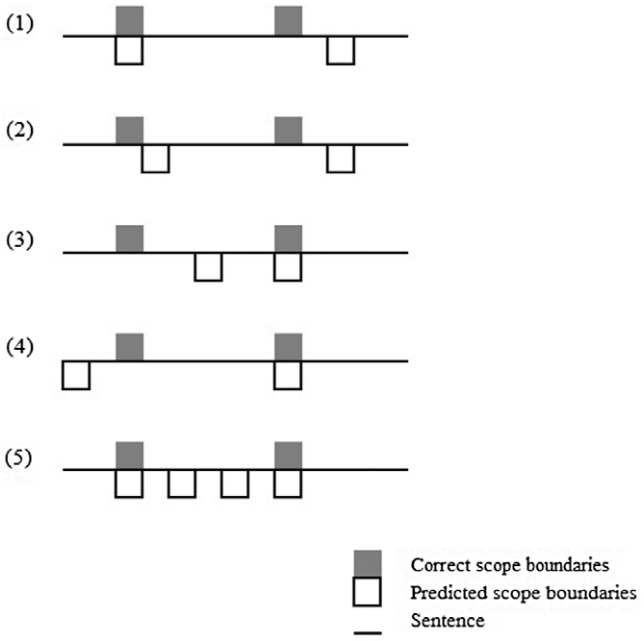

FIG. 6. Errors in scope detection task.

The most frequent errors occurred when the system identifying the scope of the signals can be divided into a wide range of categories, as shown in Figure 6.

1. The beginning of the scope is correct, but the system incorrectly extends the scope beyond the end of the sentence. For example, in the phrase "viral or reactive airways disease," the scope of the signal "or" is the words "viral" and "reactive," whereas our system identified as scope these words but also "airways" and "disease." For negation, this type of error represents 42.4% of the total; for speculation, it is 45.4%.
2. The scope identified by our system begins after the correct scope and is extended beyond the end of the sentence. For example, the scope of the signal "or" in the phrase "considerations include community acquired or atypical pneumonia such as mycoplasma" is formed by the words "community," "acquired," "atypical," but our system identifies the words "acquire," "atypical," "pneumonia," "such," "as," "mycoplasma." This type of error comprises 3.8% of the total in negation and 7.5% in speculation.
3. The end of the scope is correct but it begins after the correct position. For example, in the sentence "This may represent areas of atelectasis and/or pneumonia," our system identifies the word "pneumonia" as scope while the correct scope is formed by the words "areas," "of," "atelectasis," "pneumonia." This error represents 21.5% of the total in negation and 13.6% in speculation.
4. The scope identified by our system is correct, except that it includes words that appear before the signal. This type of error appears in 8.3% of negation signals and in

DOI: 10.1002/asi

10.6% of speculation signals. An example would be the sentence "There is no focal, lobar consolidation to suggest bacterial pneumonia." The correct scope is "bacterial pneumonia," but our system includes also the word "focal" in the scope.

5. The scope identified by our system begins and ends correctly, but is not identified as belonging to scope all the words that compose it (incorrectly omitted some words). This error represents 8.3% of the total for negation and 6.0% for speculation. For example, our system identified as scope the words "viral reactive airways disease" instead of "viral small airways reactive airways disease" (in this case, it omitted exactly two words, "small" and "airways").

Discussion

Our global system, when we use the previously identified signals, shows a higher level of performance, especially in PCS measure. Likewise, in the case of speculation, the best approach (SVM RBF classifier) identifies correctly nearly twice as many scopes as the Morante and Daelemans system (2009a, 2009b). Also, the SVM classifier has shown better performance than the C4.5 classifier in almost all cases.

Other works that we described in the Background section in the same tasks are not directly comparable with our approach because the evaluation corpora are different. For this reason, the only comparisons that we made are with the results published by Morante and Daelemans (2009a, 2009b). Moreover, the systems mentioned in the Background section are not publically available excepting NegEx. The discussion about the results of NegEx appears in the Results section.

The results obtained in the signal detection task by our system are not directly comparable with NegEx because NegEx makes a UMLS concept-level detection. Also, the database of regular expressions with which the NegEx was designed does not work as well with documents of different genre. However, it gives us an estimation of how much more efficient our system is than a system based on regular expressions as powerful as NegEx.

Comparisons with the system developed by Morante and Daelemans are possible because the collection of documents used in the test phase is the same as we used. The only difference is that they used the whole collection whereas we used one-third randomly selected documents.

Our work focused on clinical documents because this research is part of a project described in Buenaga et al. (2010). Therefore, the collection used to train and evaluate the system was the subcollection of clinical documents of the BioScope corpus. For this reason, the proposed approach may not be generalizable to other domains because the expectations in terms of effectiveness could be different if it was used in a corpus with other features, such as scientific texts.

Our results are not comparable with the results obtained for the system developed by Agarwal and Yu (2010) because the test process performed in both cases is different.

Conclusion

In this article, we have described a machine-learning system that identifies the negation signals and their scope in clinical texts. The system was trained and tested using the clinical subcollection of the BioScope corpus. The results show the superiority of the machine-learning-based approach regarding the use of regular expressions. In the detection of negation expressions, the proposed system improves the $F1$ measure of NegEx by 30%. In speculation detection, the proposed system beats the $F1$ measure of the best system by almost 10%.

Moreover, we compared our results with those obtained by the machine-learning system developed by Morante and Daelemans (2009a, 2009b). In the case of negation, our global system correctly identifies approximately 20% more than the scopes identified by the other system. In speculation, this difference is greater and our global system correctly identifies nearly twice the number of scopes.

Future research will be aimed at measuring the robustness of the system when different types of texts from the same domain are applied. To this end, the system will be tested on the article and abstract subcollections of the BioScope corpus.

Acknowledgments

This work was partially funded by the Spanish Ministry of Science and Innovation, the Spanish Government Plan E, and the European Union through ERDF (TIN2009-14057-C03-03). We would like to thank four anonymous reviewers for their valuable suggestions.

References

Agarwal, Sh., & Yu, H. (2010). Biomedical negation scope detection with conditional random fields. Journal of the American Medical Information Association, 17(6), 696–701.

Averbuch, M., Karson, T., Ben-Ami, B., Maimon, O., & Rokach, L. (2004). Context-sensitive medical information retrieval. In M. Fieschi, E. Coiera, & Y.-C.J. Li (Eds.), MEDINFO 2004: Proceedings of the 11th World Congress on Medical Informatics (pp. 1–8). Amsterdam, The Netherlands: IOS Press.

Buenaga, M., Fernández Riverola, F., Maña, M., Puertas, E., Glez-Peña, D., & Mata, J. (2010). Medical-Miner: Integración de conocimiento textual explícito en técnicas de minería de datos para la creación de herramientas traslacionales en medicina [Medical-Miner: Integrating explicit knowledge in data mining techniques for the development of translational medicine tools]. Procesamiento del Lenguaje Natural, 47, 319–320.

Chang, C.-C., & Lin, C.-J. (2001). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2(27), 1–27.

Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., & Buchanan, B.G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Information, 34(5), 301–310.

Collier, N., Park, H.S., Ogata, N., Tateishi, Y., Nobata, C., Ohta, T., . . . Tsujii, J. (1999). The GENIA project: Corpus-based knowledge acquisition and information extraction from genome research papers. In H.S. Thompson & A. Lascarides (Eds.), Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-99) (pp. 8–12). Stroudsburg, PA: Association for Computational Linguistics.

DOI: 10.1002/asi

Councill, I., McDonald, R., & Velikovich, L. (2010). What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis. In R. Morante & C. Sporleder (Eds.), Proceedings of the Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP '10). Stroudsburg, PA: Association for Computational Linguistics.

Dadvar, M., Hauff, C., & de Jong, F. (2011). Scope of negation detection in sentiment analysis. In Dutch-Belgian Information Retrieval Workshop (pp. 16–20). Amsterdam, The Netherlands: IOS Press.

Denny, J.C., Miller, R.A., Waitman, L.R., Arrieta, M.A., & Peterson, J.F. (2008). Identifying QT prolongation from ECG impressions using a general-purpose natural language processor. International Journal of Medical Informatics, 78(Suppl 1), S34–S42.

Denny, J.C., Choma, N.N., Peterson, J.F., Miller, R.A., Bastarache, L., Li, M., & Peterson, N.B. (2012). Natural language processing improves identification of colorectal cancer testing in the electronic medical record. Medical Decision Making, 32(1), 188–197.

Elkin, P.L., Brown, S.H., Bauer, B.A., Husser, C.S., Carruth, W., Bergstrom, L.R., & Wahner-Roedler, D.L. (2005). A controlled trial of automated classification of negation from clinical notes. BMC Medical Information Decision Making, 5(1), 13.

Garcia, S., Fernandez, A., & Herrera, F. (2009). Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems. Applied Soft Computing, 9, 1304–1314.

Goldin, I.M., & Chapman, W.W. (2003). Learning to detect negation with "not" in medical texts. In SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference. New York, NY: ACM Press.

Huang, Y., & Lowe, H.J. (2007). A novel hybrid approach to automated negation detection in clinical radiology reports. Journal of the American Medical Information Association, 14(3), 304–311.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). A practical guide to support vector classification (Technical report). Taiwan: National Taiwan University, Department of Computer Science.

Mierswa, I., Lemmen, F., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: Rapid prototyping for complex data mining tasks. In L. Ungar, M. Craven, D. Dunopulos, & T. Eliassi-Rad (Eds.), Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06) (pp. 935–940). New York, NY: ACM Press.

Mitchell, K.J., Becich, M.J., Berman, J.J., Chapman, W.W., Gilbertson, J., Gupta, D., . . . Crowley, R.S. (2004). Implementation an evaluation of a negation tagger in a pipeline-based system for information extract from pathology reports. Medinformation, 11(Pt 1), 663–667.

Morante, R., Liekens, A., & Daelemans, W. (2008). Learning the scope of negation in biomedical texts. In M. Lapata & T.H. NG (Eds.), Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (pp. 715–724). Stroudsburg, PA: Association of Computational Linguistics.

Morante, R., & Daelemans, W. (2009a). A metalearning approach to processing the scope of negation. In S. Stevenson & X. Carreras (Eds.), Proceedings of the 13th Conference on Computational Natural Language Learning (pp. 21–29). Stroudsburg, PA: Association of Computational Linguistics.

Morante R, & Daelemans W. (2009b). Learning the scope of hedge cues in biomedical texts. In K.B. Cohen, D. Demner-Fushman, S. Ananiadon, J. Pestian, J. Tsujii, & B. Webber (Eds.), Proceedings of the Workshop on BioNLP (pp. 28–36). Stroudsburg, PA: Association of Computational Linguistics.

Mutalik, P.G., Deshpande, A., & Nadkarni, P.M. (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 Information Association, 8(6), 598–609.

Pestian, J., Brew, C., Matykiewicz, P., Hovermale, D.J., Johnson, N., Cohen, K.B., & Duch, W. (2007). A shared task involving multi-label classification of clinical free text. In K.B. Cohen, D. Demner-Fushman, C. Friedman, L. Hirschman, & J. Pestian (Eds.), Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing (pp. 97–104). Stroudsburg, PA: Association of Computational Linguistics.

Quinlan, J.R. (1986). Induction of decision trees. Machine Learning, 1, 81–106.

Quinlan, J.R. (1993). C4.5: Programs for machine learning. Waltham, MA: Morgan Kaufmann.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Survey, 34(1), 1–47.

Toutanova, K., & Manning, C.D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In H. Schütze & K.-Y. Su (Eds.), Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (pp. 63–70). Stroudsburg, PA: Association of Computational Linguistics.

Van Rijsbergen, C.J. (1979). Information retrieval. London, England: Butterworths-Heinemann.

Vince, V., Szarvas, G., Farkas, R., Móra, G., & Csirik, J. (2008). The BioScope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. In D. Demner-Fushman, S. Ananiadon, K.B. Cohen, J. Pestian, J. Tsujii, & B. Webber (Eds.), Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP '08) (pp. 38–45). Stroudsburg, PA: Association of Computational Linguistics.

Witten, I.H., & Frank, E. Data mining: Practical machine learning tools and techniques (2nd ed.). Waltham, MA: Morgan Kaufmann.