A metalearning approach to processing the scope of negation

Roser Morante, Walter Daelemans
CNTS - Language Technology Group
University of Antwerp
Prinsstraat 13, B-2000 Antwerpen, Belgium
{Roser.Morante,Walter.Daelemans}@ua.ac.be

Abstract

Finding negation signals and their scope in text is an important subtask in information extraction. In this paper we present a machine learning system that finds the scope of negation in biomedical texts. The system combines several classifiers and works in two phases. To investigate the robustness of the approach, the system is tested on the three subcorpora of the BioScope corpus representing different text types. It achieves the best results to date for this task, with an error reduction of 32.07% compared to current state of the art results.

1 Introduction

In this paper we present a machine learning system that finds the scope of negation in biomedical texts. The system works in two phases: in the first phase, negation signals are identified (i.e., words indicating negation), and in the second phase the full scope of these negation signals is determined. Although the system was developed and tested on biomedical text, the same approach can also be used for text from other domains.

Finding the scope of a negation signal means determining at sentence level the sequence of words in the sentence that is affected by the negation. This task is different from determining whether a word is negated or not. For a sentence like the one in Example (1) taken from the BioScope corpus (Szarvas et al., 2008), the system detects that *lack, neither*, and *nor* are negation signals; that *lack* has as its scope *lack of CD5 expression*, and that the discontinuous negation signal *neither ... nor* has as its scope *neither to segregation of human autosome 11, on which the CD5 gene has been mapped, nor to deletion of the CD5 structural gene*.

(1) <sentence id="S334.5">Analysis at the phenotype and genetic level showed that <xcope id"X334.5.3"><cue type="negation" ref="X334.5.3">lack</cue> of CD5 expression</xcope> was due <xcope id="X334.5.1"> <cue type="negation" ref="X334.5.1">neither</cue> to segregation of human autosome 11, on which the CD5 gene has been mapped, <cue type="negation" ref="X334.5.1">nor</cue> to deletion of the CD5 structural gene</xcope>.</sentence>

Predicting the scope of negation is relevant for text mining and information extraction purposes. As Vincze et al. (2008) put it, extracted information that falls in the scope of negation signals cannot be presented as factual information. It should be discarded or presented separately. Szarvas et al. (2008) report that 13.45% of the sentences in the abstracts section of the BioScope corpus and 12.70% of the sentences in the full papers section contain negations. A system that does not deal with negation would treat the facts in these cases incorrectly as positives. Additionally, information about the scope of negation is useful for entailment recognition purposes.

The approach to the treatment of negation in NLP presented in this paper was introduced in Morante et al. (2008). This system achieved a 50.05 percentage of correct scopes but had a number of important shortcomings. The system presented here uses a different architecture and different classification task definitions, it can deal with multiword negation signals, and it is tested on three subcorpora of the BioScope corpus. It achieves an error reduction of

Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pages 21–29,
Boulder, Colorado, June 2009.

32.07% compared to the previous system.

The paper is organised as follows. In Section 2, we summarise related work. In Section 3, we describe the corpus on which the system has been developed. In Section 4, we introduce the task to be performed by the system, which is described in Section 5. Results are presented and discussed in Section 6. Finally, Section 7 puts forward some conclusions.

2 Related work

Negation has been a neglected area in open-domain natural language processing. Most research has been performed in the biomedical domain and has focused on detecting whether a medical term is negated or not, whereas in our approach we focus on detecting the full scope of negation signals.

Chapman et al. (2001) developed NegEx, a regular expression based algorithm for determining whether a finding or disease mentioned within narrative medical reports is present or absent. The reported results are 94.51% precision and 77.84% recall. Mutalik et al. (2001) developed Negfinder, a rule-based system that recognises negated patterns in medical documents. It consists of two tools: a lexical scanner that uses regular expressions to generate a finite state machine, and a parser. The reported results are 95.70% recall and 91.80% precision.

Sanchez-Graillet and Poesio (2007) present an analysis of negated interactions in 50 biomedical articles and a heuristics-based system that extracts such information. The preliminary results reported range from 54.32% F-score to 76.68%, depending on the method applied. Elkin et al. (2005) describe a rule-based system that assigns to concepts a level of certainty as part of the generation of a dyadic parse tree in two phases: First a preprocessor breaks each sentence into text and operators. Then, a rule based system is used to decide if a concept has been positively, negatively, or uncertainly asserted. The system achieves 97.20% recall and 98.80% precision.

The systems mentioned above are essentially based on lexical information. Huang and Lowe (2007) propose a classification scheme of negations based on syntactic categories and patterns in order to locate negated concepts, regardless of their distance from the negation signal. Their hybrid system that combines regular expression matching with grammatical parsing achieves 92.60% recall and 99.80% precision. Additionally, Boytcheva et al. (2005) incorporate the treatment of negation in a system, MEHR, that extracts from electronic health records all the information required to generate automatically patient chronicles. They report 57% of negations correctly recognised.

The above-mentioned research applies rule-based algorithms to negation finding. Machine learning techniques have been used in some cases. Averbuch et al. (2004) developed an algorithm that uses information gain to learn negative context patterns. Golding and Chapman (2003) experiment with Naive Bayes and Decision Trees to distinguish whether a medical observation is negated by the word *not* in a corpus of hospital reports. They report a maximum of 90% F-score.

Goryachev et al. (2006) compare the performance of four different methods of negation detection, two regular expression-based methods and two classification-based methods trained on 1745 discharge reports. They show that the regular expression-based methods show better agreement with humans and better accuracy than the classification methods. Like in most of the work mentioned, the task consists in determining whether a medical term is negated. Rokach et al. (2008) present a new pattern-based algorithm for identifying context in free-text medical narratives.The originality of the algorithm lies in that it automatically learns patterns similar to the manually written patterns for negation detection.

We are not aware of any research that has focused on learning the full scope of negation signals outside biomedical natural language processing.

3 Negation in the BioScope Corpus

The system has been developed using the BioScope corpus (Szarvas et al., 2008; Vincze et al., 2008)[1], a freely available resource that consists of medical and biological texts. In the corpus, every sentence is annotated with information about negation and speculation. The annotation indicates the boundaries of the scope and the keywords, as shown in (1) above. In the annotation, scopes are extended to the

[1]Web page: `www.inf.u-szeged.hu/rgai/bioscope`.

biggest syntactic unit possible, so that scopes have the maximal length, and the negation signal is always included in the scope. The annotation guidelines and the inter-annotator agreement information can be found on the web page.

	Clinical	Papers	Abstracts
#Documents	1954	9	1273
#Sentences	6383	2670	11871
#Words	41985	60935	282243
#Lemmas	2320	5566	14506
Av. length sentences	7.73	26.24	26.43
% Sent. 1-10 tokens	75.85	11.27	3.17
% Sent. 11-20 tokens	20.99	27.67	30.49
% Sent. 21-30 tokens	2.94	29.55	35.93
% Sent. 31-40 tokens	0.15	17.00	19.76
% Sent. $>$ 40 tokens	0.01	0.03	10.63
%Negation sentences	13.55	12.70	13.45
#Negation signals	877	389	1848
Av. length scopes	4.98	8.81	9.43
Av. length scopes to the right	4.84	7.61	8.06
Av. length scopes to the left	6.33	5.69	8.55
% Scopes to the right	97.64	81.77	85.70
% Scopes to the left	2.35	18.22	14.29

Table 1: Statistics about the subcorpora in the BioScope corpus and the negation scopes ("Av". stands for *average*).

The BioScope corpus consists of three parts: clinical free-texts (radiology reports), biological full papers and biological paper abstracts from the GENIA corpus (Collier et al., 1999). Table 1 shows statistics about the corpora. Negation signals are represented by one or more tokens.

Only one negation signal (*exclude*) that occurs in the papers subcorpus does not occur in the abstracts subcorpus, and six negation signals (*absence of, exclude, favor, favor over, may, rule out* that appear in the clinical subcorpus do not appear in the abstracts subcorpus. The negation signal *no* (determiner) accounts for 11.74 % of the negation signals in the abstracts subcorpus, 12.88 % in the papers subcorpus, and 76.65 % in the clinical subcorpus. The negation signal *not* (adverb) accounts for 58.89 % of the negation signals in the abstracts subcorpus, 53.22 % in the papers subcorpus, and 6.72 % in the clinical subcorpus.

The texts have been processed with the GENIA tagger (Tsuruoka and Tsujii, 2005; Tsuruoka et al., 2005), a bidirectional inference based tagger that analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags in a tab-separated format. Additionally, we converted the annotation about scope of negation into a token-per-token representation, following the standard format of the 2006 CoNLL Shared Task (Buchholz and Marsi, 2006), where sentences are separated by a blank line and fields are separated by a single tab character. A sentence consists of a sequence of tokens, each one starting on a new line.

4 Finding the scope of negation

We model the scope finding task as two consecutive classification tasks: a first one that consists of classifying the tokens of a sentence as being at the beginning of a negation signal, inside or outside. This allows the system to find multiword negation signals.

The second classification task consists of classifying the tokens of a sentence as being the first element of the scope, the last, or neither. This happens as many times as there are negation signals in the sentence. We have chosen this classification model after experimenting with two additional models that produced worse results: in one case we classifed tokens as being inside or outside of the scope. In another case we classified chunks, instead of tokens, as being inside or outside of the scope.

5 System description

The two classification tasks (identifying negation signals and finding the scope) are implemented using supervised machine learning methods trained on part of the annotated corpus.

5.1 Identifying negation signals

In this phase, a classifier predicts whether a token is the first token of a negation signal, inside a negation signal, or outside of it. We use IGTREE as implemented in TiMBL (version 6.1.2) (Daelemans et al., 2007). TiMBL[2] is a software package that contains implementations of memory-based learning algorithms like IB1 and IGTREE. We also experimented with IB1, but it produced lower results.

[2]TiMBL can be downloaded from the web page `http://ilk.uvt.nl/timbl/`.

The classifier was parameterised by using gain ratio for feature weighting. The instances represent all tokens in the corpus and they have features of the token (lemma) and of the token context: word form, POS, and chunk IOB tag[3] of one token to the left and to the right; word form of the second token to the left and to the right. According to the gain ratio scores, the most informative feature is the lemma of the token, followed by the chunk IOB tag of the token to the right, and the features relative to the token to the left.

The test file is preprocessed using a list of negation signals extracted from the training corpus, that are unambiguous in the training corpus. The list comprises the following negation signals: *absence, absent, fail, failure, impossible, lack, loss, miss, negative, neither, never, no, none, nor, not, unable, without*. Instances with this negation signals are directly assigned their class. The classifier predicts the class of the rest of tokens.

5.2 Scope finding

In this phase three classifiers predict whether a token is the first token in the scope sequence, the last, or neither. A fourth classifier is a metalearner that uses the predictions of the three classifiers to predict the scope classes. The three object classifiers that provide input to the metalearner were trained using the following machine learning methods:

- Memory-based learning as implemented in TiMBL (version 6.1.2) (Daelemans et al., 2007), a supervised inductive algorithm for learning classification tasks based on the k-nearest neighbor classification rule (Cover and Hart, 1967). In this lazy learning approach, all training data is kept in memory and classification of a new item is achieved by extrapolation from the most similar remembered training items.
- Support vector machines (SVM) as implemented in SVM^{light}V6.01 (Joachims, 1999). SVMs are defined on a vector space and try to find a decision surface that best separates the data points into two classes. This is achieved by using quadratic programming techniques. Kernel functions can be used to map the original vectors to a higher-dimensional space that is linearly separable.
- Conditional random fields (CRFs) as implemented in CRF++-0.51 (Lafferty et al., 2001). CRFs define a conditional probability distribution over label sequences given a particular observation sequence rather than a joint distribution over label and observation sequences, and are reported to avoid the label bias problem of HMMs and other learning approaches.

The memory-based learning algorithm was parameterised by using overlap as the similarity metric, gain ratio for feature weighting, using 7 k-nearest neighbors, and weighting the class vote of neighbors as a function of their inverse linear distance. The SVM was parameterised in the learning phase for classification, cost factor of 1 and biased hyperplane, and it used a linear kernel function. The CRFs classifier used regularization algorithm L2 for training, the hyper-parameter and the cut-off threshold of features were set to 1.

An instance represents a pair of a negation signal and a token from the sentence. This means that all tokens in a sentence are paired with all negation signals that occur in the sentence. Negation signals are those that have been classified as such in the previous phase. Only sentences that have negation signals are selected for this phase.

We started with a larger, extensive pool of 131 features which encoded information about the negation signal, the paired token, their contexts, and the tokens in between. Feature selection experiments were carried out with the memory-based learning classifier. Features were selected based on their gain ratio, starting with all the features and eliminating the least informative features. We also performed experiments applying the feature selection process reported in Tjong Kim Sang et al. (2005), a bi-directional hill climbing process. However, experiments with this method did not produce a better selection of features.

The features of the first three classifers are:

- Of the negation signal: Chain of words.
- Of the paired token: Lemma, POS, chunk IOB tag, type of chunk; lemma of the second and third tokens to the left; lemma, POS, chunk IOB tag, and type of chunk of the first token to the left and three tokens to the right; first word, last word, chain of words, and chain of POSs of the chunk of the paired token and of two chunks to the left and two chunks to the

[3] Tags produced by the GENIA tagger that indicate if a token is inside a certain chunk, outside, or at the beginning.

right.

- Of the tokens between the negation signal and the token in focus: Chain of POS types, distance in number of tokens, and chain of chunk IOB tags.
- Others: A feature indicating the location of the token relative to the negation signal (pre, post, same).

The fourth classifier, a metalearner, is also a CRF as implemented in CRF++. The features of this classifier are:

- Of the negation signal: Chain of words, chain of POS, word of the two tokens to the right and two tokens to the left, token number divided by the total number of tokens in the sentence.
- Of the paired token: Lemma, POS, word of two tokens to the right and two tokens to the left, token number divided by the total number of tokens in the sentence.
- Of the tokens between the negation signal and the token in focus: Binary features indicating if there are commas, colons, semicolons, verbal phrases or one of the following words between the negation signal and the token in focus:

  Whereas, but, although, nevertheless, notwithstanding, however, consequently, hence, therefore, thus, instead, otherwise, alternatively, furthermore, moreover.
- About the predictions of the three classifiers: prediction, previous and next predictions of each of the classifiers, full sequence of previous and full sequence of next predictions of each of the classifiers.
- Others: A feature indicating the location of the token relative to the negation signal (pre, post, same).

Negation signals in the BioScope corpus always have one consecutive block of scope tokens, including the signal token itself. However, the classifiers only predict the first and last element of the scope. We need to process the output of the classifers in order to build the complete sequence of tokens that constitute the scope. We apply the following post-processing:

(2) - If one token has been predicted as FIRST and one as LAST, the sequence is formed by the tokens between first and last.

- If one token has been predicted as FIRST and none has been predicted as LAST, the sequence is formed by the token predicted as FIRST.

- If one token has been predicted as LAST and none as FIRST, the sequence will start at the negation signal and it will finish at the token predicted as LAST.

- If one token has been predicted as FIRST and more than one as LAST, the sequence will end with the first token predicted as LAST after the token predicted as FIRST, if there is one.

- If one token has been predicted as LAST and more than one as FIRST, the sequence will start at the negation signal.

- If no token has been predicted as FIRST and more than one as LAST, the sequence will start at the negation signal and will end at the first token predicted as LAST after the negation signal.

6 Results

The results provided for the abstracts part of the corpus have been obtained by performing 10-fold cross validation experiments, whereas the results provided for papers and clinical reports have been obtained by training on the full abstracts subcorpus and testing on the papers and clinical reports subcorpus. The latter experiment is therefore a test of the robustness of the system when applied to different text types within the same domain.

The evaluation is made using the precision and recall measures (Van Rijsbergen, 1979), and their harmonic mean, F-score. In the negation finding task, a negation token is correctly classified if it has been classified as being at the beginning or inside the negation signal. We also evaluate the percentage of negation signals that have been correctly identified. In the scope finding task, a token is correctly classified if it has been correctly classified as being inside or outside of the scope of all the negation signals that there are in the sentence. This means that when there is more than one negation signal in the sentence, the token has to be correctly assigned a class for as many negation signals as there are. Additionally, we evaluate the percentage of correct scopes (PCS). A scope is correct if all the tokens in the sentence have been assigned the correct scope class for a specific negation signal. The evaluation in terms of precision and recall measures takes as unit a token, whereas the evaluation in terms of PCS takes as unit a scope.

6.1 Negation signal finding

An informed baseline system has been created by tagging as negation signals the tokens with the words: *absence, absent, fail, failure, impossible, instead of, lack, loss, miss, negative, neither, never, no, none, nor, not, rather than, unable, with the exception of, without*. The list has been extracted from the training corpus. Baseline results and inter-annotator agreement scores are shown in Table 2.

Corpus	Prec.	Recall	F1	Correct	IAA
Abstracts	100.00	95.17	97.52	95.09	91.46
Papers	100.00	92.46	96.08	92.15	79.42
Clinical	100.00	97.53	98.75	97.72	90.70

Table 2: Baseline results of the negation finding system and inter-annotator agreement (IAA) in %.

Table 3 shows the results of the system, which are significantly higher than the results of the baseline system. With a more comprehensive list of negation signals it would be possible to identify all of them in a text.

Corpus	Prec.	Recall	F1	Correct
Abstracts	100.00	98.75	99.37	98.68
Papers	100.00	95.72	97.81	95.80
Clinical	100.00	98.09	99.03	98.29

Table 3: Results of the negation finding system in %.

The lower result of the papers subcorpus is caused by the high frequency of the negation signal *not* in this corpus (53.22 %), that is correct in 93.68 % of the cases. The same negation signal is also frequent in the abstracts subcorpus (58.89 %), but in this case it is correct in 98.25 % of the cases. In the clinical subcorpus *not* has low frequency (6.72 %), which means that the performance of the classifer for this negation signal (91.22 % correct) does not affect so much the global results of the classifier. Most errors in the classification of *not* are caused by the system predicting it as a negation signal in cases not marked as such in the corpus. The following sentences are some examples:

(3) However, programs for tRNA identification [...] do not necessarily perform well on unknown ones.
The evaluation of this ratio is difficult because not all true interactions are known.
However, the Disorder module does not contribute significantly to the prediction.

6.2 Scope finding

An informed baseline system has been created by calculating the average length of the scope to the right of the negation signal in each corpus and tagging that number of tokens as scope tokens. We take the scope to the right for the baseline because it is much more frequent than the scope to the left, as is shown by the statistics contained in Table 1 of Section 3.

Corpus	Prec.	Recall	F1	PCS	PCS-2	IAA
Abstracts	76.68	78.26	77.46	7.11	37.45	92.46
Papers	69.34	66.92	68.11	4.76	24.86	70.86
Clinical	86.85	74.96	80.47	12.95	62.27	76.29

Table 4: Baseline results of the scope finding system and inter-annotator agreement (IAA) in %.

Baseline results and inter-annotator agreement scores are presented in Table 4. The percentage of correct scopes has been measured in two ways: PCS measures the proportion of correctly classified tokens in the scope sequence, whereas PCS-2 measures the proportion of nouns and verbs that are correctly classifed in the scope sequence. This less strict way of computing correctness is motivated by the fact that being able to determine the concepts and relations that are negated (indicated by content words) is the most important use of the negation scope finder. The low PCS for the three subcorpora indicates that finding the scope of negations is not a trivial task. The higher PCS for the clinical subcorpus follows a trend that applies also to the results of the system. The fact that, despite a very low PCS, precision, recall and F1 are relatively high indicates that these measures are in themselves not reliable to evaluate the performance of the system.

The upper-bound results of the metalearner system assuming gold standard identification of negation signals are shown in Table 5.

Corpus	Prec.	Recall	F1	PCS	PCS-2
Abstracts	90.68	90.68	90.67	73,36	74.10
Papers	84.47	84.95	84.71	50.26	54.23
Clinical	91.65	92.50	92.07	87.27	87.95

Table 5: Results of the scope finding system with gold-standard negation signals.

The results of the metalearner system are presented in Table 6. Results with gold-standard nega-

tion signals are especially better for the clinical subcorpus because except for *lack, negative* and *not*, all negation signals score a PCS higher than 90 %. Thus, in the clinical subcorpus, if the negation signals are identified, their scope will be correctly found. This does not apply to the abstracts and papers subcorpus.

Corpus	Prec.	Recall	F1	PCS	PCS-2
Abstracts	81.76	83.45	82.60	66.07	66.93
Papers	72.21	69.72	70.94	41.00	44.44
Clinical	86.38	82.14	84.20	70.75	71.21

Table 6: Results of the scope finding system with predicted negation signals.

In terms of PCS, results are considerably higher than baseline results, whereas in terms of precision, recall and F1, results are slightly higher. Compared to state of the art results (50.05 % PCS in (anonymous reference) for the abstracts subcorpus), the system achieves an error reduction of 32.07 %, which shows that the system architecture presented in this paper leads to more accurate results.

Evaluating the system in terms of a more relaxed measure (PCS-2) does not reflect a significant increase in its performance. This suggests that when a scope is incorrectly predicted, main content tokens are also incorrectly left out of the scope or added. An alternative to the PCS-2 measure would be to mark in the corpus the relevant negated content words and evaluate if they are under the scope.

Results also show that the system is portable to different types of documents, although performance varies depending on the characteristics of the corpus. Clinical reports are easier to process than papers and abstracts, which can be explained by several factors. One factor is the length of sentences: 75.85 % of the sentences in the clinical reports have 10 or less words, whereas this rate is 3.17 % for abstracts and 11.27 % for papers. The average length of a sentence for clinical reports is 7.73 tokens, whereas for abstracts it is 26.43 and for papers 26.24. Shorter sentences imply shorter scopes. In the scope finding phase, when we process the output of the classifiers to build the complete sequence of tokens that constitute the scope, we give preference to short scopes by choosing as LAST the token classified as LAST that is the closest to the negation signal. A way to make the system better portable to texts with longer sentences would be to optimise the choice of the last token in the scope.

	Abstracts		Papers		Clinical	
	#	PCS	#	PCS	#	PCS
absence	57	56.14	-	-	-	-
absent	13	15.38	-	-	-	-
can not	28	42.85	16	50.00	-	-
could not	14	57.14	-	-	-	-
fail	57	63.15	13	38.46	-	-
lack	85	57.64	20	45.00	-	-
negative	-	-	-	-	17	0.00
neither	33	51.51	-	-	-	-
no	**207**	**73.42**	**44**	**50.00**	**673**	**73.10**
nor	43	44.18	-	-	-	-
none	7	57.14	10	0.00	-	-
not	**1036**	**69.40**	**200**	**39.50**	**57**	**50.87**
rather than	20	65.00	12	41.66	-	-
unable	30	40.00	-	-	-	-
without	82	89.02	24	58.33	-	-

Table 7: PCS per negation signal for negation signals that occur more than 10 times in one of the subcorpus.

Another factor that causes a higher performance on the clinical subcorpus is the frequency of the negation signal *no* (76.65 %), which has also a high PCS in abstracts, as shown in Table 7. Typical example sentences with this negation signal are shown in (4). Its main characteristics are that the scope is very short (5 tokens average in clinical reports) and that it scopes to the right over a noun phrase.

(4) No findings to account for symptoms.
No signs of tuberculosis.

The lower performance of the system on the papers subcorpus compared to the abstracts subcorpus is due to the high proportion of the negation signal *not* (53.22 %), which scores a low PCS (39.50), as shown in Table 7. Table 7 also shows that, except for *can not*, all negation signals score a lower PCS on the papers subcorpus. This difference can not be caused by the sentence length, since the average sentence length in the abstracts subcorpus (26.43 tokens) is similar to the average sentence length in the papers subcorpus (26.24). The difference may be related to the difference in the length of the scopes and their direction. For example, the average length of the scope of *not* is 8.85 in the abstracts subcorpus and 6.45 in the papers subcorpus. The scopes to the

left for *not* amount to 23.28 % in the papers subcorpus and to 16.41 % in the abstracts subcorpus, and the average scope to the left is 5.6 tokens in the papers subcorpus and 8.82 in the abstracts subcorpus.

As for the results per negation signal on the abstracts corpus, the negation signals that score higher PCS have a low (*none*) or null (*absence, fail, lack, neither, no, rather than, without*) percentage of scopes to the left. An exception is *not* with a high score and 16.41% of scopes to the left. The negation signals with lower PCS have a higher percentage of scopes to the left (*absent, can not, nor, unable*). A typical error for the negation signal *unable* is exemplified by the sentence *VDR DNA-binding mutants were unable to either bind to this element in vitro or repress in vivo*, in which the gold scope starts at the beginning of the sentence, where the predicted scopes starts at the negation signal.

6.2.1 Results of the metalearner versus results of the first three classifiers

The choice of a metalearner approach has been motivated by the significantly higher results that the metalearner produces compared to the results of the first three classifiers. The results of each of the classifiers independently are presented in Table 8.

Algor.	Ev.	Abstracts	Papers	Clinical
TiMBL	Prec.	78.85	68.66	82.25
	Rec.	80.54	66.29	78.56
	F1	79.69	67.46	80.36
	PCS	56.80	33.59	**70.87**
	PCS-2	57.99	37.30	71.21
CRF	Prec.	78.49	68.94	93.42
	Rec.	80.16	66.57	80.24
	F1	79.31	67.73	86.33
	PCS	59.90	36.50	59.51
	PCS-2	60.04	38.88	59.74
SVM	Prec.	77.74	68.01	93.80
	Rec.	79.35	65.66	85.16
	F1	78.54	66.82	89.27
	PCS	56.80	33.33	**82.45**
	PCS-2	57.59	35.18	82.68

Table 8: Results for the first three classifiers of the scope finding system.

PCS results show that the metalearner system performs significantly better than the three classifiers for the abstracts and papers subcorpora, but not for the clinical subcorpus, in which case TiMBL and SVM produce higher scores, although only the SVM results are significantly better with a difference of 11.7 PCS. An analysis in detail of the SVM scores per negation signal shows that the main difference between the scores of the metalearner and SVM is that the SVM is good at predicting the scopes of the negation signal *no* when it occurs as the first token in the sentence, like in (4) above. When *no* occurs in other positions, SVM scores 1.17 PCS better.

We plan to perform experiments with the three classifiers using the features of the metalearner that are not related to the predictions, in order to check if the three classifiers would perform better.

7 Conclusions

In this paper we have presented a metalearning approach to processing the scope of negation signals. Its performance is evaluated in terms of percentage of correct scopes on three test sets. With 66.07 % PCS on the abstracts corpus the system achieves 32.07 % of error reduction over current state of the art results. The architecture of the system is new for this problem, with three classifiers and a metalearner that takes as input the output of the first classifiers. The classification task definition is also original.

We have shown that the system is portable to different corpora, although performance fluctuates depending on the characteristics of the corpora. The results per corpus are determined to a certain extent by the scores of the negation signals *no* and *not*, that are very frequent and difficult to process in some text types. Shorter scopes are easier to learn as reflected in the results of the clinical corpus, where *no* is the most frequent negation signal. We have also shown that the metalearner performs better than the three first classifiers, except for the negation signal *no* in clinical reports, for which the SVM classifier produces the highest scores.

Future research will deal with a more detailed analysis of the errors by each of the three initial classifiers compared to the errors of the metalearner in order to better understand why the results of the metalearner are higher. We also would like to perform feature analysis, and test the system on general domain corpora.

Acknowledgments

Our work was made possible through financial support from the University of Antwerp (GOA project BIOGRAPH). We are grateful to four anonymous reviewers for their valuable comments and suggestions.

References

M. Averbuch, T. Karson, B. Ben-Ami, O. Maimon, and L. Rokach. 2004. Context-sensitive medical information retrieval. In *Proc. of the 11th World Congress on Medical Informatics (MEDINFO-2004)*, pages 1–8, San Francisco, CA. IOS Press.

S. Boytcheva, A. Strupchanska, E. Paskaleva, and D. Tcharaktchiev. 2005. Some aspects of negation processing in electronic health records. In *Proc. of International Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries*, pages 1–8, Borovets, Bulgaria.

S. Buchholz and E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proc. of the X CoNLL Shared Task*, New York. SIGNLL.

W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B.G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*, 34:301–310.

N. Collier, H.S. Park, N. Ogata, Y. Tateisi, C. Nobata, T. Sekimizu, H. Imai, and J. Tsujii. 1999. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. In *Proceedings of EACL-99*.

T. M. Cover and P. E. Hart. 1967. Nearest neighbor pattern classification. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, 13:21–27.

W. Daelemans, J. Zavrel, K. Van der Sloot, and A. Van den Bosch. 2007. TiMBL: Tilburg memory based learner, version 6.1, reference guide. Technical Report Series 07-07, ILK, Tilburg, The Netherlands.

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

l. M. Goldin and W.W. Chapman. 2003. Learning to detect negation with 'Not' in medical texts. In *Proceedings of ACM-SIGIR 2003*.

S. Goryachev, M. Sordo, Q.T. Zeng, and L. Ngo. 2006. Implementation and evaluation of four different methods of negation detection. Technical report, DSG.

Y. Huang and H.J. Lowe. 2007. A novel hybrid approach to automated negation detection in clinical radiology reports. *J Am Med Inform Assoc*, 14(3):304–311.

T. Joachims, 1999. *Advances in Kernel Methods - Support Vector Learning*, chapter Making large-Scale SVM Learning Practical, pages 169–184. MIT-Press, Cambridge, MA.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML 2001*, pages 282–289.

R. Morante, A. Liekens, and W. Daelemans. 2008. A combined memory-based semantic role labeler of english. In *Proc. of the EMNLP 2008*, pages 715–724, Honolulu, Hawaii.

A.G. Mutalik, A. Deshpande, and P.M. Nadkarni. 2001. Use of general-purpose negation detection to augment concept indexing of medical documents. a quantitative study using the UMLS. *J Am Med Inform Assoc*, 8(6):598–609.

L. Rokach, R.Romano, and O. Maimon. 2008. Negation recognition in medical narrative reports. *Information Retrieval Online.*

O. Sanchez-Graillet and M. Poesio. 2007. Negation of protein-protein interactions: analysis and extraction. *Bioinformatics*, 23(13):424–432.

G. Szarvas, V. Vincze, R. Farkas, and J. Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scopein biomedical texts. In *Proc. of BioNLP 2008*, pages 38–45, Columbus, Ohio, USA. ACL.

E. Tjong Kim Sang, S. Canisius, A.van den Bosch, and T. Bogers. 2005. Applying spelling error correction techniques for improving semantic role labelling. In *Proc. of CoNLL 2005*, pages 229–232.

Y. Tsuruoka and J. Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proc. of HLT/EMNLP 2005*, pages 467–474.

Y. Tsuruoka, Y. Tateishi, J. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii, 2005. *Advances in Informatics - 10th Panhellenic Conference on Informatics*, volume 3746 of *Lecture Notes in Computer Science*, chapter Part-of-Speech Tagger for Biomedical Text, Advances in Informatics, pages 382–392. Springer, Berlin/Heidelberg.

C.J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworths, London.

V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9((Suppl 11)):S9.