Learning the scope of hedge cues in biomedical texts

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

Identifying hedged information in biomedical literature is an important subtask in information extraction because it would be misleading to extract speculative information as factual information. In this paper we present a machine learning system that finds the scope of hedge cues in biomedical texts. The system is based on a similar system that finds the scope of negation cues. We show that the same scope finding approach can be applied to both negation and hedging. To investigate the robustness of the approach, the system is tested on the three subcorpora of the BioScope corpus that represent different text types.

1 Introduction

Research on information extraction of biomedical texts has grown in the recent years. Most work concentrates on finding relations between biological entities, like genes and proteins (Krauthammer et al., 2002; Mitsumori et al., 2006; Krallinger et al., 2008a; Krallinger et al., 2008b). Determining which information has been hedged in biomedical literature is an important subtask of information extraction because extracted information that falls in the scope of hedge cues cannot be presented as factual information. It should be discarded or presented separately with lower confidence. The amount of hedged information present in texts cannot be underestimated. Vincze et al. (2008) report that 17.70% of the sentences in the abstracts section of the BioScope corpus and 19.44% of the sentences in the full papers section contain hedge cues. Light et al. (2004) estimate that 11% of sentences in MEDLINE abstracts contain speculative fragments. Szarvas (2008) reports that 32.41% of gene names mentioned in the hedge classification dataset described in Medlock and Briscoe (2007) appears in a speculative sentence.

In this paper we present a machine learning system that finds the scope of hedge cues in biomedical texts. Finding the scope of a hedge cue means determining at sentence level which words in the sentence are affected by the hedge cue. The system combines several classifiers and works in two phases: in the first phase hedge cues (i.e., words indicating speculative language) are identified, and in the second phase the full scope of these hedge cues is found. This means that for a sentence like the one in Example (1) taken from the BioScope corpus (Szarvas et al., 2008), the system performs two actions: first, it detects that *suggest, might*, and *or* are hedge signals; second, it detects that *suggest* has as its scope *expression of c-jun, jun B and jun D genes might be involved in terminal granulocyte differentiation or in regulating granulocyte functionality*, that *might* has as its scope *be involved in terminal granulocyte differentiation or in regulating granulocyte functionality*, and that *or* has as its scope *in regulating granulocyte functionality*.

(1) These results <xcope id="X7.5.3" ><cue type= "speculation" ref="X7.5.3"> **suggest** </cue> that <xcope id= "X7.5.2">expression of c-jun, jun B and jun D genes <cue type= "speculation" ref= "X7.5.2"> **might** </cue> be involved <xcope id="X7.5.1">in terminal granulocyte differentiation <cue type= "speculation" ref="X7.5.1" >**or**</cue> in regulating granulocyte functionality </xcope></xcope></xcope>.

Proceedings of the Workshop on BioNLP, pages 28–36,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

Contrary to current practice to only detect modality, our system also determines the part of the sentence that is hedged. We are not aware of other systems that perform this task. The system is based on a similar system that finds the scope of negation cues (Morante and Daelemans, 2009). We show that the system performs well for this task and that the same scope finding approach can be applied to both negation and hedging. To investigate the robustness of the approach, the system is tested on three subcorpora of the BioScope corpus that represent different text types. Although the system was developed and tested on biomedical text, the same approach can also be applied to text from other domains.

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

Hedging has been broadly treated from a theoretical perspective. The term *hedging* is originally due to Lakoff (1972), who introduces it in relation to prototype theory. Palmer (1986) defines a term related to hedging, *epistemic modality*, which expresses the speaker's degree of commitment to the truth of a proposition. Saurí et al. (2006) research the modality of events, which "expresses the speaker's degree of of commitment to the events being referred to in a text". They treat a wide spectrum of modal types and present the codification of modality information with the specification language TimeML, which allows to mark modality cues at a lexical level and at a syntactic level.

As for research that focuses specifically on scientific texts with descriptive purposes, Hyland (1998) describes hedging in scientific research articles, proposing a pragmatic classification of hedge expressions based on an exhaustive analysis of a corpus. The catalogue of hedging cues includes modal auxiliaries, epistemic lexical verbs, epistemic adjectives, adverbs, and nouns. Additionally, it includes also a variety of non–lexical cues. Light et al. (2004) analyse the use of speculative language in MEDLINE abstacts. They studied the expression of levels of belief (hypothesis, tentative conclusions, hedges, and speculations) and annotated a corpus of abstracts in order to check if the distinction between high speculative, low speculative and definite sentences could be made reliably. They found that the speculative vs. definite distinction was reliable, but the distinction between low and high speculative was not. Thompson et al. (2008) report on a list of words and phrases that express modality in biomedical texts and put forward a categorisation scheme. The list and the scheme are validated by annotating 202 MEDLINE abstracts.

Some NLP applications incorporate modality information. Friedman et al. (1994) develop a medical text processor "that translates clinical information in patient documents into controlled vocabulary terms". The system uses a semantic grammar that consists of rules that specify well-formed semantic patterns. The extracted findings are assigned one of five types of modality information: *no, low certainty, moderate certainty, high certainty* and *cannot evaluate*. Di Marco and Mercer (2005) use hedging information to classify citations. They observe that citations appear to occur in sentences marked with hedging cues.

Work on hedging in the machine learning field has as a goal to classify sentences into speculative or definite (non speculative). Medlock and Briscoe (2007) provide a definition of what they consider to be hedge instances and define hedge classification as a weakly supervised machine learning task. The method they use to derive a learning model from a seed corpus is based on iteratively predicting labels for unlabeled training samples. They report experiments with SVMs on a dataset that they make publicly available[1]. The experiments achieve a recall/precision break even point (BEP) of 0.76. They apply a bag-of-words (BOG) approach to sample representation. Medlock (2008) presents an extension of this work by experimenting with more features (part-of-speech (PoS), lemmas, and bigrams). Experiments show that the PoS representation does not yield significant improvement over the results in

[1] Available at http://www.benmedlock.co.uk/hedgeclassif.html.

Medlock and Briscoe (2007), whereas with a lemma representation the system achieves a peak performance of 0.8 BEP, and with bigrams of 0.82 BEP. Szarvas (2008) follows Medlock and Briscoe (2007) in classifying sentences as being speculative or non-speculative. Szarvas develops a MaxEnt system that incorporates bigrams and trigrams in the feature representation and performs a complex feature selection procedure in order to reduce the number of keyword candidates. It achieves up to 0.85 BEP and 85.08 F1 by using an external dictionary. Kilicoglu and Bergler (2008) apply a linguistically motivated approach to the same clasification task by using knowledge from existing lexical resources and incorporating syntactic patterns. Additionally, hedge cues are weighted by automatically assigning an information gain measure and by assigning weights semi–automatically depending on their types and centrality to hedging. The system achieves results of 0.85 BEP.

As mentioned earlier, we are not aware of research that has focused on learning the scope of hedge signals inside or outside of the biomedical domain, which makes a direct comparison with the approaches described here impossible.

3 Hedge cues in the BioScope Corpus

The system has been developed using the BioScope corpus (Szarvas et al., 2008; Vincze et al., 2008)[2], 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 biggest syntactic unit possible, so that scopes have the maximal length, and the speculation cue is always included in the scope.

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. Hedge cues are represented by one or more tokens, as (2) shows, where the hedge cues that appear in the three corpora are listed. The complete list of all hedge cues comprises 176 cues. In the same corpora the number of negation cues is lower, 38.

(2) apparent, apparently, appear, assume, can, consider, consistent with, could, either, indicate, likely, may, no evidence, not, or, perhaps, possible, possibly, presumably, probable, probably, should, suggestion, support, think, unclear, whether, would

35 hedge cues that occur in the clinical reports subcorpus do not occur in the abstracts subcorpus, and 34 hedge cues that appear in the papers subcorpus do not appear in the abstracts subcorpus. Only 15.90% of the total of hedge cues appear in the three subcorpora. The most frequent hedge cues in the abstracts subcorpus are *may* (19.15 %), *appear* (5.30 %), and *or* (4.45 %); in the papers subcorpus, *suggest* (10.26 %), *may* (9.97 %), and *might* (5.86 %); and in the clinical subcorpus, *or* (24.27 %), *suggest* (5.62 %), and *evaluate for* (5.27 %).

	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
%Hedge sentences	13.39	19.44	17.70
# Hedge cues	1189	714	2769
Av. length scopes	5.92	14.37	16.27
Av. length scopes to the right	5.15	13.00	15.44
Av. length scopes to the left	2.46	5.94	5.60
% Scopes to the right	73.28	76.55	82.45
% Scopes to the left	26.71	23.44	17.54

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

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.

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

4 Finding the scope of hedge cues

We model this task in the same way that we modelled the task for finding the scope of negation (Morante and Daelemans, 2009), i.e., as two consecutive classification tasks: a first one that consists of classifying the tokens of a sentence as being at the beginning of a hedge signal, inside or outside. This allows the system to find multiword hedge cues. 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 hedge cues in the sentence.

5 System description

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

5.1 Identifying hedge cues

In this phase, a classifier predicts for all tokens in a sentence whether a token is the first token of a hedge cue (B-cue), inside a hedge cue (I-cue), or outside of it (O-cue). For sentence (3) the system assigns the B-cue class to *indicate*, the I-cue class to *that* and the O-cue class to the rest of tokens.

(3) These results *indicate that* a component or components of NF–AT have the potential to reconstitute NF(P)

The instances represent all tokens in the corpus and they have features about the token: lemma, word, part-of-speech (POS) and IOB[3] chunk tag; and features about the token context: Word, POS and IOB chunk tag of 3 tokens to the right and 3 to the left.

We use IGTREE as implemented in TiMBL (version 6.1.2) (Daelemans et al., 2007). We also experimented with IB1, but it produced lower results. The classifier was parameterised by using gain ratio for feature weighting. According to the gain ratio scores, the most informative features are the lemma and word of the token in focus, followed by the word of the token to the right and of the token to the left.

We performed two experiments. In one, the test file is preprocessed using a list of hedge cues extracted from the training corpus. The list comprises the following hedge cues listed in (4). Instances with these hedge cues are directly assigned their class. The classifier predicts the class of the rest of tokens. In the other experiment we don't preprocess the test file.

(4) appear, apparent, apparently, believe, either, estimate, hypothesis, hypothesize, if, imply, likely, may, might, or, perhaps, possible, possibly, postulate, potential, potentially, presumably, probably, propose, putative, should, seem, speculate, suggest, support, suppose, suspect, think, uncertain, unclear, unkwown, unlikely, whether, would

5.2 Scope finding

In this phase three classifiers predict for all tokens in the sentence whether a token is the first token in the scope sequence (F-scope), the last (L-scope), or neither (NONE). For the sentence in 3, the classifiers assign the class F-scope to *indicate*, L-scope to *NF(P)*, and NONE to the rest of tokens. A fourth classifier is a metalearner that uses the predictions of the three classifiers to predict the scope classes. An instance represents a pair of a hedge cue and a token from the sentence. This means that all tokens in a sentence are paired with all hedge cues that occur in the sentence. Hedge cues are those that have been classified as such in the previous phase. Only sentences that have hedge cues are selected for this phase. 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 (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.

[3] *I* stands for 'inside', *B* for 'beginning', and *O* for 'outside'.

- Conditional random fileds (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 in this case 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.

We have used the same features used for the system that finds the scope of negation. The features of the first three classifers are:

- Of the hedge 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 right.
- Of the tokens between the hedge cue 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 hedge cue (pre, post, same).

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

- Of the hedge 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 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 of tokens in the sentence.
- Of the tokens between the hedge cue and the token in focus: Binary features indicating if there are commas, colons, semicolons, verbal phrases or one of the following words between the hedge cue 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 hedge cue (pre, post, same).

Hedge cues in the BioScope corpus always scope over a consecutive block of tokens, including the cue 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 postprocessing:

(5) - 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 hedge cue 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 hedge signal.

- If no token has been predicted as FIRST and more than one as LAST, the sequence will start at the hedge cue and will end at the first token predicted as LAST after the hedge 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. We report micro F1.

In the hedge finding task, a hedge token is correctly classified if it has been classified as being at the beginning or inside the hedge signal. We also evaluate the percentage of hedge cues 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 hedge cues that there are in the sentence. This means that when there is more than one hedge cue in the sentence, the token has to be correctly assigned a class for as many hedge 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 hedge 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 Hedge cue finding

An informed baseline system has been created by tagging as hedge cues the tokens with the words listed in (4) above. The list has been extracted from the training corpus. The results are shown in Table 2.

Corpus	Prec.	Recall	F1	% Correct
Abstracts	55.62	71.77	62.67	70.91
Papers	54.39	61.21	57.60	64.46
Clinical	66.55	40.78	50.57	51.38

Table 2: Baseline results of the hedge finding system.

The fact that the results are lower for the papers and clinical subcorpora can be explained by the fact that the list of cues has been extracted from the training corpus.

Table 3 shows the results of the system. The results of the system for abstracts and papers are higher than baseline, but for clinical they are lower. This is due to the fact that in the baseline system the hedge cue *or* that accounts for 24.53 % of the hedge cues is 100 % correct, whereas the system achieves only 0.72 % of correct predictions. The score obtained by *or* is also the reason why the system produces lower results for the clinical subcorpus.

Corpus	Prec.	Recall	F1	% Correct
Abstracts	90.81	79.84	84.77	78.67
Papers	75.35	68.18	71.59	69.86
Clinical	88.10	27.51	41.92	33.36

Table 3: Results of the hedge finding system without preprocessing.

Table 4 shows the results of the system with preprocessing. In terms of % of correct cues, the system that uses a preprocessed test set gets higher scores, but in terms of F1 it gets lower results, except for the clinical subcorpus. The drop in F1 of this system is caused by a drop in precision due to the excess of false positives.

Corpus	Prec.	Recall	F1	% Correct
Abstracts	60.74	94.83	74.05	96.03
Papers	56.56	84.03	67.61	88.60
Clinical	71.25	52.33	60.34	64.49

Table 4: Results of the hedge finding system with preprocessing.

In the abstracts subcorpus the hedge cue that has the biggest proportion of false positives is *or*. Of the 1062 accurrences of *or*, in 88.32% of the cases *or* is not a hedge cue. The system that uses preprocessing produces 938 false positives and 4 false negatives, whereas the other system produces 21 false positives and 108 false negatives. In the papers subcorpus, the hedge cues *if, or, can, indicate* and *estimate* cause 67.38% of the false positives. In the clinical subcorpus the hedge cues *evidence, evidence of, no* and *appear* cause 88.27% of the false positives. In contrast with the abstracts subcorpus, the hedge cue *or* has only 5 false positives and scores an F1 of 99.10. So, in the clinical corpus *or* is not ambiguous, whereas in the abstracts subcorpus it is very ambiguous. An example of *or* as hedge cue in the clinical subcorpus is shown in (6). An example of *or* as hedge cue in the abstracts subcorpus is shown in (7), and as a non cue in (8).

(6) Findings compatible with reactive airway disease or viral lower respiratory tract infection.

(7) Nucleotide sequence and PCR analyses demonstrated the presence of novel duplications or deletions involving the NF-kappa B motif.

(8) In nuclear extracts from monocytes or macrophages, induction of NF-KB occurred only if the cells were previously infected with HIV-1.

Compared to negation cues, hedge cues are more varied and more ambiguous. Both the system without and with preprocessing for negation finding performed better than the hedge finding system.

6.2 Scope finding

An informed baseline system has been created by calculating the average length of the scope to the right of the hedge cue 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. Baseline results are presented in Table 5. The low PCS for the three subcorpora indicates that finding the scope of hedge cues is not a trivial task. 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.

Corpus	Prec.	Recall	F1	PCS
Abstracts	78.92	62.19	69.56	3.15
Papers	72.03	50.43	59.33	2.19
Clinical	64.92	25.10	36.20	2.72

Table 5: Baseline results of the scope finding system.

The upper-bound results of the metalearner system assuming gold standard identification of hedge cues are shown in Table 6.

Corpus	Prec.	Recall	F1	PCS	PCS-2
Abstracts	89.71	89.09	89.40	77.13	78.21
Papers	77.78	77.10	77.44	47.94	58.21
Clinical	79.16	78.13	78.64	60.59	63.94

Table 6: Results of the scope finding system with gold-standard hedge signals.

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 speculated (indicated by content words) is the most important use of the hedge scope finder.

Results show that the system achieves a high percentage of fully correct scopes, and that, although performance is lower for the papers and clinical corpora, the system is portable. Table 7 shows the results of the negation scope finding system also with gold standard negation cues. The comparison of results shows that for abstracts and papers the scores are higher for the hedge system, which means that the system can be used for finding both types of scope.

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

The results of the hedge system with predicted hedge cues are presented in Table 8. The hedge cues have been predicted by the system without the preprocessing step presented in Subsection 6.1.

Corpus	Prec.	Recall	F1	PCS	PCS-2
Abstracts	85.77	72.44	78.54	65.55	66.10
Papers	67.97	53.16	59.66	35.92	42.37
Clinical	68.21	26.49	38.16	26.21	27.44

Table 8: Results of the scope finding system with predicted hedge signals.

In terms of PCS, which is a scope based measure, results are considerably higher than baseline results, whereas in terms of precision, recall and F1, which are token based measures, results are lower. 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.

Results also show that the system based on predicted hedge cues performs lower for all corpora, which is also a trend observed for the negation scope finding system. The difference in performance for abstracts and papers follows the same trends as in the negation system, whereas the drop in performance for the clinical subcorpus is bigger. This can be explained by the results obtained in the cues finding phase, where the clinical subcorpus obtained only 41.92% F1. However, gold standard results show that if the hedge cues are identified, then the system is portable.

	Abstracts		Papers		Clinical	
	#	PCS	#	PCS	#	PCS
appear	143	58.04	39	28.20	-	-
can	48	12.5	25	0.00	22	0.00
consistent with	-	-	-	-	67	0.00
could	67	11.94	28	14.28	36	22.22
either	28	0.00	-	-	-	-
evaluate for	-	-	-	-	86	3.84
imply	21	90.47	-	-	-	-
indicate	23	73.91	-	-	-	-
indicate that	276	89.49	-	-	-	-
likely	59	59.32	36	30.55	63	66.66
may	516	81.39	68	54.41	107	80.37
might	72	73.61	40	35.00	-	-
or	120	0.00	-	-	276	0.00
possible	50	66.00	24	54.16	26	80.76
possibly	25	52.00	-	-	-	-
potential	45	28.88	-	-	-	-
potentially	21	52.38	-	-	-	-
propose	38	63.15	-	-	-	-
putatitve	39	17.94	-	-	-	-
rule out	-	-	-	-	61	0.00
suggest	613	92.33	70	62.85	64	90.62
think	35	31.42	-	-	-	-
unknown	26	15.38	-	-	-	-
whether	96	72.91	-	-	-	-
would	-	-	21	28.57	-	-

Table 9: PCS per hedge cue for hedge cues that occur more than 20 times in one of the subcorpus.

Table 9 shows the PCS results per hedge cue. The cues that get better scores in the clinical and papers subcorpora are cues that appear in the abstracts subcorpus and get a good score. Cues that occur in the clinical subcorpus and do not occur in the abstracts (training) subcorpus, get 0.00 score or close to 0.00, whereas cues that appear in both subcorpora tend to get a similar or better score in the clinical subcorpus. This is a trend that we also observed in the negation scope finding system. As with that system, we also observed that the papers subcorpus tends to get lower scores than the abstracts subcorpus.

The results of the system based on gold standard hedge cues showed that the system can be applied to negation scope finding and hedge scope finding, but these results show that the results of the second phase of the system depend on the results of the first phase of the system, and that finding hedge cues is a domain dependent task. The cues that are not present in the training data cannot be learned in the test data and the same applies to their scope. This observation is consistent with the observation that the portability of hedge classifiers is limited, made by Szarvas (Szarvas, 2008).

7 Conclusions

In this paper we have presented a metalearning approach to processing the scope of hedge cues, based on a system that finds the scope of negation cues. We have shown that the same system can find both the scope of negation and hedge cues. The performance of the system is evaluated in terms of percentage of correct scopes on three text types.

In the hedge finding phase, the system achieves an F1 of 84.77% in the abstracts subcorpus. Existing systems that classify sentences as speculative or not reach an 85.00 BEP. Although the tasks are different, we consider that the results of our system are competitive. In the scope finding phase, the system that uses predicted hedge cues achieves 65.55% PCS in the abstracts corpus, which is very similar to the result obtained by the negation scope finding system with predicted negation cues (66.07% PCS). However, the results for the papers and clinical subcorpora are considerably lower than the results for the abstracts subcorpus in the two phases. In the case of the negation scope finding system, the evaluation on the clinical subcorpus yielded a 4.23% PCS higher result, whereas in the case of the hedge scope finding system the results are almost 30.00% PCS lower, confirming the observation that the portability of hedge classifers is limited. Future research will focus on trying to improve the first phase of the system and anlysing errors in depth in order to get insights into how to get a better performance.

Acknowledgments

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

References

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.

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 *Proc. of EACL 1999*.

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.

C. Di Marco and R.E. Mercer, 2005. *Computing attitude and affect in text: Theory and applications*, chapter Hedging in scientific articles as a means of classifying citations. Springer-Verlag, Dordrecht.

C. Friedman, P. Alderson, J. Austin, J.J. Cimino, and S.B. Johnson. 1994. A general natural–language text processor for clinical radiology. *JAMIA*, 1(2):161–174.

K. Hyland. 1998. *Hedging in scientific research articles*. John Benjamins B.V, Amsterdam.

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

H. Kilicoglu and S. Bergler. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9(Suppl 11):S10.

M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia. 2008a. Overview of the protein–protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9(Suppl 2):S4.

M. Krallinger, A. Valencia, and L. Hirschman. 2008b. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biology*, 9(Suppl 2):S8.

M. Krauthammer, P. Kra, I. Iossifov, S.M. Gomez, G. Hripcsak, V. Hatzivassiloglou, C. Friedman, and A.Rzhetsky. 2002. Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics*, 18(Suppl 1):S249–57.

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.

G. Lakoff. 1972. Hedges: a study in meaning criteria and the logic of fuzzy concepts. *Chicago Linguistics Society Papers*, 8:183–228.

M. Light, X.Y.Qiu, and P. Srinivasan. 2004. The language of bioscience: facts, speculations, and statements in between. In *Proc. of the BioLINK 2004*, pages 17–24.

B. Medlock and T. Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proc. of ACL 2007*, pages 992–999.

B. Medlock. 2008. Exploring hedge identification in biomedical literature. *JBI*, 41:636–654.

T. Mitsumori, M. Murata, Y. Fukuda, K Doi, and H. Doi. 2006. Extracting protein-protein interaction information from biomedical text with svm. *IEICE - Trans. Inf. Syst.*, E89-D(8):2464–2466.

R. Morante and W. Daelemans. 2009. A metalearning approach to processing the scope of negation. In *Proc. of CoNLL 2009*, Boulder, Colorado.

F.R. Palmer. 1986. *Mood and modality*. CUP, Cambridge, UK.

R. Saurí, M. Verhagen, and J. Pustejovsky. 2006. Annotating and recognizing event modality in text. In *Proc. of FLAIRS 2006*, pages 333–339.

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

G. Szarvas. 2008. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proc. of ACL 2008*, pages 281–289, Columbus, Ohio, USA. ACL.

P. Thompson, G. Venturi, J. McNaught, S. Montemagni, and S. Ananiadou. 2008. Categorising modality in biomedical texts. In *Proc. of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining 2008*, pages 27–34, Marrakech. LREC.

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 *LNCS*, 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.