Resolving Speculation and Negation Scope in Biomedical Articles with a Syntactic Constituent Ranker

Jonathon Read **Erik Velldal** **Stephan Oepen** **Lilja Øvrelid**
Department of Informatics, University of Oslo
{jread,erikve,oe,liljao}@ifi.uio.no

Abstract

We discuss how the scope of speculation and negation can be resolved by learning a ranking function that operates over syntactic constituent subtrees. An important assumption of this method is that scope aligns with constituents, and hence we investigate instances of disalignment. We also show how the method can be combined with an existing scope-resolution system based on manually-crafted rules over dependency structures. While both systems achieve encouraging results, combining the two improves performance beyond either in isolation. Furthermore, coupling this hybrid scope approach with an SVM cue classifier achieves the best published results on data from the CoNLL-2010 Shared Task.

1 Introduction

Detecting instances of speculation and negation is an important task when processing scientific and technical text. Being motivated by the need to recognize certainty regarding biomedical events, research in this area is particularly prevalent in the biomedical domain (see Kim et al. (2009) and Farkas et al. (2010), for example).

Approaches to scope resolution for speculation and negation typically involve two stages. Firstly it is necessary to detect cues, e.g. *may* or *instead of* in the examples below:

(1) {The unknown amino acid ⟨may⟩ be used by these species}

(2) Samples of the protein pair space were taken {⟨instead of⟩ considering the whole space} as this was more computationally tractable.

The second stage involves resolving the scope of the cues; in (1) the entire sentence is speculation, whereas a phrase is negated in (2).

This paper presents a novel data-driven approach to scope resolution, which we propose to solve by learning a discriminative ranking function to score candidate scopes on the basis of syntactic (HPSG-based) constituent trees. The remainder of the paper is structured as follows. Other approaches to the scope resolution task are discussed in Section 2. Section 3 describes the data sets and the parser configuration we use to obtain constituent trees. Section 4 details our approach, while Section 5 considers an important assumption of our approach—that the annotated scope of a cue corresponds to a syntactic constituent. The experiments evaluating our scope resolution method are described in Section 6—including experiments with a *hybrid approach*, incorporating the scope predictions of the rule-based system of Velldal et al. (2010). Finally, Section 7 presents our conclusions and directions for future work.

2 Related Work

Although there exists a body of earlier work on identifying uncertainty on the *sentence level*, the task of resolving the *in-sentence scope* of speculation cues was first pioneered by Morante and Daelemans (2009a), facilitated by the BioScope corpus (Vincze et al., 2008). In this system, the tasks of cue and scope identification are both treated as sequence labeling tasks using a BIO scheme (i.e. labeling tokens as being at the Beginning, Inside, or Outside). While the initial classifier of Morante and Daelemans (2009a) uses only token-level lexical information, this is extended with syntactic features in the memory-based learner of Morante et al. (2010), achieving the best performance on the scope-level evaluation of the CoNLL-2010 Shared Task (Farkas et

al., 2010), which was devoted to the task of speculation scope resolution. In fact, all the top performing scope-resolution systems in the shared task rely on syntactic information, e.g. the output from a dependency parser (Morante et al., 2010; Velldal et al., 2010), a tag sequence grammar (Rei and Briscoe, 2010), or a constituent analysis in combination with dependency triplets (Kilicoglu and Bergler, 2010). Most of the systems rely on learning a token-level classifier using a BIO labeling scheme (Farkas et al., 2010), although some of the top performers employ manually constructed rules (Velldal et al., 2010; Kilicoglu and Bergler, 2010) or combine rules with machine learning (Rei and Briscoe, 2010).

For negation, the systems typically cited as the current top performers for scope resolution on the BioScope data are those of Morante and Daelemans (2009b) and Councill et al. (2010). The system of Morante and Daelemans (2009b) combines the predictions of three learners—a memory-based model, a Support Vector Machine (SVM) classifier and a Conditional Random Fields (CRF) classifier—using lexical features, such as PoS and chunk tags. Councill et al. (2010) employ a CRF learner with features from a dependency parser (such as the path to the negation cue).

3 Data Sets and Parsing

We use the BioScope corpus (Vincze et al., 2008) consisting of biomedical abstracts (11,871 sentences) and full papers (2,670 sentences). We will refer to these subsets as BSA and BSP respectively (and BS collectively). For evaluation purposes, the organizers of the CoNLL-2010 Shared Task (Farkas et al., 2010) provided an additional set of full papers annotated with speculation (BSE, 5,003 sentences), which we reserve for held-out testing of our speculation models for strict comparison with the Shared Task participants. As BSE does not include negation annotations we have instead set aside every tenth sentence of BSA and BSP for held-out testing for the negation task (BS_{eval}), using the remainder for development (BS_{dev}).

For parsing we employ analyses licensed by the LinGO English Resource Grammar (ERG) (Flickinger, 2002), a general-purpose, wide-coverage grammar couched in the framework of Head-Driven Phrase Structure Grammar (HPSG). To parse biomedical text using the ERG, a lattice of tokens annotated with parts of speech and named entity hypotheses contributed by the GENIA tagger (Tsuruoka et al., 2005) is input to the PET HPSG chart parser (Callmeier, 2002). As the ERG has not previously been adapted to the biomedical domain, unknown word handling in the parser plays an important role. Here we build on a set of somewhat under-specified 'generic' lexical entries for common open-class categories provided by the ERG (thus complementing the 35,000-entry lexicon that comes with the grammar), which are activated on the basis of PoS and NE annotation. Other than these, there are no robustness measures in the parser, such that syntactic analysis will fail in certain cases, viz. when the ERG is unable to derive a complete, well-formed syntactic structure for the entire input string. In this configuration, the parser returns at least one derivation for 91.2% of all utterances in BSA, and 85.6% and 81.4% for BSP and BSE, respectively.

Usually, the grammar will license several thousand parses for a sentence. Much like for other unification-based grammars and in a sense all statistical parsers, ERG parses are therefore ranked by an underlying parse selection model. The PET parser first constructs a packed forest of candidate analyses and then applies a discriminative parse ranking model to selectively enumerate an n-best list of top-ranked candidates (Zhang et al., 2007). To make the parse selection more tuned to the biomedical domain, we re-trained the discriminative model, combining gold-standard out-of-domain data from existing ERG treebanks with a fully automated procedure seeking to take advantage of syntactic annotations in the GENIA Treebank (see MacKinlay et al. (2011) for details).

4 A Constituent Ranker

In this section we first define the learning problem underlying our ranking approach, including how we identify candidate scopes. We then go on to present the various feature functions used by the learner, also showing how the ranker can take the predictions of other systems into account.

4.1 Defining the Learning Problem

Our novel approach for the resolution of speculation and negation is abstractly related to statistical *parse selection*, and in particular discriminative parse selection for unification based grammars, as discussed above. The overall goal is to learn a function for ranking syntactic structures, based on training data that annotates which tree(s) are correct and incorrect for each sentence. In our case however, rather than discriminating between complete analyses for a given sentence, we want to learn a ranking function over candidate *subtrees* (i.e. constituents) within a parse (or possibly even within several parses). Figure 1 exemplifies this idea—the scope of the cue *may* includes the subject NP, due to the interaction of a subject control verb and the passive construction. Starting from the cue and working through the tree bottom-up, there are three candidate constituents to determine scope (marked in bold), each projecting onto a substring of the full utterance, and each including at least the cue. For this particular context and cue, determining the correct scope requires ranking the top constituent (the one labeled as the subject – head construction of a main clause) as the most likely candidate scope.

The training data is defined as follows. Given a parsed BioScope sentence, the subtree that corresponds to the annotated scope for a given speculation cue will be labeled as correct. Any other remaining constituents that also span the cue are labeled as incorrect. Note that, in the case of *multiword* cues (e.g. *either. . . or*), the candidate constituents must span all the cue words. We then attempt to learn a linear SVM-based scoring function that reflects these preferences, using the implementation of ordinal ranking in the SVM$^{\text{light}}$ toolkit—see Joachims (2002) for details. Note that, for some sentences the highest-ranked parse may not actually contain a subtree that aligns with the annotated scope. We therefore also experiment with training on data from the n-best parses.

4.2 Features

As shown in Figure 1, the HPSG derivation trees are labeled with identifiers of grammar components (syntactic rules, lexical rules, and categories of lexical entries); when taken together in the exact configuration governed by the derivation,

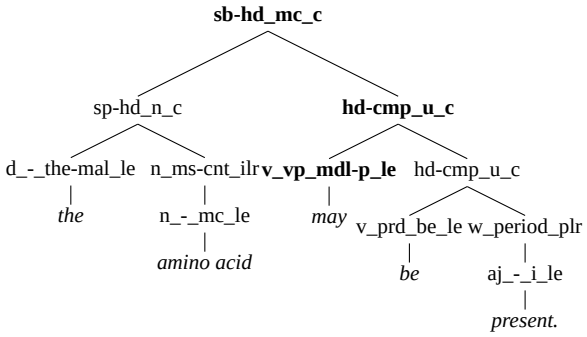

Figure 1: Derivation tree for our running example, with candidate consituents highlighted in bold. Internal nodes are labeled with ERG rule identifiers; common HPSG constructions near the top, and lexical rules closer to the leafs. The preterminals are so-called LE types, corresponding to fine-grained parts of speech with close to a thousand lexical distinctions.

these components fully and unambiguously determine the corresponding HPSG sign. For our purposes, however, we take advantage of abstractions provided in the hierarchy of HPSG constructions and lexical categories, and the systematic naming conventions for these components in the ERG.

In the following we detail three classes of features inherent to the ranker; recording properties of (*i*) paths in the tree, (*ii*) surface order and (*iii*) rule-like linguistic features. We then describe a fourth class of features that lets us incorporate the predictions of other scope models.

Given our working hypothesis that speculation scopes are aligned with syntactic constituents, the most natural features to employ are the location of constituents within trees. We define these in terms of full paths of nodes from cues to candidate constituents, including both lexicalized and unlexicalized versions. We also include a generalized version recording only the cue and the candidate. As traversal from the cue to the candidate can involve many nodes (which are not captured by the generalized path features) we also record bigrams of nodes and their parents.

The second type of features we utilize involves surface properties of scope candidates. These include: the enumeration of bigrams of the preterminal lexical types, the cue position within the candidate (in tertile bins relative to the candidate length), and the candidate size (in quartile bins relative to the sentence length). We also record

whether punctuation was present at the end of the terminal preceding the candidate or at the end of the right-most terminal of the candidate.

The third family of features extend on the previously described path-features by adding information about specific linguistic phenomena described in the BioScope annotation guidelines (Vincze et al., 2008). These include detection of: passivization; subject control verbs occurring with passivized verbs; subject raising verbs; and predicative adjectives. Furthermore, these features are only activated when the subject of the construction is not an expletive pronoun.

In addition to the features derived directly from the constituent trees, we can also make the ranker incorporate the predictions of other *external* scope resolution systems. We incorporate this information by appending a flag to the path features described above, indicating whether a candidate matches the start, end, or entirety of the span predicted by the auxiliary system. Section 6.4 reports on experiments with augmenting the ranker in this way, incorporating the scope predictions of the system of Øvrelid et al. (2010) which is based on a small set of manually crafted rules operating over dependency structures.

5 Constituent and Scope Alignment

The constituent ranking approach proposed in the previous section makes explicit an assumption that scope boundaries align with the boundaries of syntactically meaningful units. This assumption is motivated by general BioScope annotation principles, as Vincze et al. (2008) suggest: "*The scope of a keyword can be determined on the basis of syntax. [...] When marking the scopes of negative and speculative keywords, we extended the scope to the largest syntactic unit possible [...].*" To determine the degree to which ERG analyses conform with this expectation we computed the ratio of alignment between scopes and constituents in BioScope. To improve alignment we apply a small set of heuristics that slacken the requirements for alignment, removing: (a) utterance-final punctuation; (b) citation markers; (c) the left-most terminal if it is an adverb and not the cue; (d) all terminals to the left of the cue when it is a noun; and, in the case of negation only, (e) removing the left-most terminal when it is an auxiliary. Together, these rules improve alignment over parsed sentences from 74.10% to 80.54% for speculation (in BSP), and from 50.91% to 77.45% for negation (in BSP_{dev}).

The degree of alignment observed between scopes and constitutents produced by the parser increases when searching the n-best derivations. Alignment of speculation scope with constituents in the first-best parse is 84.37% for BSA and 80.54% for BSP. Including the top fifty-best derivations improves alignment to 92.21% and 88.93%, respectively. Taken together with an observed parser coverage of 85.6% for BSP, these results mean that for only about 76% of all utterances in the BioScope full papers can the ranker be expected to identify a constituent matching the gold-standard speculation scope. A similar analysis indicates that the ranker can potentially identify 77% of instances of negation.

6 Experiments

In Section 6.1 we first describe the evaluation metric we employ. Section 6.2 then reports experiments with tuning the configuration of the ranker. This involves assessing the contribution of the various feature families, as well as optimizing the number of parses to include from the n-best when extracting candidate scopes in training and testing. The performance of the constituent ranker is contrasted with an altogether different approach, namely the system of Øvrelid et al. (2010) which is based on manually defined rules operating on dependency structures. In Section 6.4 we then present a *hybrid approach*, combining the predictions of the system of Øvrelid et al. (2010) with the constituent ranker.

6.1 Evaluation Measures

We report precision, recall and F_1 computed using the official scoring software from the CoNLL-2010 Shared Task (modified to also support negation). Under this rather strict metric a *true positive* requires that both the predicted cue and scope exactly match the gold annotation. A *false positive* occurs when a predicted cue and/or its scope does not match the gold annotation. Importantly, however, when a cue is correctly identified but its scope is not (or vice-versa), both a false positive and a *false negative* is incurred (Farkas et al.,

Features	Speculation	Negation
Baseline	26.76	21.73
Path	78.10	79.28
Path+Surface	79.93	78.88
Path+Linguistic	83.72	89.47
All	85.30	89.24

Table 1: The performance of the ranker (F_1) when using various combinations of feature families, compared with a random choice baseline.

2010). Of course, a false negative will also be counted when the system fails to detect an annotated cue/scope pair at all.

In an end-to-end evaluation, the cue detection performance carries through to the scope-level performance. In order to better assess scope resolution itself we will first be reporting results using gold-standard cues. This entails that the number of false negatives and false positives will always be equal, and thus so will precision, recall and F_1. Hence, we only report F_1 when evaluating scope resolution in isolation.

6.2 Tuning the Ranker

We conducted several experiments designed to find an optimal configuration of features. Table 1 lists the results of combinations of the feature families on the BSP data set when using gold cues, reporting 10-fold cross-validated F_1 scores with respect to only the instances of speculation and negation that are aligned with constituents in parsed sentences. The table also lists a baseline performance, which is calculated as the mean ambiguity of each instance (i.e. the averaged reciprocal of the number of candidates). The feature optimization results indicate that, when ranking candidates for speculation scope resolution, each feature family is informative, and that the best result can be obtained by using all three in conjunction. This is not the case in the negation task, though, where the surface features are counter-productive. In both cases, however, the gain in ranker performance obtained from the 'rule-like' linguistic feature family is particularly noteworthy as our current system includes only four such features, suggesting that subtle syntactic configurations are an important component for scope resolution.

As discussed in Section 5 searching the best-ranked parses can greatly increase the number of aligned constituents and thus improve the upper-bound potential of the ranker. We therefore experimented with *training* using the *first* aligned constituent in the n-best derivations. At the same time we varied the m-best derivations used in *testing*, employing features from *all* m derivations. We found that performance did not vary greatly, but that the best result for both speculation and negation was achieved for $n = 1$ and $m = 3$. (Such optimization over n-best lists of ERG parses will play a much greater role in a hybrid approach to scope resolution developed in Section 6.4 below.) Note that all our models use the default regularization parameter computed by SVM^{light}.

6.3 Ranker Performance

Table 2 summarizes the cross-validated performance of the constituent ranker on the speculation data, coupled with the 'default scope' baseline in the case of unparsed items. Our default scope is a simple heuristic that assumes the scope starts with the cue and extends to the end of the sentence. The table also lists the performance of the dependency analysis-based scope resolution rules manually developed by Øvrelid et al. (2010). Note that we extended the rules detailed in Øvrelid et al. (2010) to account for negation. This involved formulating a small set of additional rules targeting parts-of-speech not covered by the existing rule set: determiners (*no*), nouns (*lack*, *failure*), and a special treatment of negative adverbs (*not*) whose scope ranges over the verbal structure to which they are attached. In order to apply the rules we also parse the data with the stacked dependency parser of Øvrelid et al. (2010).

We note that the constituent ranker is superior to the dependency rules on BS but inferior on BSE (though still much stronger than the baseline). We attribute this to BSE being comprised entirely of papers and therefore tending to contain more complex constructions (with both coverage and quality of parses being correspondingly lower), whereas BS is dominated by abstracts. Table 3 lists the results for negation, where the dependency rules consistently outperform the ranker, although it is again stronger than the baseline.

Data	Configuration	F_1
BS	Default	64.89
	Constituent Ranker	73.67
	Dependency Rules	73.39
	Combined	78.69
BSE	Default	46.95
	Constituent Ranker	59.15
	Dependency Rules	66.60
	Combined	69.60

Table 2: Speculation scope resolution on gold cues

Data	Configuration	F_1
BS_{dev}	Default	48.07
	Constituent Ranker	67.46
	Dependency Rules	70.11
	Combined	73.98
BS_{eval}	Default	51.92
	Constituent Ranker	65.39
	Dependency Rules	68.75
	Combined	71.15

Table 3: Negation scope resolution on gold cues

6.4 A Hybrid Approach

Both our constituent ranker and the dependency analysis-based rules of Øvrelid et al. (2010) perform well in isolation. However, they do not necessarily perform well on the same test items. Consequently we investigated the effects of combining their predictions in a new hybrid set-up. When ERG parses are available we apply the ranker as before, but also add separate features incorporating the predictions of the rule-based system (in the manner described in Section 4.2). In cases where no ERG parse is available, we simply retain the prediction of the dependency rules (instead of reverting to the default scope as before).

As the addition of the rule prediction features to the ranker can influence the effectiveness of using multiple parses, we repeated our examination of the effects of using the best-ranked parses for training (n) and testing (m). Figure 2 plots the effect for speculation. We found that this combined set-up performs best for $n = 5$ and $m = 20$ for speculation and $n = 25$ and $m = 20$ for negation.

Tables 2 and 3 compare the performance of each approach across each data set when resolving the scope of gold speculation and negation cues. The results indicate that the combined approach consistently outperforms both the dependency rules and the constituent ranker in isolation.

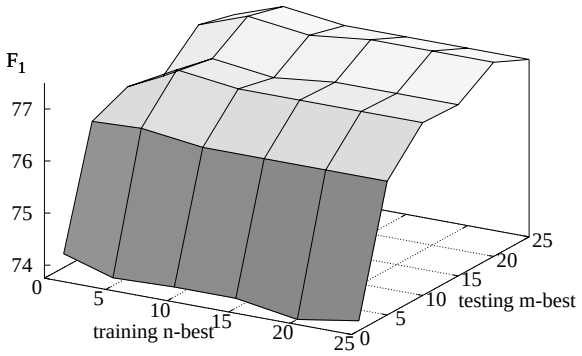

Figure 2: Cross-validated F_1 scores of the ranker combined with the dependency rules over gold cues for parsed sentences from BSP, varying the maximum number of parses employed for training and testing.

A cursory error analysis conducted over aligned items in BSP indicated that there are a number of instances where the predicted scope is correct according to the BioScope annotation guidelines, while the gold annotation is incorrect. In this typical example our system prediction (indicated with { }) includes the subject of the passive verb, whereas the gold annotation does not (indicated with | |):

(3) ...{both RAG1 and RAG2 |⟨might⟩ have evolved from a transposase (TPase) that ...|}

We also note some instances where the 'rule-like' features are activated on the correct constituent, but the ranker nevertheless selects a different candidate. In a strictly rule-based system, these features would act as hard constraints and yield superior results in these cases. Therefore, these instances seem a prime source of inspiration for further improvements to the ranker in future work.

6.5 End-to-End Evaluation

The results given so far have assumed *gold standard cues*, in order to focus on the scope component. To be able to compare our results to those of other systems, we here report on end-to-end experiments using the state-of-the-art cue classifier described by Velldal (2011). This is a simple SVM-based token-classifier that approaches cue detection as a disambiguation problem restricted to words that have previously been observed as

System	Prec	Rec	F1
Morante et al. (2010)	59.62	55.18	57.32
Combined	62.00	57.02	59.41

Table 4: End-to-end evaluation on held-out speculation data (BSE), compared to the top performer of the CoNLL-2010 Shared Task.

Data	Configuration	Rec/PCS
BSA (10-Fold)	Morante et al. (2009b)	66.07
	Combined	73.03
BSP (Held-out)	Morante et al. (2009b)	41.00
	Combined	71.55

Table 5: End-to-end evaluation on negation following the methodology of Morante et al. (2009b).

cues in the training data (i.e. effectively treating the set of cue words as a closed class), and using only simple n-gram features over lemmas and forms.The classifier currently achieves a cue-level F_1 of 80.80 for speculation in BSE and 96.00 for negation in BS_{eval}.

Table 4 lists the end-to-end performance of our system (SVM cue classification with combined constituent ranking and dependency rules for scope resolution) when analyzing speculation, compared to the best-performing system in the CoNLL-2010 shared task (Morante et al., 2010). In terms of both precision and recall, our hybrid system achieves superior performance on BSE.

In the case of negation, comparing results to other systems is difficult due to differences in evaluation metrics and data-splits. However, in order to facilitate comparison with the results of Morante and Daelemans (2009b), we have attempted to replicate their experimental set-up as closely as possible, performing 10-fold cross-validation on the BioScope abstracts and held-out testing on the papers (training on the abstracts). The results[1] are given in Table 5, showing the recall measure only, which corresponds to the Percentage of Correct Scopes (PCS) reported by Morante and Daelemans (2009b). While our hybrid approach compares very favorably to the system of Morante and Daelemans (2009b), it should be noted that the results are not strictly comparable since we have reserved 10% of the data for held-out testing on BS_{eval}. Furthermore, comparison on the papers subset is made slightly unfair by the fact that our system has likely benefited from optimization on this data during development.

For this reason, the results reported by Councill et al. (2010) when performing 5-fold cross-validation on the BSP data, can perhaps here constitute a more similar point of comparison. When training and testing on BSP, Councill et al. (2010) achieve a PCS score of 53.7. This can be compared to the recall of 70.38 obtained in our own end-to-end experiments with 10-fold cross-validation on BSP_{dev} (although this included additional sentences from BSA in training, again making the comparison somewhat unfair, but still the best comparison possible at this time).

7 Conclusion

This paper has presented a novel approach to the task of resolving the in-sentence scope of speculation and negation. Whereas most existing research applies sequence-labeling techniques we instead estimate an SVM-based discriminative ranking function for selecting subtrees from HPSG-based constituent structures. We showed that while this technique achieves good performance on its own, combining it with an existing rule-based system operating over dependency parses achieves the best performance yet on the CoNLL-2010 Shared task of speculation detection (to the best of our knowledge). The same combined approach is readily adapted to also resolve the scope of instances of negation, again achieving state-of-the-art performance.

An important premise of our subtree ranking approach, as discussed in Section 5, is that scopes align with constituents. Thus we expect improvements in our alignment procedure, as well as in the underlying domain-adapted parse selection model, to carry through to improve the overall performance of our approach.

Among the many feature types employed by the ranker, those which encode 'rule-like' linguistic knowledge stand out as contributing most substantially to ranker performance. As these fea-

[1] As the results reported in Morante and Daelemans (2009b) were inaccurate, we instead refer to values obtained from personal communication with the authors.

tures were derived from the BioScope annotation manual it is not clear whether they are portable to other domains or scope relation tasks. Future work will investigate how similar features may be discovered automatically.

Acknowledgments

We are grateful to the creators of the BioScope corpus and the organizers of the 2010 CoNLL Shared Task. We thank Andrew MacKinlay and Dan Flickinger for their help in adapting ERG parse selection to the biomedical domain. We also thank the anonymous reviewers and our colleagues for their valuable comments. Large-scale experimentation was conducted with the TITAN HPC facilities at the University of Oslo.

References

Ulrich Callmeier. 2002. Preprocessing and encoding techniques in PET. In Stephan Oepen, Daniel Flickinger, J. Tsujii, and Hans Uszkoreit, editors, *Collaborative Language Engineering. A Case Study in Efficient Grammar-based Processing*. CSLI Publications, Stanford, CA.

Isaac G. Councill, Ryan McDonald, and Leonid Velikovich. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, Uppsala, Sweden.

Richard Farkas, Veronika Vincze, György Móra, Janos Csirik, and György Szarvas. 2010. The CoNLL 2010 Shared Task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the 14th Conference on Natural Language Learning*, Uppsala, Sweden.

Dan Flickinger. 2002. On building a more efficient grammar by exploiting types. In Stephan Oepen, Dan Flickinger, Jun'ichi Tsujii, and Hans Uszkoreit, editors, *Collaborative Language Engineering: A Case Study in Efficient Grammar-based Processing*. CSLI Press.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM International Conference on Knowledge Discovery and Data Mining*, Alberta, Canada.

Halil Kilicoglu and Sabine Bergler. 2010. A high-precision approach to detecting hedges and their scopes. In *Proceedings of the 14th Conference on Natural Language Learning*, Uppsala, Sweden.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kanu, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*.

Andrew MacKinlay, Rebecca Dridan, Dan Flickinger, Stephan Oepen, and Timothy Baldwin. 2011. Using external treebanks to filter parse forests for parse selection and treebanking. In *In Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand.

Roser Morante and Walter Daelemans. 2009a. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the BioNLP 2009 Workshop*, Boulder, Colorado.

Roser Morante and Walter Daelemans. 2009b. A metalearning approach to processing the scope of negation. In *Proceedings of the 13th Conference on Natural Language Learning*, CO, USA.

Roser Morante, Vincent Van Asch, and Walter Daelemans. 2010. Memory-based resolution of in-sentence scope of hedge cues. In *Proceedings of the 14th Conference on Natural Language Learning*, Uppsala, Sweden.

Marek Rei and Ted Briscoe. 2010. Combining manual rules and supervised learning for hedge cue and scope detection. In *Proceedings of the 14th Conference on Natural Language Learning*, Uppsala, Sweden.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a robust Part-of-Speech tagger for biomedical text. In *Advances in Informatics*. Springer, Berlin, Germany.

Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2010. Resolving speculation: MaxEnt cue classification and dependency-based scope rules. In *Proceedings of the 14th Conference on Natural Language Learning*, Uppsala, Sweden.

Erik Velldal. 2011. Predicting speculation: A simple disambiguation approach to hedge detection in biomedical literature. *Journal of Biomedical Semantics*, 2(5).

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. *BMC Bioinformatics 2008*, 9(11).

Yi Zhang, Stephan Oepen, and John Carroll. 2007. Efficiency in unification-based n-best parsing. In *Proceedings of the 10th International Conference on Parsing Technologies*, Prague, Czech Republic.

Lilja Øvrelid, Erik Velldal, and Stephan Oepen. 2010. Syntactic scope resolution in uncertainty analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China.