Corpus-based approaches to processing the scope of negation cues: an evaluation of the state of the art

Roser Morante, Sarah Schrauwen and Walter Daelemans
CLiPS-CL University of Antwerp
{Roser.Morante,Walter.Daelemans}@ua.ac.be
Sarah.Schrauwen@student.ua.ac.be

Abstract

In this paper we summarize existing work on the recently introduced task of processing the scope of negation and modality cues; we analyse the scope model that existing systems can process, which is mainly the model reflected in the annotations of the biomedical corpus on which the systems have been trained; and we point out aspects of the scope finding task that would be different based on observations from a corpus from a different domain and nature.

1 Introduction

Negation and modality are complex aspects of the semantics of language. *Modality* was introduced by Jespersen (1924), who distinguishes between two categories of mood that later have been named as *deontic modality* and *epistemic modality*. Lyons (1996) describes *epistemic modality* as concerned with matters of knowledge and belief, "the speaker's opinion or attitude towards the proposition that the sentence expresses or the situation that the proposition describes". Palmer (1986) defines it as expressing the speaker's degree of commitment to the truth of a proposition. *Polarity* is a discrete category that can take two values: positive and negative. Positive polarity is used by speakers to put information as a fact in the world, whereas negative polarity is used to put information as a counterfact, a fact that does not hold in the world. Negation is a linguistic resource used to express negative polarity.

Although the treatment of these topics in computational linguistics is relatively new compared to other areas like machine translation, parsing or semantic role labeling, incorporating information about modality and polarity has been shown to be useful for a number of applications, such as biomedical text processing (Di Marco and Mercer, 2005; Chapman et al., 2001), opinion mining and sentiment analysis (Wilson et al., 2005), recognizing textual entailment (Snow et al., 2006), and automatic style checking (Ganter and Strube, 2009). In general, the treatment of modality and negation is very relevant for computational applications that process factuality (Saurí, 2008). For example, information extraction systems may be confronted with fragments of texts like the one presented in (1)[1], which contains two negation cues[2] (*not, un-*) and one speculation cue (*likely*) that affect the factuality of the events being expressed:

(1) The atovaquone/proguanil combination has **not** been widely used yet in West Africa so it is **unlikely** that the patient was initially infected with an atovaquone-resistant strain.

So far two main tasks have been addressed within the natural language processing (NLP) community: (i) the detection of various forms of polarity and modality and (ii) the identification of the scope of negation and modality cues. In this paper we reflect on the achievements of the recently introduced scope finding task (Section 2), we analyse the scope model that existing systems can process (Section 3), and we point out aspects of the scope finding task that would be different based on observations from a corpus from a different domain (Section 4).

[1] Example to be found in http://www.biomedcentral.com/content/pdf/1475-2875-1-1.pdf [last consulted 8-10-2010]

[2] A *cue* is the lexical marker that expresses negation or modality.

2 Achievements in scope processing

In the last years, several corpora have been annotated with information related to modality and polarity, which have made it possible to develop machine learning systems. Annotation has been performed at different levels: word (Hassan and Radev, 2010), expression (Baker et al., 2010; Toprak et al., 2010), sentence (Medlock and Briscoe, 2007), event (Saurí and Pustejovsky, 2009), discourse relation (Prasad et al., 2006), text (Amancio et al., 2010), and scope of negation and modality cues (Vincze et al., 2008). Thanks to the existence of the BioScope corpus, the scope processing task was recently introduced. BioScope is a freely available resource, that consists of three parts of medical and biological texts annotated with negation and hedge cues and their scope.

The scope processing task is concerned with determining at a sentence level which tokens are affected by modality and negation cues. It was first modelled as a classification problem by Morante et al. (2008). Later on several systems have been trained on the same corpus (Morante and Daelemans, 2009; Özgür and Radev, 2009; Agarwal and Yu, 2010; Li et al., 2010). Councill et al. (2010) process scopes of negation cues in a different corpus of product reviews, but this corpus is not publicly available.

The CoNLL Shared Task 2010 on *Learning to detect hedges and their scope in natural language text* (Farkas et al., 2010) boosted research on the topic. It consisted of identifying sentences containing uncertainty and recognizing speculative text spans inside sentences. Participating systems would, for example, produce the tagged sentence in (2)[3], in which *propose, suggest* and *possible* are identified as hedge cues and their scope is marked in agreement with the gold standard.

(2) We $[_{propose}$ **propose** that the existence of the alternative alignments, specific to distinct groups of genes, $[_{suggest}$ **suggests** presence of different synchronization modes between the two organisms and $[_{possible}$ **possible** functional decoupling of particular physiological gene networks in the course of evolution $_{possible}]_{suggest}]_{propose}]$.

The best system (Morante et al., 2010) for hedge scope finding in the CoNLL ST 2010 scores 57.32 F-score. Although the results are lower than the scores obtained in other well established tasks (i.e. semantic role labeling, dependency parsing), we can say that setting the first step towards automatic scope processing is an achievement. However, it can be useful to revise the characteristics of the scopes that the systems learn to process, not from a technical machine learning perspective, but from the linguistic annotation perspective, since the annotation model that systems learn determines the quality of the system output and the knowledge that can be inferred from the scopes.

3 Scope model based on the BioScope corpus

Most existing scope labelers have been trained on the BioScope corpus. Thus, the model of scope that these systems learn is determined by the characteristics of scope as they have been annotated in BioScope. Additionally, the systems have been trained for a specific domain, biomedical texts, but it might be the case that negation and speculation cues require different annotation specifications for texts from other domains. In this section we analyze the characteristics of the scope model in the BioScope corpus based on the guidelines (BioScope, 2008) and we propose some changes for further annotation work that we are carrying out. We mark in italics the statements from the BioScope guidelines and we comment on them.

– *The scope is always a continuous sequence of tokens and the cue is included in the scope*. Although most scopes in the corpus are continuous, examples such as (3), in which sentence adverbs do not belong to the scope, suggest that the scopes should be annotated as discontinuous if necessary:

(3) $[_{not}$ The number of glucocorticoid receptors per cell (Ro) and the binding affinity (Kd) for dexamethasone were$_{not}]$, however, $[_{not}$**not** significantly different $_{not}]$

– *Scopes can be determined on the basis of syntax and they extend to the biggest unit possible. If necessary, complements and adjuncts are included in the scope*. It would be useful to furhter specify how different syntactic constructions (coordination, subordination, etc.) should be annotated.

[3] In the examples below, cues will be marked in bold and their scope between brackets.

– *The scope of negative auxiliaries, adjectives and adverbs usually starts with the cue and ends at the end of the phrase, clause or sentence*. In (4) the scope extends to the right of *not*. In our view, the scope should include the subject because the subject contributes to the meaning of the event being negated. If, as Lyons (1996) suggest, we paraphrase the negative connective in (4) with the formula *it is not the case that*, we obtain (5), where the subject is under the scope of the formula.

(4) Once again, the Disorder module does $[_{not}$ **not** contribute positively to the prediction $_{not}]$

(5) Once again, it is not the case that the Disorder module does contribute positively to the prediction

– *Passive voice changes the scope of the cue because the object of the active construction is the subject of the passive construction*. According to the BioScope guidelines, the scope of *not* in 6 and 7 would be annotated differently. As indicated above, we consider that the subject of the active sentence is also under the scope of the negation, so in our view both sentences should be analyzed equally.

(6) $[_{not}$ Levels of RNA coding for the receptor were **not** modulated by exposure to high levels of ligand $_{not}]$

(7) Exposure to high levels of ligand does $[_{not}$ **not** modulate levels of RNA coding for the receptor $_{not}]$

– *Negative conjunctions generally scope over the syntactic unit whose members it coordinates. However, if the complex negative keyword occurs within the subject of the sentence, its scope is extended to the whole sentence*. (8) is the example provided in the guidelines, but paraphrasing the sentence with the *it is not the case* formula as in (9) shows that the subject should also be included in the scope.

(8) In contrast, sodium salicylate (1 mM) inhibited $[_{neither-nor}$**neither** adhesion **nor** expression of these adhesion molecules $_{neither-nor}]$

(9) In contrast, it is not the case that sodium salicylate (1 mM) inhibits either adhesion or expression of these adhesion molecules

– *Prepositions scope over the following (noun) phrase*. (10) is the example provided in the guidelines, where *without* scopes over a noun phrase. Nevertheless, *without* can be followed by a verb phrase, as in (11). In this case, one could argue that the logical subject of the verb should be included in the scope of the preposition, since the negation can be paraphrased as in (12).

(10) $[_{without}$ Mildly hyperinflated lungs **without** focal opacity $_{without}]$

(11) $[_{without}$ CD28 costimulation $_{without}]$ augments IL-2 secretion of activated lamina propria T cells by increasing mRNA stability $[_{without}$ **without** enhancing IL-2 gene transactivation $_{without}]$

(12) It is not the case that CD28 costimulation enhances IL-2 gene transactivation

Possible improvements in the BioScope annotation model are pointed out in Vincze (2010), namely the treatment of elliptic constructions, and discontinuous and intersecting scopes. An additional improvement would be to annotate affixal negation. We consider that (13) is equivalent to (14) and should receive the same analysis, since they can be paraphrased as in (15):

(13) Actually, $[_{un}$ tRNASec and tRNAPyl have **un**usual secondary structures 515 $_{un}]$

(14) Actually, $[_{not}$ tRNASec and tRNAPyl do **not** have usual secondary structures 515 $_{not}]$

(15) Actually, it is not the case that tRNASec and tRNAPyl have usual secondary structures 515

4 Annotating scopes in a different domain

The existing scope labelers have been trained on biomedical texts. However, it is reasonable to expect that texts from other domains contain different phenomena that would affect the systems performance. We are currently analysing negations and their scopes in a complete different corpus, *The Hound of the Baskervilles* (HB) by Conan Doyle. This corpus has been annotated with coreference and semantic roles for the *SemEval Task Linking Events and Their Participants in Discourse* (Ruppenhofer et al., 2010), and will be further annotated with negation and modality cues. Phenomena in this corpus show that whereas the scope of cues can be determined in a similar way as it is determined in biomedical texts, identifying

negation cues in certain contexts, which is the first part of the scope finding task, is not only a matter of lexical lookup:

– Not all negative affixes are negation cues. For example the affix *un-* in *unspoken* does not negate its root morpheme. *Unspoken* does not mean 'not spoken', but 'understood without the need for words'. Consequently, in (16) *unspoken* is not a negation cue.

(16) All my unspoken instincts, my vague suspicions, suddenly took shape and centred upon the naturalist

– Fixed expressions like *could not help* in the sentence below do not negate the modified event.

(17) Why about Sir Henry in particular? I could not help asking

– Negation words in tag questions do not have a negation function, but a pragmatic function, since the speaker seeks confirmation from the addressee. A similar case are negation words in dialogue checks like *don't you think* in (19).

(18) You have been inside the house, have you not, Watson?

(19) Don't you think, Watson, that you are away from your charge rather long?

– Negation words in exclamative particles do not have a negation function. In (20), *don't tell me* does not express a negated event. This is a multiword construction used to express surprise.

(20) "Don't tell me that it is our friend Sir Henry!"

– Some modality cues, such as *no doubt*, contain false negation cues. In (21) *no doubt* is a fixed expression that expresses certainty, no event is negated. It is an expression that acts at the discourse level conveying information about the attitude of the speaker towards his statement.

(21) Partly it came no doubt from his own masterful nature, which loved to dominate and surprise those who were around him

– The context influences the effect of the negation cue. The volitive verb *wish* in (22) and the conditional construction in (23) cancel the negative effect of *not*.

(22) Your mission to-day has justified itself, and yet I could almost wish that you had not left his side

(23) In fact, if you had not gone to-day it is exceedingly probable that I should have gone to-morrow

5 Conclusions and future work

In this paper we have briefly presented the achievements in processing the scope of negation and modality cues. There are currently several systems that can process scopes in biomedical texts, however there is a lack of annotated resources, since there is only one publicly available corpus. We have also pointed out that the quality of the systems output depends not only on the technical aspects of the systems, but also on the linguistic model contained in the annotations. Based on annotation work on a literary corpus, we have pointed out some difficulties that existing systems could face in detecting cues.

We are currently annotating texts by Conan Doyle with negation cues and their scopes. For defining the guidelines we take the model of the BioScope corpus as a starting point and we include modifications based on the observations made above. The annotated corpus and the guidelines will be publicly available.

Apart from annotating more data, further work will focus on computing the factuality of statements based on the scopes of negation and modality cues and other contextual features, and studying the interaction between negation and modality.

Acknowledgements

This study was made possible through financial support from the University of Antwerp (GOA project BIOGRAPH). We would like to thank four anonymous reviewers for their suggestions.

References

Agarwal, S. and H. Yu (2010). Detecting hedge cues and their scope in biomedical literature. *Journal of Biomedical Informatics 710.016/j.jbi.2010.08.003*.

Amancio, D. R., R. Fabbri, O. N. Oliveira Jr., M. Nunes, and L. Costa (2010, July). Distinguishing between positive and negative opinions with complex network features. In *Proc. of TextGraphs-5 - 2010 Workshop on Graph-based Methods for Natural Language Processing*, Uppsala, Sweden, pp. 83–87. ACL.

Baker, K., M. Bloodgood, B. Dorr, N. Filardo, L. Levin, and C. Piatko (2010). A modality lexicon and its use in automatic tagging. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valetta, Malta, pp. 1402–1407. European Language Resources Association (ELRA).

BioScope (2008). Annotation guidelines. http://www.inf.u-szeged.hu/rgai/project/nlp/bioscope/Annotation guidelines2.1.pdf.

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

Councill, I., R. McDonald, and L. Velikovich (2010, July). What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proc. of the Workshop on Negation and Speculation in Natural Language Processing*, Uppsala, Sweden, pp. 51–59. University of Antwerp.

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

Farkas, R., V. Vincze, G. Szarvas, G. Móra, and J. Csirik (Eds.) (2010, July). *Proc. of the Fourteenth Conference on Computational Natural Language Learning*. Uppsala, Sweden: ACL.

Ganter, V. and M. Strube (2009). Finding hedges by chasing weasels: Hedge detection using wikipedia tags and shallow linguistic features. In *Proc. of the ACL-IJCNLP 2009 Conference Short Papers*, Suntec, Singapore, pp. 173–176.

Hassan, A. and D. Radev (2010, July). Identifying text polarity using random walks. In *Proc. of the 48th Annual Meeting of the ACL*, Uppsala, Sweden, pp. 395–403. ACL.

Jespersen, O. (1924). *The philosophy of grammar*. London: Allen and Unwin.

Li, J., Q. Zhu, and G. Zhou (2010). A unified framework for scope learning via simplified shallow semantic parsing. In *Proc. of EMNLP 2010*.

Lyons, J. (1996). *Semantics*. Cambridge: CUP.

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

Morante, R. and W. Daelemans (2009). Learning the scope of hedge cues in biomedical texts. In *Proc. of BioNLP 2009*, Boulder, Colorado, pp. 28–36.

Morante, R., A. Liekens, and W. Daelemans (2008). Learning the scope of negation in biomedical texts. In *Proc. of the EMNLP 2008*, Honolulu, Hawaii, pp. 715–724.

Morante, R., V. Van Asch, and W. Daelemans (2010, July). Memory-based resolution of in-sentence scopes of hedge cues. In *Proc. of the Fourteenth Conference on Computational Natural Language Learning*, Uppsala, Sweden, pp. 40–47. ACL.

Özgür, A. and D. Radev (2009). Detecting speculations and their scopes in scientific text. In *Proc. of EMNLP 2009*, Singapore, pp. 1398–1407.

Palmer, F. (1986). *Mood and modality*. Cambridge, UK: CUP.

Prasad, R., N. Dinesh, A. Lee, A. Joshi, and B. Webber (2006). Annotating attribution in the penn discourse treebank. In *SST '06: Proc. of the Workshop on Sentiment and Subjectivity in Text*, Morristown, NJ, USA, pp. 31–38. ACL.

Ruppenhofer, J., C. Sporleder, R. Morante, C. Baker, and M. Palmer (2010, July). Semeval-2010 task 10: Linking events and their participants in discourse. In *Proc. of the 5th International Workshop on Semantic Evaluation*, Uppsala, Sweden, pp. 45–50. ACL.

Saurí, R. (2008). *A factuality profiler for eventualities in text*. Ph. D. thesis, Waltham, MA, USA.

Saurí, R. and J. Pustejovsky (2009). FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation 43*(3), 227–268.

Snow, R., L. Vanderwende, and A. Menezes (2006). Effectively using syntax for recognizing false entailment. In *Proc. of HLT NAACL*, Morristown, NJ, USA, pp. 33–40. ACL.

Toprak, C., N. Jakob, and I. Gurevych (2010, July). Sentence and expression level annotation of opinions in user-generated discourse. In *Proc. of the 48th Annual Meeting of the ACL*, Uppsala, Sweden, pp. 575–584. ACL.

Vincze, V. (2010, July). Speculation and negation annotation in natural language texts: what the case of bioscope might (not) reveal. In *Proc. of the Workshop on Negation and Speculation in Natural Language Processing*, Uppsala, Sweden, pp. 28–31. University of Antwerp.

Vincze, V., 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.

Wilson, T., P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan (2005). Opinionfinder: a system for subjectivity analysis. In *Proc. of HLT/EMNLP on Interactive Demonstrations*, Morristown, NJ, USA, pp. 34–35. ACL.