A review corpus annotated for negation, speculation and their scope

Natalia Konstantinova[1], Sheila C.M. de Sousa[1], Noa P. Cruz[2], Manuel J. Maña[2], Maite Taboada[3] and Ruslan Mitkov[1]

[1]Research Group in Computational Linguistics, University of Wolverhampton (UK)
[2]Departamento de Tecnologías de la Información, Universidad de Huelva (Spain)
[3]Simon Fraser University, Department of Linguistics (Canada)
E-mail: {n.konstantinova,sheila.castihomonteirodesousa, R.Mitkov}@wlv.ac.uk
{noa.cruz,manuel.mana}@dti.uhu.es
mtaboada@sfu.ca

Abstract

This paper presents a freely available resource for research on handling negation and speculation in review texts. The SFU Review Corpus, consisting of 400 documents of movie, book, and consumer product reviews, was annotated at the token level with negative and speculative keywords and at the sentence level with their linguistic scope. We report statistics on corpus size and the consistency of annotations. The annotated corpus will be useful in many applications, such as document mining and sentiment analysis.

Keywords: corpus, annotation, negation and speculation

1. Introduction

Processing negation and speculation can be useful for several NLP applications such as information extraction, paraphrasing, recognizing textual entailment, opinion mining and sentiment analysis.

For example, many authors have studied the role of negation in the sentiment analysis task, where it is one of the most common linguistic means that can lead to a change in polarity. Councill et al. (2010) describe a system that can identify exactly the scope of negation in free text. The authors concluded that performance was improved dramatically by introducing negation scope detection. In more recent work, Dadvar et al. (2011) investigated the problem of determining the polarity of sentiment in movie reviews when negation words occur in the sentences. The authors also observed significant improvements on the classification of the documents after applying negation detection.

Distinguishing between objective and subjective facts is also crucial for sentiment analysis. Speculation is a linguistic expression that tends to correlate with subjectivity (also known as private state). Pang & Lee (2004) showed that subjectivity detection in the review domain helps to improve polarity classification.

To the best of our knowledge, there are no publicly available standard corpora of reasonable size from the review domain annotated with negation and speculation. This motivated our annotation of the SFU corpus, which is widely used in the domain of sentiment analysis and opinion mining.

In this paper, we expand the previous work developed by Konstantinova & de Sousa (2011). The corpus is now annotated in its entirety and available, and the annotation guidelines are fully developed as well.

The paper is organised in the following way: Section 2 describes related research; Section 3 focuses on corpus characteristics and provides some statistics of the SFU review corpus. Section 4 provides more details about annotation guidelines and also presents statistics about negation and speculation cues in the annotated corpus.

Section 5 discusses agreement analysis and the most regular cases of disagreements between the annotators. Section 6 provides information about the corpus and guidelines availability. The paper finishes with conclusions and future work (Section 7).

2. Related work

Even though negation and speculation detection has gained much attention in recent years, open access annotated resources are rare and relatively small in size. Most of the work has been done for the biomedical domain, where there are several annotated corpora. The GENIA Event corpus (Kim et al., 2008) contains annotation of biological events with negation and two types of uncertainty. Medlock and Briscoe (2007) based their system on a corpus consisting of six papers from genomics literature, which were annotated for speculation. Settles et al. (2008) constructed a corpus where sentences were classified as either speculative or definite; however, no keywords were marked in the corpus. Vincze et al. (2008) developed standard corpora of reasonable size with information about negative/speculative keywords and their scope.

The research community is trying to explore other domains as well: Morante et al. (2011a) discuss the need for corpora which cover different domains than biomedical. The authors point out that the existing guidelines should be adapted to new domains and they annotated literary texts by Conan Doyle, although only for the case of negation (Morante et al., 2011b).

We are aware of only one corpus in the review domain described in Councill et al. (2010), however it was annotated only for negation, but not speculation. This

corpus is also rather small in size, containing only 2111 sentences in total, out of which 679 contain negation.
Therefore, our corpus is the first one with an annotation of negative/speculative information and their scope in the review domain.

3. Corpus characteristics

The Simon Fraser University Review corpus (Taboada et al., 2006) was chosen for our annotation of negation and speculation. This corpus consists of 400 documents (50 of each type) of movie, book, and consumer product reviews from the website Epinions.com. Each text was assigned a label based on whether it is a positive or negative review. All the texts differ in size and are written by different people (more information about the size of the corpus can be found in Table 1). As shown in this table, there are appreciable differences in the length of the documents depending on the domain but not in the length of sentences, so sentence complexity in the entire corpus is comparable.

Domain	#Sentences	Av. length document	#Words	Av. length sentences
Books	1,596	31.92	32,908	20.61
Cars	3,027	60.54	58,481	19.32
Computers	3,036	60.72	51,668	17.01
Cookware	1,504	30.08	27,323	18.16
Hotels	2,129	42.58	40,344	18.95
Movies	1,802	36.04	38,507	21.36
Music	3,110	62.20	54,058	17.38
Phones	1,059	21.18	18,828	17.77
Total	**17,263**	**43.15**	**303,289**	**17.56**

Table 1: Statistics of the SFU Review Corpus. Av. length document is shown in number of sentences. Av. length sentences are shown in number of words.

4. Annotation guidelines

The entire corpus was annotated by one linguist. A second linguist annotated 10% of the documents, randomly selected and in a stratified way, with the aim of measuring inter-annotator agreement (Section 5 provides more details about this analysis).
The guidelines presented in this paper have been adapted from the existing Bioscope corpus guidelines (Vincze et al., 2008) in order to fit the needs of the review domain.

4.1 General remarks

There are several general principles to be followed when annotating negation and speculation:

- Only sentences with some instance of speculative language or negation should be considered.
- Questions should not be annotated at all.
- The min-max strategy should be followed during annotation, following the BioScope corpus guidelines (Vincze et al., 2008):
  - When annotating keywords, try to choose the minimal unit which expresses negation or speculation (special attention should be paid to distinguishing complex cues and sequences of several keywords).
  - When annotating scope, try to annotate the maximum number of words affected by the phenomenon.
- Cue words are not included in the scope.
- Transitional words (e.g. *in addition*, *not to mention*, etc.) should not be included in the scope.
- When unsure of the scope, annotate only a keyword.
- When unsure what category the keyword should be assigned to (whether it expresses negation or speculation), use the 'undecided' label.

As mentioned earlier we did not agree with the BioScope guidelines completely and introduced some modifications. These main changes are summarised below:

- We did not include cue words in their scope;
- A different scheme for annotating coordination was used;
- Embedded scopes were quite a frequent case;
- We had a case of ‘no scope’ both in the case of negation and speculation.

More information about the differences with the BioScope principles can be found in Konstantinova & de Sousa (2011).

The nature of the review domain texts introduces a greater possibility of encountering difficult cases than in the biomedical domain. Some of these special cases are discussed in Section 5. More detail can be found in the full version of the guidelines (see Section 6).

4.2 Negation

Statistical analysis of the annotated corpus revealed that out of the total amount of 17,263 sentences 18% contained negation cues. However it should be noted that this proportion of negation cues varies slightly depending on the domain as shown in Table 2.

Domain	# Cues	% Negated sentences
Books	406	22.7
Cars	576	17.1
Computers	590	17.2
Cookware	376	21.3
Hotels	387	16.3
Movies	490	23.7
Music	470	13.4
Phones	232	19.5
Total	**3527**	**18.1**

Table 2: Negation statistics in the SFU Review Corpus

The total amount of distinct negation cues in our corpus amounted to 53, with the top 10 most frequent cues shown in Table 3. It is interesting to note that the first two cues for negation (*not* and *no*) constitute more than 55% of the total frequency of all the cues found in the corpus, while the remaining 51 cues cover only 45%.

Cue	Frequency	Percentage
Not	1419	40.23
No	524	14.85
Don't	296	8.39
Never	248	7.03
Doesn't	154	4.36
Without	151	4.28
Didn't	119	3.37
Isn't	89	2.52
Can't	68	1.92
Wasn't	57	1.61

Table 3: The most frequent negation keywords in the SFU Review Corpus

4.3 Speculation

In the case of speculation, the statistical analysis described in Table 4 shows that, out of a total of 17,263 sentences, around 22% are speculative. Therefore the proportion of speculative sentences in the annotated corpus is higher than negative ones. This can be explained by the nature of the corpus which consists of reviews, which are subjective and where speculation is extensively used to express opinions.

Domain	#Cues	%Speculative sentences
Books	370	17.2
Cars	1068	26.0
Computers	944	23.2
Cookware	583	27.3
Hotels	695	23.7
Movies	648	26.0
Music	643	15.1
Phones	408	27.4
Total	**5359**	**22.7**

Table 4: Speculation statistics in the SFU Review corpus

More than 100 different cues were used in our corpus for expressing speculation. This number is considerably higher than the number of negation cues encountered during the annotation. In addition, as described in Table 5, the amount of occurrences of each cue was equally distributed across all the cues, so the top most frequent cues did not represent the majority of speculation cases as happened for negation.

Cue	Frequency	Percentage
If	876	16.34
Or	820	15.30
Can	765	14.27
Would	594	11.08
Could	299	5.57
Should	213	3.97
Think	211	3.93
May	157	2.92
Seems	150	2.79
Probably	121	2.25

Table 5: The most frequent speculative keywords in the SFU Review Corpus

5. Agreement analysis

Initially, 20% of the corpus was annotated by the first annotator. This study showed that it is not possible to follow the BioScope annotation guidelines and special guidelines for this domain should be developed. Therefore the knowledge acquired during this initial annotation was used to adapt the existing guidelines and study problematic cases. We then used the developed guidelines to annotate the whole corpus. In order to reveal some possible weaknesses of the annotation, another expert annotator was involved at a later stage. The second annotator worked with 10% of the documents from the original collection, selected randomly. The annotation was done according to the guidelines used by the first annotator. During the annotation process, the annotators were not allowed to communicate with each other. However, after the annotation was finished a disagreement analysis was carried out and the two annotators met to discuss the guidelines and the most problematic cases. This stage helped to refine and finalize the guidelines, which are freely available online (see Section 6). The corpus annotation carried out by the first linguist was corrected at the final stage in order to ensure that it follows the established version of the guidelines.

In addition to the described research we also measured inter-annotator agreement using F-measure and Kappa, treating the second annotator as the gold standard. Table 6 illustrates the results obtained for inter-annotator agreement regarding the scope for both negation and speculation in terms of F-measure. The results are slightly higher than those reported for the subcollection of full papers of the BioScope corpus[1]. As can be seen from Table 6, the speculation phenomenon is more problematic for annotation and more prone to disagreements. This is due to the fact that speculation is a fuzzy category (Konstantinova and De Sousa, 2011). Table 6 also illustrates that the left scope is easier for annotators to agree on and the right one poses problems and drops the results for the full scope F-measure. In most cases the left scope started just after the cue word and therefore if the annotators agreed on the cue word there was only a small amount of disagreement about the left scope. However, the right scope was more difficult to decide on as it depended on the understanding of the text and the annotators' decision about the part of the text affected by the phenomenon, for both negation and speculation.

In addition to F-measure, we calculated the inter-annotator agreement in terms of Kappa (Cohen and Jacob, 1960); results are shown in Table 7. In the case of scope, we counted the agreements (at word level) between the two linguists for all scopes in the sentences that have negation or speculation cues. We considered two types of agreement: (1) Both linguists annotated the word as belonging to the same scope and (2) The word was annotated as being outside of any scope.

The agreement presented in Table 7 is considered quite high (Landis and Koch, 1977) and therefore we can be confident that the corpus is annotated correctly, and that the annotation is reproducible.

The next subsection will provide some insight into the cases of disagreement revealed during the analysis of annotations done by the two annotators.

	F-measure	
	Negation	Speculation
Cues	**92.79**	**89.18**
Full scope	**81.88**	**70.20**
Left scope	**97.17**	**88.04**
Right scope	**82.11**	**78.78**

Table 6: Inter-annotator agreement in terms of F-measure

	Kappa	
	Negation	Speculation
Cues	**0.927**	**0.890**
Scope	**0.872**	**0.867**

Table 7: Inter-annotator agreement in terms of Kappa

5.1 Disagreement cases

As mentioned in the previous section, disagreement cases were analyzed and both annotators decided on correct guidelines for every aspect. Most of the disagreement cases were simply the result of human error, when one of the annotators accidentally missed a word or included a word that did not belong either in the scope or as a part of a cue word.

However, other cases of disagreement can be explained mostly by the lack of clear guidelines about some issues at the beginning of the annotation.

As noted earlier, speculation is a difficult phenomenon not always clear to identify because its notion is fuzzy and therefore prone to misinterpretation. When cases of disagreement were caused by a different understanding of the phenomenon, further discussion between the two annotators helped to achieve a consistent annotation.

Cases involving the cue words *appear* and *seem* were a common source of disagreement due to the lack of clear initial guidelines. Afterwards, the annotators agreed that when the object of the sentence was also modified by the cue word, it should be included in the scope. The following example illustrates a case with the word *appear* where the keyword is included in square brackets, and the scope in curly brackets.

[1] See http://www.inf.u-szeged.hu/rgai/bioscope

Example [1]: Alex Cross is back as a D.C. detective who again pairs up with the FBI to help solve a series of horrific murders {that [appear]$_{spec}$ to have been committed by vampires} .

Here, the object *that* should be included in the scope of *appear* as it is modified by the cue word. This becomes obvious after transforming a sentence from the passive voice into active voice. Thus, the sentence in example [1] can be reformulated into *It appears that the murders have been committed by vampires.*

In general, scope was one of the biggest causes of disagreement in the annotations (Table 6). Example 3 (below) illustrates a problematic case with several keywords and difficult sentence structure. Consider the two annotations below made by annotator 1 (A1) and annotator 2 (A2):

Example [3]: **(A1)** Well , [if]$_{spec}$ {you 're an Alex Cross follower , you [might]$_{spec}$ {as well read it}} because you're [probably] $_{spec}$ {already hooked} (as I am) and will want to read the next one (which there almost certainly will be--why [not] [if] $_{spec}$ {he can get away with marketing this amateurish crap and still stay on the bestseller list} ?) .

(A2): Well , [if]$_{spec}$ {you 're an Alex Cross follower} , you [might]$_{spec}$ {as well read it because you 're [probably] $_{spec}$ {already hooked} (as I am) and will want to read the next one}} (which there almost certainly will be--why [not] [if] $_{spec}$ {he can get away with marketing this amateurish crap and still stay on the bestseller list} ?) .

In this example we have two main disagreement points: while A1 considered as scope of the cue word *if* the part *you're an Alex Cross follower, you [might] as well read it*, A2 considered only *you're an Alex Cross follower*. Also A1 considered the scope for *might* only the part '*as well read it'*, while A2 included more: *as well read it because you're [probably] already hooked (as I am) and will want to read the next one*.
In this case it was agreed that the annotation of A1 was more accurate since the cue word *if* modifies the sentence up to *as well read it*, thus, having an embedded scope with the cue word *might*, and also the scope for *might* should not include the part *because you're [probably] already hooked (as I am) and will want to read the next one*. It becomes clearer that *because…* is not affected by the cue word *might* if one adds a period right after *read it*. The sentence would not change its meaning.
All the above mentioned cases were discussed by the two annotators in order to find a common point of view reflected in the guidelines. Once the final guidelines were established the corpus annotation was corrected in order to produce a consistently annotated corpus.

6. Corpus availability

The corpus is available, in plain raw text and annotated form, from the SFU Review Corpus site (http://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html). The detailed guidelines for the annotators can be found there as well.

7. Conclusions

We have presented a freely available corpus in the review domain, annotated for negation, speculation and their scope. The annotation followed exhaustive guidelines, and was validated through an inter-annotator reliability study. We believe that the guidelines are sound, and that the corpus will be useful for sentiment analysis, negation recognition, and many other tasks in text analysis.

8. Acknowledgements

The authors wish to thank the reviewers for their comments, which helped to enhance the paper. This work also benefited from the input of Veronika Vincze and Wilker Aziz, who provided valuable ideas and assistance to our research.

9. References

Councill, I.; McDonald, R. and Velikovich, L. (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.*

Dadvar, M.; Hauff, C. and Jong, de F. (2011). Scope of negation detection in sentiment analysis. *Dutch-Belgian Information Retrieval Workshop*. Amsterdam, the Netherlands.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20* (1): 37–46.

Konstantinova, N. and De Sousa, S. (2011). Annotating negation and speculation: the case of the review domain. In *Student Workshop of the International Conference on Recent Advances in Natural Language Processing*. Hissar, Bulgaria.

Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33** (1): 159–174.

Medlock, B. and Briscoe, T. (2007). Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the ACL*, pages 992– 999, Prague, Czech Republic.

Morante, R.; Schrauwen, S. and Daelemans, W. (2011a). Corpus-based approaches to processing the scope of negation cues: an evaluation of the state of the art. In *Proceedings of the Ninth International Conference on Computational Semantics, IWCS '11*, pages 350–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

Morante, R.; Schrauwen, S. and Daelemans, W. (2011b). Annotation of negation cues and their scope. *Guidelines v1.0. Technical Report Series CTR-003*, CLiPS, University of Antwerp, Antwerp.

Settles, B.; Craven, M. and Friedland, L. (2008). Active learning with real annotation costs. *In Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, pages 1–10.

Taboada, M.; Anthony, C. and Voll, K. (2006). Methods

for creating semantic orientation dictionaries. *In Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 427–432, Genoa, Italy.

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