Negation and Speculation Detection for Improving Information Retrieval Effectiveness

Noa P. Cruz Díaz
Universidad de Huelva
E.T.S de ingeniería. Ctra. Palos de la Frontera s/n. 21819. Palos de la Frontera (Huelva)
noa.cruz@dti.uhu.es

The thesis proposed here intends to assist information retrieval and text mining tasks through the negation and speculation detection focusing on two different areas. In the biomedical domain, the existence of an annotated corpus with this kind of information has made possible the development of an effective system to automatically detect these language forms. In the review domain, we have annotated for negation, speculation and their scope a set of reviews.

Information retrieval. Negation and speculation detection. Biomedical and review domains.

1. INTRODUCTION

Negation and speculation modify the meaning of the phrases in their scope. This means, negation denies or rejects statements transforming a positive sentence into a negative one, e.g., “It isn’t scary, but it is enthralling”. Speculation is used to express that some fact is not known with certainty, e.g., “The treatment seems to be successful”. These two phenomena are interrelated (de Haan, 1997) and have similar characteristics in the text.

Identifying these linguistic phenomena is a very important problem for a wide range of information retrieval (IR) and text mining tasks. We focus on two domains of preference: biomedical domain and review domain. In the first one, negation/speculation detection can help in tasks like Protein-Protein interaction or Drug-Drug interactions whereas in the second one, opinion mining, sentiment analysis and polarity identification are examples of improvable tasks through the identification of negation/speculation. In both domains we develop negation/speculation detection systems based on machine learning techniques with the aim to improve the effectiveness of these kinds of applications.

In the biomedical domain, there are many machine learning approaches developed on detecting negative and speculative information. In addition, most of them use the BioScope corpus (Vincze et al., 2008) which is the same collection used in our experiments.

One of the most representative works in this regard is the research conducted by Morante and Daelemans (2009a). The performance showed for the system in all the sub-collections of the BioScope corpus is high. The authors (2009b) extended their research to include speculation detection showing that the same scope finding approach can be applied to both negation and speculation.

Also using the BioScope corpus, recently, Velldal et al. (2012) combined manually crafted rules with machine learning techniques. The results obtained by this system can be considered as the state-of-the-art.

However, our combination of novel features together with the classification algorithm choice improves the results to date for the sub-collection of clinical documents (Cruz et al., 2012).

On the other hand, the impact of negation and speculation detection on sentiment analysis has not been sufficiently considered compared to the biomedical domain. In fact, we are not aware of any available standard corpora of reasonable size annotated with negation and speculation. This issue together with the fact that identification of this kind of information in reviews can help the opinion mining task motivated our work of annotation of the SFU Review Corpus (Konstantinova et al., 2012) which is the first one with an annotation of negative/speculative information and their linguistic scope in the review domain. It will allow us to develop a system in the same way we did for the biomedical domain.

2. WORK DONE

2.1 Biomedical domain

Proceedings of the 5th BCS IRSG Symposium on Future Directions in Information Access

The main goal of the thesis is focused on developing a system based on machine learning techniques that identifies negation and speculation cues and their scope for improving the effectiveness of IR tasks. The system is trained and evaluated on the clinical texts of the BioScope corpus. This part represents the major portion of the corpus and is the densest in negative and speculative information.

Our system is modelled in two consecutive classification phases. In the first one, a classifier decides whether each token in a sentence is a cue or not. In the second one, another classifier decides, for every sentence that has cues, if the other words in the sentence are inside or outside the scope of each cue.

We used different features in each of the two phases into which the task was divided. They encode information about the cue, the paired token, their contexts and the tokens between. As classification algorithms, we experimented with Naïve Bayes, C4.5 and Support Vector Machine.

We trained and evaluated the system with the sub-collection of clinical documents of the BioScope corpus. This was done by randomly dividing the sub-collection into three parts, using two thirds for training and one third for evaluating.

The results obtained are higher than those previously published for clinical documents. Cruz et al. (2012) show a complete description of the system and an extensive analysis of these results.

2.2 Review domain

The novelty in this work is derived from the annotation of the SFU corpus with negation and speculation information.

This corpus is widely used in the field of sentiment analysis and opinion mining and consists of 400 reviews from the website Epinions.com. In total, more than 17,000 sentences were annotated by one linguistic adapting the existing Bioscope corpus guidelines in order to fit the needs of the review domain.

We followed the general principles used to annotate the BioScope corpus. However, we introduced main changes which are summarized below:

(i) We do not include cue words in their scope.
(ii) A different scheme for annotating coordination is used.
(iii) Embedded scopes are quite a frequent case.
(iv) We have a case of 'no scope' both in negation and speculation.

In addition, the nature of the review domain texts introduces a greater possibility of encountering difficult cases than in the biomedical domain. More detail about these cases can be found in Konstantinova et al. (2012).

With the aim of measuring inter-annotator agreement and correcting these problematic cases, a second linguist annotated 10% of the documents, randomly selected and in a stratified way. The agreement obtained is consider high.

This corpus is freely downloadable[1] and the annotation guidelines are fully available as well.

3. FUTURE WORK

Future research directions include improving the performance of the systems both in the biomedical domain and in the review domain. We will carry this out in two aspects. Firstly, in the cue detection phase we plan to use external lexicons. Secondly, in the scope detection phase, it will be necessary to explore new features derived from deeper syntactic analysis because as Huang and Lowe notes (2007), structure information stored in parse trees helps identifying the scope or as Vincze (2008) points out, the scope of a cue can be determined on the basics of syntax. In fact, initial results obtained with the SFU corpus using features extracted via dependency graphs are good and improvable in the future by adding more syntactic information.

As a last point, in the review domain, we plan to explore if correct annotation of negation/speculation improves the results of the SO-CAL system (Taboada et al., 2008; Taboada et al., 2011) using our system as a recognizer for this kind of information, rather than the search heuristics that the SO-CAL system is currently using. On the other hand, in the biomedical domain, we intend to integrate negation/speculation detection in a clinical record retrieval system. An initial work in this direction can be found in Cordoba et al. (2011).

4. REFERENCES

Córdoba, JM; et al. Medical-Miner at TREC 2011 Medical Records Track. TREC. 2011.

Cruz Diaz, N; Maña Lopez, M; Vazquez, J; Álvarez V. 2012. A machine-learning approach to negation and speculation detection in clinical texts. JASIST, 63(7), 1398–1410.

[1] http://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html

De Haan F. 1997. The interaction of modality and negation: a typological study. Garland Publishing, New York, USA.

Huang, Y., and Lowe, H.J. (2007). A novel hybrid approach to automated negation detection in clinical radiology reports. Journal of the American Medical Information Association, 14(3), pages 304–311.

Konstantinova, N; de Sousa, S; Cruz, N; Maña, M; Taboada, M; Mitkov, R. 2012. A review corpus annotated for negation, speculation and their scope. Proceedings of the Eight International Conference on LREC. Istanbul (Turkey). May 2012. European Language Resources Association (ELRA).

Morante, R; Daelemans, W. 2009a. A metalearning approach to processing the scope of negation. Proceedings of the Thirteenth CoNLL. Sofia (Bulgaria). August 2009. Association for Computational Linguistics. p. 21–29.

Morante, R; Daelemans, W. 2009b. Learning the scope of hedge cues in biomedical texts. Proceedings of the Workshop on Current Trends in BioNLP. Association for Computational Linguistics, Boulder (Colorado). June 2009. p. 28–36.

Taboada M., Anthony C., Brooke J., Grieve J. and Voll, K. 2008. SO-CAL: Semantic Orientation CALculator. Simon Fraser University, Vancouver.

Taboada M., Brooke J., Tofiloski M., Voll K. and Stede M. 2011. Lexicon-Based Methods for Senti-ment Analysis. Computational Linguistics 37 (2), pages 267-307.

Velldal,E; Ovrelid, L; Read, J; Oepen, S. 2012. Speculation and negation: Rules, rankers, and the role of syntax. Computational Linguistics, 38(2), 369-410.

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