Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 107-110 recibido 28-11-14 revisado 27-01-15 aceptado 10-02-15

Negation and Speculation Detection in Clinical and Review Texts[1]

Detección de la Negación y la Especulación en Textos Médicos y de Opinión

Noa P. Cruz Díaz
Dpto. de Tecnologías de la Información. Universidad de Huelva
Ctra. Huelva - Palos de la Frontera s/n. 21819 La Rábida (Huelva)
noa.cruz@dti.uhu.es

Abstract: PhD Thesis written by Noa P. Cruz Díaz at the University of Huelva under the supervision of Dr. Manuel J. Maña López. The author was examined on 10th July 2014 by a committee formed by the doctors Manuel de Buenaga (European University of Madrid), Mariana Lara Neves (University of Berlin) and Jacinto Mata (University of Huelva). The PhD Thesis was awarded Summa cum laude (International Doctorate).

Keywords: Negation and speculation detection, machine learning, biomedicine, sentiment analysis.

Resumen: Tesis doctoral realizada por Noa P. Cruz Díaz en la Universidad de Huelva bajo la dirección del Dr. Manuel J. Maña López. El acto de defensa tuvo lugar el jueves 10 de julio de 2014 ante el tribunal formado por los doctores Manuel de Buenaga (Universidad Europea de Madrid), Mariana Lara Neves (Universidad de Berlín) y Jacinto Mata (Universidad de Huelva). Obtuvo mención internacional y la calificación de Sobresaliente Cum Laude por unanimidad.

Palabras clave: Detección de la negación y la especulación, aprendizaje automático, biomedicina, análisis de sentimientos.

1 Introduction

Negation and speculation are complex expressive linguistic phenomenas which have been extensively studied both in linguistic and philosophy (Saurí, 2008). They modify the meaning of the phrases in their scope. Negation denies or rejects statements transforming a positive sentence into a negative one, e.g., Mildly hyperinflated lungs without focal opacity. Speculation, also known as hedging, it is used to express that some fact is not known with certainty, e.g., Atelectasis in the right mid zone is, however, possible. These two phenomenas are interrelated (De Haan, 1997) and have similar characteristics in the text: they both have scope, so affect part of the text which is denoted by the presence of negation or speculation cue words.

Nowadays, negation and speculation detection is an emergent task in Natural Language Processing (henceforth, NLP). In recent years, several challenges and shared tasks have included the extraction of these language forms such as the BioNLP'09 Shared Task 3 (Kim et al., 2009), the CoNLL-2010 Shared Task (Farkas et al., 2010) or the SEM 2012 Shared Task (Morante and Blanco, 2012).

Detecting uncertain and negative assertions is relevant in a wide range of applications such as information extraction (henceforth, IE), interaction detection, opinion mining, sentiment analysis, paraphrasing and recognising textual entailment (Farkas et al., 2010; Konstantinova et al., 2012; Morante and Daelemans, 2009a; Morante and Daelemans, 2009b). For all of these tasks it is crucial to know when a part of the text should have the opposite meaning (in the case of negation) or should be treated as subjective and non-factual (in the case of speculation). This part of the text is what is known as scope.

At first glance, negation and speculation might seem easy to deal with. The problem could be broken down into finding negative and hedge cues and determining their scope. However, it is much more problematic.

[1] This thesis can be downloaded from http://www.sepln.org/wp-content/uploads/2014/09/NEGATION-AND-SPECULATION-Q9.pdf

ISSN 1135-5948

Negation and speculation play a remarkable role towards understanding text and pose considerable challenges. They interact with many other phenomenas and they are used for so many different purposes that a deep analysis is needed (Blanco and Moldovan, 2011b).

This thesis is focused on the two domains in which negation and hedging have drawn more attention: the biomedical domain and the review domain. In the first one, negation and speculation detection can help in tasks like Protein-Protein interaction or Drug-Drug interaction. This particular area has been the focus of much current research, mainly due to the availability of the BioScope corpus (Szarvas et al., 2008); a collection of clinical documents, full papers and abstracts annotated for negation, speculation and their scope. In the review domain; opinion mining, sentiment analysis and polarity identification are examples of improvable tasks through the identification of negation and speculation. In all these tasks, distinguishing between objective and subjective facts is crucial and therefore negative and speculative information must be taken into account. Despite its importance and the interest of some authors to explore other areas apart from biomedical (Morante and Daelemans, 2012), the impact of negation and speculation detection in the review domain has not been sufficiently considered compared to the biomedical domain.

2 Contributions

The aim of this thesis is to contribute to the ongoing research on negation and speculation in the Language Technology community. In the medical domain, a system based on machine-learning techniques that identifies negation and speculation cues and their scope in clinical texts is proposed (Cruz et al., 2012).

Additionally, and due to the tokenization problems encountered during the pre-processing of the BioScope corpus and the lack of guidance in this respect, this thesis closely describes this issue and provide both a comprehensive overview analysis and evaluation of tokenization tools. This means, the first comparative evaluation study of tokenizers in the biomedical domain which could help developers to choose the best tokenizer to use.

In the sentiment analysis and opinion mining domains, and contrary to what happens in the biomedical field, there are no publicly available standard corpora of reasonable size annotated with negation and hedging. Therefore, the first step was the participation in the annotation process of the SFU Review corpus with negative and speculative keywords and their linguistic scope. It represents the first corpus annotated with this kind of information in the review domain. Next, using the corpora previously described as well as following the approach used in the biomedical domain, a system to automatically detect negation and hedge cues and their scope is presented.

3 Structure of the thesis

An outline of the thesis is described below.

Chapter 2 begins with an introduction to the definition of negation and speculation from different perspectives, including a classification of the different types of each of them. After briefly motivating the importance of processing these language forms, this chapter presents the related work that inspired and motivated our work, both in the biomedical domain and in sentiment analysis.

Chapter 3 is dedicated to the tokenization problem in the biomedical domain with the aim of helping developers in the decision of choosing the best tokenizer to use. Therefore, this chapter provides an analysis of the problematic cases that the nature of the biomedical field introduces as well as a comprehensive comparative study of the available tools. Finally, it includes the evaluation of the 2 tokenizers that show better features and more accuracy and consistency in the previous study.

Chapter 4 is an in-depth description of the negation and speculation detection system for the clinical domain, explaining every step of the development process. It also presents the corpora used to build the system that accompanies it. Finally, this chapter describes how the system is evaluated and gives details about the experimentation, showing the results obtained and the discussion and error analysis around them.

Chapter 5 presents the developed system for the negation and hedging detection in review texts. It includes the description of the corpora used to train and test the system and the methodology followed. The corpora have been previously annotated for this task so their annotation process is also specified. It describes the evaluation process; the experiments performed as well as it details the system performance. A discussion and error analysis are also presented in this chapter.

Chapter 6 sums up the outcomes of the work done in this thesis and discuss the possibilities for future work.

4 Conclusions

This thesis tackles negation and speculation treatment in computational linguistics in the two fields which have received more attention: biomedical and review.

In the biomedical domain, a machine-learning system that identifies the negation/speculation cues and their scope in clinical texts has been developed, using the clinical sub-collection of the BioScope corpus as a learning source and for evaluation purposes. For this reason, the proposed approach may not be generalisable to other domains because the expectations in terms of effectiveness could be different if it was used in a corpus with other features, such as scientific texts. The proposed approach achieves an F_1 of 97.3% and 94.9% in negation and speculation cue detection, respectively. In the scope recognition, the system reports F_1 values of 90.9% in negation and 71.9% in speculation. These results show the superiority of the machine-learning-based approach regarding the use of regular expressions. In fact, in the detection of negation expressions, the developed system outstrips the F_1 of NegEx (Chapman et al., 2001) by 30%. In speculation, the proposed method beats the F_1 of the best system by more than 10%. In addition, compared to other approaches based on machine-learning techniques, the developed global system correctly determines approximately 20% more than the scopes identified by Morante and Daelemans (2009b) in negation. In speculation, this difference is greater and the proposed approach correctly recognises nearly twice the number of scopes identified by Morante and Daelemans (2009a). This means improving the results to date for the sub-collection of clinical documents. However, much still remains to be done since scope detector performance is far from having reached the level of well established tasks such as parsing, especially in speculation detection.

Also in the biomedical field, this thesis includes a comprehensive overview study of tokenization tools. Choosing the right tokenizer in this domain is a non-trivial task so this contribution aims to provide a valuable guideline for NLP developers in the biomedical field to select the appropriate tokenizer as the first phase of a text mining task. Specifically, all the biomedical domain difficulties, together with what could be considered to be the correct tokenization in each of these difficult cases are detailed. The process followed to create the list of tools for tokenizing texts to analyse is also explained, including a description of the technical, functional and usability criteria employed to asses each of these tokenizers. After analyzing 21 tools according to the criteria, 13 of them are tested on a set of 28 sentences from the BioScope corpus. Finally, the two tokenizers that show better features and more accuracy and consistency in the examples tested in the previous phase are evaluated in a subset of sentences of this corpus. This contribution means, as far as we are aware, the first comparative evaluation carried out on tokenizers in the biomedical field.

In the review domain, although negation and speculation recognition can help to improve the effectiveness of sentiment analysis and opinion mining tasks, there is just a few works on detecting negative information. Besides, there is, as far as we are aware, no work in identifying speculation. Therefore, this thesis aims to fill this gap through the development of a system which automatically identifies both negation and speculation keywords and their scope. It means the first attempt to detect speculation in the review domain. The novelty of this contribution also lies in the fact that, to the best of our knowledge, this is the first system trained and tested on the SFU Review corpus (Konstantinova et al., 2012). This corpus is extensively used in opinion mining and consists of 400 documents annotated with negative and speculative information. Overall, the results are competitive and the system is portable. In fact, the results reported in the cue detection task (92.37% and 89.64% in terms of F_1 for negation and speculation, respectively) are encouragingly high. In the case of the speculation, the results are comparable to those obtained by a human annotator doing the same task. In the scope detection task, the results are promising and the system correctly identifies 80.26% full scopes in negation and 71.43% in speculation. The proposed approach outstrips the baseline by as much as about 20% in the negation cue detection and improves it up by roughly 13% in scope detection.

Acknowledgements

This thesis has been funded by the University of Huelva (PP10-02 PhD Scholarship), the Spanish Ministry of Education and Science (TIN2009-14057-C03-03 Project) and the Andalusian Ministry of Economy, Innovation and Science (TIC 07629 Project).

References

Blanco, E. and Moldovan, D. I. 2011b. Some issues on detecting negation from text. FLAIRS Conference.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F. and Buchanan, B. G. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34, 301-310.

Cruz Díaz, N. P., Maña López, M. J., Vázquez, J. M. and Álvarez, V. P. 2012. A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the American Society for Information Science and Technology*, 63(7), 1398-1410.

De Haan, F. 1997. *The interaction of modality and negation: A typological study Taylor and Francis.*

Farkas, R., Vincze, V., Móra, G., Csirik, J. and Szarvas, G. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. En *Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task*, páginas 1-12.

Kim, J., Ohta, T., Pyysalo, S., Kano, Y. and Tsujii, J. 2009. Overview of BioNLP'09 shared task on event extraction. En *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, páginas 1-9.

Konstantinova, N., de Sousa, S., Cruz, N., Maña, M. J., Taboada, M. and Mitkov, R. 2012. A review corpus annotated for negation, speculation and their scope. LREC, 3190-3195

Morante, R. and Blanco, E. 2012. * SEM 2012 shared task: Resolving the scope and focus of negation. En *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, páginas 265-274.

Morante, R. and Daelemans, W. 2009a. Learning the scope of hedge cues in biomedical texts. En *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, páginas 28-36.

Morante, R. and Daelemans, W. 2009b. A metalearning approach to processing the scope of negation. En *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, páginas 21-29.

Morante, R. and Daelemans, W. 2012. ConanDoyle-neg: Annotation of negation in conan doyle stories. En *Proceedings of the Eighth International Conference on Language Resources and Evaluation*.

Saurí, R. 2008. *A factuality profiler for eventualities in text.*

Szarvas, G., Vincze, V., Farkas, R. and Csirik, J. 2008. The BioScope corpus: Annotation for negation, uncertainty and their scope in biomedical texts.