Detecting Negated and Uncertain Information in Biomedical and Review Texts

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

Abstract

The thesis proposed here intends to assist Natural Language Processing tasks through the negation and speculation detection. We are focusing on the biomedical and review domain in which it has been proven that the treatment of these language forms helps to improve the performance of the main task. In the biomedical domain, the existence of a corpus annotated for negation, speculation and their scope has made it possible for the development of a machine learning system to automatically detect these language forms. Although the performance for clinical documents is high, we need to continue working on it to improve the efficiency of the system for scientific papers. On the other hand, in the review domain, the absence of an annotated corpus with this kind of information has led us to carry out the annotation for negation, speculation and their scope of a set of reviews. The next step in this direction will be to adapt it to this domain for the system developed by the biomedical area.

1 Introduction

Negation and speculation are complex expressive linguistic phenomena which have been extensively studied both in linguistic and philosophy (Saurí, 2008). They 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., "*Mildly hyperinflated lungs without focal opacity*". Speculation 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 phenomena are interrelated (de Haan, 1997) and have similar characteristics in the text.

From a natural language processing (NLP) perspective, identification of negation and speculation is a very important problem for a wide range of applications such as information extraction, interaction detection, opinion mining, sentiment analysis, paraphrasing and recognizing textual entailment.

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 implies that a simple approach like a bag of words could be not enough so an in-depth analysis of the text would be necessary. Therefore, for improving the effectiveness of these kinds of applications, we aim to develop negation/speculation detection systems based on machine learning techniques. We focus on two domains of preference: biomedical domain and review domain.

In the biomedical domain, there are many machine learning approaches developed on detecting negative and speculative information due to the availability of the BioScope corpus, a collection of clinical documents, full papers and abstracts annotated for negation, speculation and their scope (Vincze et al., 2008), which is the same collection used in our experiments.
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).

However, the research community is trying to explore other areas such as sentiment analysis where distinguishes between objective and subjective information is also crucial and therefore must be taken into account. For example, Morante et al. (2011) discuss the need for corpora which covers different domains apart from biomedical. In fact, we are not aware of any available standard corpora of reasonable size annotated with negation and speculation in this area. This issue together with the fact that identification of this kind of information in reviews can help the opinion mining task motivated our work

Proceedings of the Student Research Workshop associated with RANLP 2013, pages 45–50,
Hissar, Bulgaria, 9-11 September 2013.

of annotation of the SFU Review Corpus (Konstantinova et al., 2012). This means that this corpus is the first one with an annotation of negative/speculative information and their linguistic scope in the review domain. In addition, it will allow us to develop a negation/speculation detection system in the same way we did for the biomedical domain.

With the aim of presenting the work carried out and the further work to be done, in my thesis in this respect, the structure of the paper has been divided in the following: Section 2 outlines related research; Section 3 describes the goals achieved in the biomedical and review domain. Section 4 discusses the future research directions in both domains. The paper finishes with the conclusions (Section 5).

2 Related Work

In the biomedical domain, which is the main focus of the thesis, there are many approaches developed on detecting negative and speculative information because of their benefits to the NLP applications. These approaches evolve from rule-based ones to machine learning techniques.

Among the first types of research, the one developed by Chapman et al. (2001) stands out. Their algorithm, NegEx, which is based on regular expressions, determines whether a finding or disease mentioned within narrative medical reports is present or absent. Although the algorithm is defined by the authors themselves as simple, it has proven to be powerful in negation detection in discharge summaries. The reported results of NegEx showed a precision of 84.5%, recall of 77.8% and a specificity of 94.5%. In 2007, the authors developed an algorithm called ConText (Chapman et al., 2007), an extension of the NegEx negation algorithm, which identify the values of three contextual features (negated, historical or hypothetical and experienced). In spite of its simplicity, the system performed well at identifying negation and hypothetical status.

Other interesting research works based on regular expressions are that of Mutalik et al. (2001), Elkin et al. (2005) and Huang and Lowe (2007) who were aware that negated terms may be difficult to identify if negation cues are more than a few words away from them. To address this limitation in automatically detecting negations in clinical radiology reports, they proposed a novel hybrid approach, combining regular expression matching with grammatical parsing. The sensitivity of negation detection was 92.6%, the PPV was 98.6% and the specificity was 99.8%.

However, the most recent works are based on machine-learning approaches. In addition, most of them use the BioScope corpus 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). Their machine-learning system consists of five classifiers. The first one decides if the tokens in a sentence are negation cues or not. Four classifiers are used to predict the scope. Exactly, three of them determine whether a token is the first token, the last, or neither in the scope sequence and the last one uses these predictions to determine the scope classes. The set of documents used for experimentation was the BioScope corpus. The performance showed for the system in all the sub-collection of the corpus was high, especially in the case of clinical reports. The authors (2009b) extended their research to include speculation detection. They showed that the same scope-finding approach can be applied to both negation and speculation. Another recent work is that developed by Agarwal and Yu (2010). In this work, the authors detected negation cue phrases and their scope in clinical notes and biological literature from the BioScope corpus using conditional random fields (CRF) as machine-learning algorithm. The best CRF-based model obtained good results in terms of F-score both for negation and speculation detection task. Also using the BioScope corpus, recently, Velldal et al. (2012) explored two different syntactic approaches to resolve the task. One of them uses manually crafted rules operating over dependency structures while the other automatically learns a discriminative ranking function over nodes in constituent trees. The results obtained by the combination of the 2 approaches can be considered as the state-of-the-art.

On the other hand, the impact of negation and speculation detection on sentiment analysis, which is the other goal of this thesis, has not been sufficiently considered compared to the biomedical domain.

Some authors have studied the role of negation. For example, Councill et al. (2010) described a system that can exactly identify the scope of negation in free text. The authors concluded that the 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. Lapponi et al. (2012) reviewed different schemes for representing negation and presented a state-of-the-art system for negation detection. By employing different configurations of their system as a component in a testbed for lexical-based sentiment classification, they demonstrated that the choice of representation has a significant effect on the performance.

For its part, speculation has not received much attention perhaps because of the absence up to this point of a corpus annotated with this information. However, it should be treated in the future because authors such as Pang and Lee (2004) showed that subjectivity detection in the review domain helps to improve polarity classification.

3 Work Done

3.1 Biomedical Domain

The machine-learning system developed for negation and speculation detection was trained and evaluated on the clinical texts of the BioScope corpus. This is a freely available resource consisting of clinical documents, full articles and abstracts with annotation of negative and speculative cues and their scope. The sub-collection of clinical documents represents the major portion of the corpus and is the densest in negative and speculative information. More specifically, it contains 1,954 documents formed by a clinical history section and an impression section, the latter, used by the radiologist to describe the diagnosis obtained from the radiographies. In terms of the percentage of negation and speculation cues, it represents 4.78% of the total of words in the sub-collection. In the others, this percentage is only about 1.7%.

Our system was modeled in two consecutive classification phases. In the first one, a classifier decided whether each token in a sentence was a cue or not. More specifically, with the aim of finding complex negation cues formed by more than one word, the classifier determined if the tokens ere at the beginning, inside or outside of the cue. In the second phase, another classifier decided, for every sentence that had cues, if the other words in the sentence were inside or outside the scope of the cue. This means repeating the process as many times as cues appeared in the sentence.

We used different sets of novel features in each of the two phases into which the task was divided. They encoded information about the cue, the paired token, their contexts and the tokens between.

As classification algorithms, we experimented with Naïve Bayes and C4.5 (Quinlan, 1986) implemented in Weka (Witten & Frank, 2005). Authors such as Garcia, Fernandez and Herrera (2009) have shown its competitiveness in terms of accuracy and its adequacy for imbalanced problems. We also used Support Vector Machine (SVM) implemented in LIBSVM (Chang and Lin, 2001) because this classifier has proven to be very powerful in text classification tasks as described by Sebastiani (2002).

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 in negation, due to the complexity of the speculation detection task, are higher than those obtained in speculation. However, our combination of novel features together with the classification algorithm choice achieves good performance values in both cases. What's more, these results are higher than those previously published.

Cruz et al. (2012) show a complete description of the system and an extensive analysis of these results.

3.2 Review Domain

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

This corpus is widely used in the field of sentiment analysis and opinion mining and consists of 400 documents (50 of each type) of movie, book, and consumer product reviews from the website Epinions.com. All the texts differ in size, are written by different people and have been assigned a label based on whether it is a positive or negative review. In total, more than 17,000 sentences were annotated by one linguistic who followed the general principles used to annotate the BioScope corpus. However, in order to fit the needs of the review domain, we introduced main changes which are summarized below:

- Keywords: Unlike the BioScope corpus, where the cue words are annotated as part of the scope, for the SFU corpus we decided not to include the cue words in the scope.

- Scope: When the annotator was unsure of the scope of a keyword only the keyword was annotated.
- Type of keyword: When the annotator was unsure what type the keyword should be assigned to (whether it expresses negation or speculation), nothing was annotated.
- Coordination: The BioScope guidelines suggest extending the scope for speculation and negation keywords to all members of the coordination. However, in the case of the review domain as the keywords were not included in the scope, the scopes were annotated separately and then linked to the keywords.
- Embedded scopes: Although keywords are not included in their own scope, a keyword can be included in the scope of other keywords and situations of embedded scopes are possible. There were also cases when the combination of different types of keywords (i.e. negation and speculation ones) resulted in the embedded scopes.
- No scope: Unlike the BioScope guidelines which mention only the cases of negation keywords without scope, situations where speculation keywords had no scope were encountered as well in the review domain.

Konstantinova & de Sousa (2011) provide an extensive description of all different cases and also give examples illustrating these rules.

In addition, the nature of the review domain texts introduces a greater possibility of encountering difficult cases than in the biomedical domain. 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. This 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. 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.

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. More detail about theses special cases can be found in Konstantinova & de Sousa (2012).

The agreement between annotators is consider high so we can be confident that the corpus is annotated correctly and that the annotation is reproducible.

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

4 Future Work

So far, the work done in the biomedical domain includes the development of a machine-learning system to detect speculation, negation and their linguistic scope in clinical texts. As we have mentioned, the result for this sub-collection is very good, especially for negation. However, the system is not so efficient for the other sub-collections of documents due to the fact that scientific literature presents more ambiguity and complex expressions.

Therefore, future research directions include, improving the performance of the system in this case. We will carry this out in two aspects. Firstly, in the cue detection phase we plan to use external sources of information which could include external lexicon such as WordNet or Freebase. 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 competitive and improvable in the future by adding more syntactic information.

In addition, we plan to integrate negation/speculation detection in a clinical record retrieval system. An initial work in this regard can be found in Cordoba et al. (2011).

We also intend to broaden this work into different areas such as sentiment analysis where the corpus annotation described in the previous section will facilitate the training of a system to automatically detect negation and speculation in the same way as we did for the biomedical domain.

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

As a last point, we intend 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. Thus, we could measure the practical impact of accurate negation/speculation detection and check as authors like Councill (2010) affirms it helps to improve the performance in sentiment predictions.

5 Conclusions

The aim of the thesis here described is to develop a system to automatically detect negation, speculation and their scope in the biomedical domain as well as in the review domain for improving NLP effectiveness. In the case of clinical documents, the system obtains a high level of performance, especially in negation. The ambiguity in scientific papers is greater and the detection becomes more complicated. Therefore, an in-depth analysis of the text is necessary to improve performance in this case.

Finally, we plan to adapt the system developed for the biomedical area to the review domain. The first step in this aspect has been the annotation of the SFU Review Corpus (Taboada et al., 2006) with negation and speculation information.

References

Shashank Agarwal and Yu Hong. 2010. Biomedical negation scope detection with conditional random fields. *Journal of the American Medical Information Association*, 17(6), pages 696–701.

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(27), pages 1–27.

Wendy Chapman, Will Bridewell, Paul Hanbury, Gegory Cooper and Bruce Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Information*, 34(5), pages 301–310.

Wendy Chapman, David Chu, John Dowling. ConText: An algorithm for identifying contextual features from clinical text. 2007. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 81-88.

Jose Manuel Córdoba et al. Medical-Miner at TREC 2011 Medical Records Track. TREC, 2011.

Isaac 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. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP'10)*, pages 51–59.

Noa P. Cruz Díaz, Manuel J. Maña López, Jacinto Mata Vázquez and Victoria Pachón Álvarez. 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), pages 1398–1410.

Maral Dadvar, Claudia Hauff and F.M.G de Jong. 2011. Scope of negation detection in sentiment analysis. *11th Dutch-Belgian Information Retrieval Workshop (DIR 2011)*, pages 16–19.

Ferdinand de Haan. 1997. *The interaction of modality and negation: a typological study*. Garland Publishing, New York, USA.

Peter Elkin, Steven Brown, Brent Bauer, Casey Husser, William Carruth, Larry Bergstrom and Dietlind Wahner-Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC Medical Information Decision Making*, 5(1), 13.

Salvador García, Alberto Fernández and Francisco Herrera. 2009. Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems. *Applied Soft Computing*, volume 9, pages 1304–1314.

Yang Huang and Henry Lowe. 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.

Natalia Konstantinova and Sheila de Sousa. Annotating Negation and Speculation: the Case of the Review Domain. *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 139-144

Natalia Konstantinova, Sheila de Sousa, Noa Cruz, Manuel Maña, Maite Taboada and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC).*

Emanuel Lapponi, Jonathon Read, Lilja Øvrelid. 2012. Representing and resolving negation for sentiment analysis. *Proceedings of the 2012 ICDM Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction.*

Roser Morante and Walter Daelemans. 2009a. A metalearning approach to processing the scope of negation. *Proceedings of the 13th Conference on*

Computational Natural Language Learning, pages 21–29.

Roser Morante and Walter Daelemans. 2009b. Learning the scope of hedge cues in biomedical texts. *Proceedings of the Workshop on BioNLP*, pages 28–36.

Roser Morante, Sarah Schrauwen and Walter Daelemans. 2011. Corpus-based approaches to processing the scope of negation cues: an evaluation of the state of the art. *Proceedings of the Ninth International Conference on Computational Semantics*, pages 350–354.

Pradeep G. Mutalik, Aniruddha Deshpande and Prakash M. Nadkarni. 2001. Use of general purpose negation detection to augment concept indexing of medical documents: A quantitative study using the UMLS. *Journal of the American Medical Information Association*, 8(6), pages 598–609.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the ACL*, pages 271-278.

John Ross Quinlan. 1986. Induction of decision trees. *Machine Learning*, volume 1, pages 81–106.

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

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Survey*, 34(1), pages 1–47.

Maite Taboada, Caroline Anthony and Kemberly Voll. 2006. Methods for creating semantic orientation dictionaries. *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 427–432.

Maite Taboada, Caroline Anthony, Julian Brooke, Jack Grieve and Kemberly Voll. 2008. SO-CAL: *Semantic Orientation CALculator*. Simon Fraser University, Vancouver.

Taboada, Julian Brooke, Milan Tofiloski, Kemberly Voll and Manfred Stede 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics* 37 (2), pages 267-307.

Eric Velldal, Lilja Ovrelid, Jonathon Read and Stephan Oepen S. 2012. Speculation and negation: Rules, rankers, and the role of syntax. *Computational Linguistics*.

Veronica Vincze, György Szarvas, Richárd Farkas and János Csirik. 2008. The Bio-Scope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9+.

Ian H. Witten and Eibe Frank. 2005. *Data mining: Practical machine learning tools and techniques* (2nd ed.). Waltham, MA: Morgan Kaufmann.