Marker words for negation and speculation in health records and consumer reviews

Maria Skeppstedt[1,2] **Carita Paradis**[3] **Andreas Kerren**[2]
[1]Gavagai AB, Stockholm, Sweden
maria@gavagai.se
[2]Computer Science Department, Linnaeus University, Växjö, Sweden
andreas.kerren@lnu.se
[3]Centre for Languages and Literature, Lund University, Lund, Sweden
carita.paradis@englund.lu.se

Abstract

Conditional random fields were trained to detect marker words for negation and speculation in two corpora belonging to two very different domains: clinical text and consumer review text. For the corpus of clinical text, marker words for speculation and negation were detected with results in line with previously reported inter-annotator agreement scores. This was also the case for speculation markers in the consumer review corpus, while detection of negation markers was unsuccessful in this genre. Also a setup in which models were trained on markers in consumer reviews, and applied on the clinical text genre, yielded low results. This shows that neither the trained models, nor the choice of appropriate machine learning algorithms and features, were transferable across the two text genres.

1 Introduction

When health professionals document patient status, they often record common symptoms that the patient is *not* showing, or reason about possible diagnoses. Clinical texts, therefore, contain a large amount of negation and speculation (Velupillai et al., 2011).

Negations and speculations are also expressed in consumer review texts, e.g., when the reviewed artefact lacks an expected feature, or when reviewers are uncertain of their opinion. Previous research shows that the proportion of sentences containing negation and speculation is even larger in consumer review texts that in clinical texts (Vincze et al., 2008; Konstantinova et al., 2012).

The BioScope corpus was one of the first clinical corpora annotated for negation and speculation (Vincze et al., 2008). The guidelines used for the BioScope corpus have later, with only a few modifications, been used for annotating consumer review texts. A qualitative analysis of the difference between the medical genres of the BioScope corpus and consumer review texts has previously been carried in order to adapt the guidelines for the genre of review texts (Konstantinova and de Sousa, 2011). To the best of our knowledge, there are, however, no previous studies in which the same machine learning algorithm is applied to both corpora and the results are compared.

2 Background

There are other medical corpora annotated with the same guidelines as the BioScope corpus (Vincze et al., 2008), e.g., a drug-drug interaction corpus (Bokharaeian et al., 2014). There are also medical corpora annotated according to other guidelines, e.g., guidelines that include more fine-grained categories, such as weaker or stronger speculation/uncertainty (Velupillai, 2012), or whether a clinical finding is conditionally or hypothetically present in the patient (Uzuner et al., 2011). Large annotated corpora are often constructed on English medical text, e.g., the *i2b2/VA challenge on concepts, assertions, and relations* corpus, but negation and speculation has also been annotated in corpora with clinical text written in, e.g., Swedish (Velupillai, 2012) and Japanese (Aramaki et al., 2014).

Examples of non-medical corpora are the previously mentioned corpus of consumer reviews (Konstantinova and de Sousa, 2011), and literary texts annotated for negation in the *SEM shared task (Morante and Blanco, 2012).

Negations and speculations are often annotated in two steps. First, *marker* words (often also referred to as *cue* words or *keywords*) for nega-

tion/speculation are annotated, and then either the *scope* of text that the marker words affects is annotated, or whether specific *focus* words occurring in the text are affected by the marker words. Focus words could, for instance, be clinical findings that are mentioned in the same sentence as the marker words. Automatic detection of negation and speculation is typically divided into two subtasks corresponding to the two annotation steps. That is, first the marker words are detected and, thereafter, the task of determining the scope or classifying the focus words is carried out.

In this study, the first of the two subtasks of negation/speculation detection is addressed, i.e., the detection of marker words for negation and speculation. This task is typically addressed using two main approaches, either a vocabulary of negation/speculation markers is compiled and tokens in the text are compared to this vocabulary in order to determine whether they are marker words (Chapman et al., 2001; Ahltorp et al., 2014), or alternatively a machine learning model is trained.

3 Materials

Two English corpora were used in the experiments, the Bioscope corpus (Vincze et al., 2008) and the SFU Review corpus annotated for negation and speculation (Konstantinova et al., 2012).

As previously mentioned, the annotation guidelines for the SFU Review corpus were an adaption of the guidelines for the BiosScope corpus, and they were, therefore, very similar. In both corpora, marker words expressing negation and speculation were annotated, as well as their scope. The general principle for the length of text to annotate as marker words was to annotate the minimal unit of text that still expresses negation or speculation. The definition of negation used for the task was "[...] the implication of the non-existence of something", while speculation was defined as "[...] the possible existence of a thing, i.e. neither its existence nor its non-existence is unequivocally stated [...]". Marker words could either be individual words that express negation or speculation on their own, e.g., "This {may} {indicate}..", or complex expressions containing several words that do not convey negation or speculation on their own, e.g., "This {raises the question of}...".

The BioScope corpus consists of three subcorpora, containing clinical text, biological full papers and biological scientific abstracts. For this study, the subcorpus containing clinical text was used, which consists of 6,400 sentences of which 14% contains negation and 13% contains speculation. The pairwise agreement rates for the three annotators involved in the project were 91/95/96 for annotating marker words for negation and 84/90/92 for marker words for speculation.

The corpus of consumer reviews was a previously complied corpus, the SFU Review corpus, to which annotations of negation and speculation were added. The corpus contains consumer generated reviews of books, movies, music, cars, computers, cookware and hotels (Taboada and Grieve, 2004; Taboada et al., 2006). The corpus consists of 17,000 sentences, of which 18% was annotated as containing negation and 22% as containing speculation. 10% of the corpus was doubly annotated to measure inter-annotator agreement, resulting in an F-score and Kappa score of 92 for negation markers and 89 for speculation markers.

There are previous studies on the detection of speculation and negation markers in these two corpora. A perfect precision and a recall of 0.98 were obtained, when training an IGTree classifier to detect negation markers on the full paper subcorpus of the BioScope corpus and evaluating it on the clinical sub-corpus (Morante and Daelemans, 2009b). Similar results for detecting negation markers in the clinical sub-corpus were achieved by a vocabulary matching system. When using the same set-up for detecting speculation markers, i.e., training on the paper sub-corpus and evaluating on the clinical, a precision of 0.88 and a recall of 0.27 were achieved (Morante and Daelemans, 2009a). For these experiments, the token to be classified, as well as its immediate neighbouring tokens were used as features. When instead training as well as evaluating on the clinical sub-corpus (a conditional random fields model with tokens as features), a precision of 0.99 and a recall of 0.87 were achieved for detecting speculation, while a rule-base vocabulary matching system achieved a precision of 0.95 and a recall of 0.96 on this task (Agarwal and Yu, 2010). Examples of other results reported are a precision/recall of 0.97/0.98 for negation markers and 0.96/0.93 for speculation markers (Cruz Díaz et al., 2012), using a C4.5 classifier and a support vector machine.

There is also previous research on the detection of which tokens that constitute negation and speculation markers in the SFU Review corpus

(Cruz et al., 2015). Experiments were conducted in which 10-fold cross-validation was applied on the entire corpus, and a feature set that included the token and its closest neighbours was used. For the most successful machine learning algorithm (a cost-sensitive support vector machine), a precision of 0.80 and a recall of 0.98 were obtained for negation and a precision of 0.91 and a recall of 0.94 were obtained for speculation. For the two other evaluated algorithms (Naive Bayes and a support vector machine with a radial basic function kernel), much lower and slightly lower results, respectively, were obtained. Both of these two lower-performing models had problems handling multi-word markers for negation that included *n't* or *not*, and results for these two models were improved by a simple rule-based post-processing algorithm specifically designed to handle these cases.

4 Experiments

Experiments consisted of training machine learning models to recognise markers for negation and speculation and, thereafter, evaluate these models. Three setups were used: i) models trained on a subset of the BioScope corpus and evaluated on another subset of the same corpus, ii) models trained on a subset of the SFU Review corpus and evaluated on another subset of this corpus, and finally iii) models trained on the SFU Review corpus and evaluated on the BioScope corpus. The rationale for performing the last experiment was the difficulty that is often associated with getting access to large amounts of clinical text, due to the sensitive content of text belonging to this genre. If it would be possible to successfully apply a model trained on non-clinical text on the clinical text genre, this might be a possible solution in cases when the amount of available clinical data is scarce.

The text segments annotated as negation- and speculation markers were coded according to the BIO-format, i.e., a token could be the *b*eginning of, *i*nside or *o*utside of a marker segment. The approach of structured prediction was taken, and the PyStruct package was used (Müller and Behnke, 2014) to train a linear conditional random fields model, using the OneSlackSSVM class. Default parameters were used (which included a regularisation parameter of 1) and a maximum of 100 passes over the dataset to find constraints. To limit the feature set, as the models were to be trained on a limited amount of data, features were restricted to the token that was to be classified, and, in addition, a minimum of two occurrences of a token in the training data was required for it to be included. As linear conditional random fields were used, the classification of a token was dependent on the classification of the two neighbouring tokens (Sutton and McCallum, 2006), making it possible to detect multi-word markers.

For all setups, the models were trained with an increasingly larger size of training data, from 600 training instances to 3,000. In each iteration, 200 new training instances were randomly selected for inclusion in the training data. The same experiment was repeated four times, each time with a new, randomly selected, subset of held-out data to use for evaluation in setups i) and ii), and (for all experiments) new random selections of training instances. Precision, recall and F-score for recognising segments that were classified as negation- or speculation markers were measured with NLTK's ChunkScore class (Bird, 2002).

5 Results and discussion

For detecting speculation markers in the SFU Review corpus, and for detecting both speculation and negation markers in the BioScope corpus when trained on text of the same genre, the method was relatively successful (Figure 1), achieving results in line with the inter-annotator agreement.[1] For detecting negation, the increase in training data size did not affect these results, while the general trend for speculation was an improvement of results with more training samples, although results remained slightly unstable.

For detecting negation in the SFU Review corpus, on the other hand, results were much lower than the measured agreement figures. Results were consistently low for all four folds (F-scores 0.70/0.75/0.76/0.74 for 3,000 training instances), and the F-score decreased with a larger training data set due to a decrease in precision, and a recall that remained low. It could be ruled out that the low results were due to the relatively small training data size, since an additional model, trained on 8,000 samples, gave even lower results (an F-score of 0.62). Multi-token negation markers including

[1] Previous machine learning results have typically been achieved using a larger training set, and, therefore, a comparison to the agreement figures was carried out, instead of a comparison to previous results.

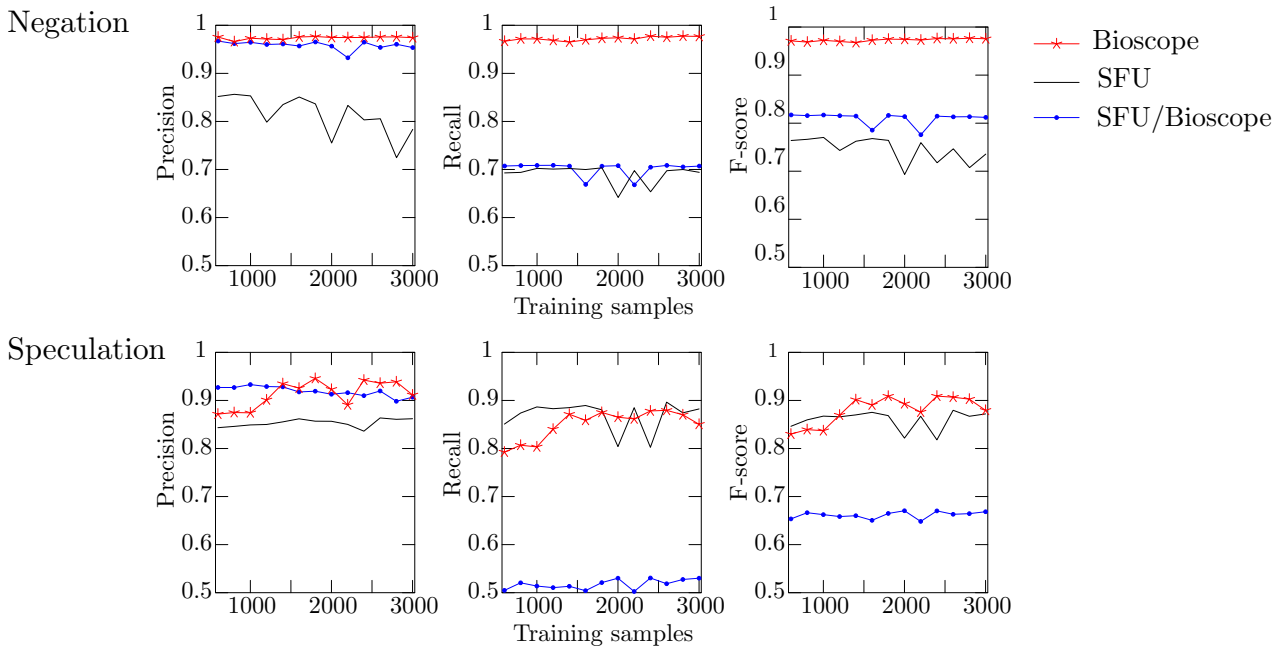

Figure 1: Average results for different number of training samples. SFU/BioScope is the model trained on the SFU Review corpus and applied on the BioScope corpus.

n't or *not* were, however, very common among false negatives and positives, and it is therefore likely that the low results for this category were due to the inability of the trained model to detect multi-token negations, i.e., the same problem that arose for two of the models trained by Cruz et al. (2015). This might, for instance, be an effect of not including the neighbouring words as features. The models were, however, in general able to detect multi-word marker words, e.g., the following complex speculation markers *I-'d-suggest*, *would-think*, *can-either*, *might-expect*, *would-feel*. There were also a number of complex expressions among the false positives for speculation, that might be considered as belonging to this class, despite not being annotated as such. Examples are *can-hope*, *can-either*, *to-think*.

Also the setting of training the model on the SFU Review corpus and evaluating it on the BioScope corpus gave low results for negation as well as for speculation. It can, however, be observed that for speculation markers, this strategy was more successful than the previously explored strategy of training a model on biomedical article texts and applying it on the clinical text genre (Morante and Daelemans, 2009a). There might thus be a larger similarity between how speculation is expressed in consumer reviews and in clinical texts, than between clinical and biomedical texts. Examining incorrectly classified segments showed that false negatives were not limited to marker words that might be more typical to the reasoning style of the clinical genre, e.g., *evaluate*, *suggest*, *indicate*, *compatible*, *consistent* and *question*, but also included general expressions such as *possible* and *probable*.

Results also show that not even lessons learnt for the choice of appropriate machine learning algorithms and features are transferable across genres, as the techniques for detecting negation that was shown successful for the BioScope corpus produced low results on the SFU Review corpus. Future work includes research on whether these findings also hold for the scope of the markers.

6 Conclusion

In the BioScope corpus, speculation and negation markers were detected with results close to previously reported annotator agreement scores. This was also the case for speculation markers in the SFU Review corpus, while detection of negation markers was unsuccessful in this genre. To train the model on consumer reviews and apply it on clinical text also yielded low results, showing that neither the trained models, nor the choice of appropriate algorithms and features, were transferable across the two text genres.

Acknowledgements

This work was funded by the StaViCTA project, framework grant "the Digitized Society – Past, Present, and Future" with No. 2012-5659 from the Swedish Research Council (Vetenskapsrådet).

References

Shashank Agarwal and Hong Yu. 2010. Detecting hedge cues and their scope in biomedical text with conditional random fields. *Journal of Biomedical Informatics*, 43(6):953 – 961.

Magnus Ahltorp, Hideyuki Tanushi, Shiho Kitajima, Maria Skeppstedt, Rafal Rzepka, and Kenji Araki. 2014. HokuMed in NTCIR-11 MedNLP-2:Automatic extraction of medical complaints from Japanese health records using machine learning and rule-based methods. In *Proceedings of NTCIR-11*, pages 158–162.

Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. 2014. Overview of the NTCIR-11 MedNLP-2 Task. In *Proceedings of NTCIR-11*, pages 147–154.

Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Behrouz Bokharaeian, Alberto Diaz, Mariana Neves, and Virginia Francisco. 2014. Exploring negation annotations in the drugddi corpus. In *Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BIOTxtM 2014)*.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*, 34(5):301–310, Oct.

Noa P. Cruz, Maite Taboada, and Ruslan Mitkov. 2015. A machine-learning approach to negation and speculation detection for sentiment analysis. *Journal of the Association for Information Science and Technology*, pages 526–558.

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):1398–1410.

Natalia Konstantinova and Sheila C. M. de Sousa. 2011. Annotating negation and speculation: the case of the review domain. In *Proceedings of the Student Research Workshop associated with The 8th International Conference on Recent Advances in Natural Language Processing, RANLP 2011, 13 September, 2011, Hissar, Bulgaria*, pages 139–144.

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğanur, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, pages 3190–3195, Istanbul, Turkey. European Language Resources Association (ELRA).

Roser Morante and Eduardo Blanco. 2012. *sem 2012 shared task: resolving the scope and focus of negation. In *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, SemEval '12*, pages 265—274.

Roser Morante and Walter Daelemans. 2009a. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2009b. A metalearning approach to processing the scope of negation. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 21–29, Morristown, NJ, USA. Association for Computational Linguistics.

Andreas C. Müller and Sven Behnke. 2014. pystruct - learning structured prediction in python. *Journal of Machine Learning Research*, 15:2055–2060.

Charles. Sutton and Andrew McCallum. 2006. An introduction to conditional random fields for relational learning. In Lise Getoor and Ben Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press.

Maite Taboada and Jack Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 158–161.

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

Özlem. Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–556.

Sumithra Velupillai, Hercules Dalianis, and Maria Kvist. 2011. Factuality Levels of Diagnoses in Swedish Clinical Text. In A. Moen, S. K. Andersen, J. Aarts, and P. Hurlen, editors, *Proc. XXIII International Conference of the European Federation for Medical Informatics (User Centred Networked Health Care)*, pages 559–563, Oslo, August. IOS Press.

Sumithra Velupillai. 2012. *Shades of Certainty – Annotation and Classification of Swedish Medical Records*. Doctoral thesis, Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden, April.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope Corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9 (Suppl 11):S9.