Active learning for detection of stance components

Maria Skeppstedt[1], **Magnus Sahlgren**[2], **Carita Paradis**[3], **Andreas Kerren**[1]
[1]Computer Science Department, Linnaeus University, Växjö, Sweden
{maria.skeppstedt,andreas.kerren}@lnu.se
[2]Swedish Institute of Computer Science, Kista, Sweden
mange@sics.se
[3]Centre for Languages and Literature, Lund University, Lund, Sweden
carita.paradis@englund.lu.se

Abstract

Automatic detection of five language components, which are all relevant for expressing opinions and for stance taking, was studied: *positive sentiment*, *negative sentiment*, *speculation*, *contrast* and *condition*. A resource-aware approach was taken, which included manual annotation of 500 training samples and the use of limited lexical resources. Active learning was compared to random selection of training data, as well as to a lexicon-based method. Active learning was successful for the categories *speculation*, *contrast* and *condition*, but not for the two sentiment categories, for which results achieved when using active learning were similar to those achieved when applying a random selection of training data. This difference is likely due to a larger variation in how sentiment is expressed than in how speakers express the other three categories. This larger variation was also shown by the lower recall results achieved by the lexicon-based approach for sentiment than for the categories *speculation*, *contrast* and *condition*.

1 Introduction

In studies of automatic detection of opinions, it is typically assumed that there are substantial resources available in the form of annotated text corpora (Konstantinova et al., 2012; Socher et al., 2013). However, such large resources of annotated data cannot always be obtained, e.g., when crowd-sourced or community annotations are not possible or not desirable (Fort et al., 2011; Xia and Yetisgen-Yildiz, 2012). The aim of this study is, therefore, to explore the possibility to detect language components that are relevant for opinion mining and stance detection, when using very limited resources of manually annotated data.

Five language components, which are relevant as topic-independent components for expressing opinions and for stance taking, were investigated: *positive* and *negative* sentiment, *speculation*, *contrast* and *condition*. Sentiment analysis is an important component of stance detection, as knowledge of whether positive or negative sentiment is expressed towards a target of interest has been shown useful for the task of binary stance detection, i.e., stance taking *for* or *against* a certain target (Mohammad et al., 2016).

Speculation, *contrast* and *condition* were assessed as important components for stance taking, as they can all be used as modifications of opinions. For instance, an expression of *contrast* could indicate that opinions of different polarities are expressed, e.g., “I did enjoy reading some of this book, but the two tales in the middle dragged too much for me to be able to really recommend this book”. A positive opinion that is expressed with *speculation* might be less positive, e.g., “His description of the 50’s seems accurate and readers might enjoy the trip back in time”. Finally, when a positive opinion is expressed in the context of a *condition*, it is not necessarily positive anymore, e.g., ”If the plot had been more gripping, more intense, this would have worked perfectly”.

2 Previous research

There is a large number of previous sentiment analysis studies, which use different techniques, corpora and task definitions (Täckström and McDonald, 2011; Wang et al., 2012). For instance, an accuracy of 0.85 was achieved when recursive neural networks were used to classify movie review sentences from the

Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media, pages 50–59, Osaka, Japan, December 12 2016.

Stanford Sentiment Treebank into the categories *positive* and *negative* sentiment. When the sentences were classified into a five-level scale of sentiment, with the category *neutral* included (Socher et al., 2013), an accuracy of 0.81 was achieved.

The three opinion modifying categories have all been defined in previous research. *Speculation* has, for instance, been defined as "the possible existence of a thing [that] is claimed – neither its existence nor its non-existence is known for sure" (Vincze, 2010). *Contrast* has been defined as "Contrast(α,β) holds when α and β have similar semantic structures, but contrasting themes, i.e. sentence topics, or when one constituent negates a default consequence of the other" (Reese et al., 2007). Finally, the category *condition* is defined within in Rhetorical Structure Theory as something which "presents a hypothetical, future, or otherwise unrealized situation" (Mann and Taboada, 2016).

There are several studies on speculation/uncertainty detection (Vincze et al., 2008; Farkas et al., 2010; Velupillai, 2012; Wei et al., 2013). On the SFU Review corpus, which consists of English consumer generated reviews of books, movies, music, cars, computers, cookware and hotels (Taboada and Grieve, 2004; Taboada et al., 2006), speculation cues, together with their scopes, have been annotated (Konstantinova et al., 2012). An F-score of 0.92 (Cruz et al., 2015) was achieved when training a support vector machine to automatically detect the annotated cues. The SFU Review corpus has also been annotated for contrasts and conditions (Taboada and Hay, 2008). Experiments have been carried out on the task of determining whether a sentence in this corpus contains an expression of *speculation*, *contrast* or *condition*. A classifier F-score of around 0.90 was achieved for *speculation*, around 0.60 for *contrast* and around 0.70 for *condition*, when using around 3,000 training samples (Skeppstedt et al., 2015).

The standard method to randomly select samples for training the machine learning models were used in all studies described above. However, instead of a random selection, it is possible to use an active selection of useful training data. Although there are some studies on the use of such active learning techniques for sentiment analysis (Li et al., 2012; Kranjc et al., 2015), few studies measure results for resource-aware approaches when using a very limited amount of manually annotated data. The usefulness of active learning for sentence-level detection of language components relevant when expressing stance, and when using a very limited amount of training data, is, therefore, the focus of this study.

3 Method

As the main resource-aware method for detecting the five categories studied, active learning was used. Lexicons of marker words for the categories were also incorporated when training the classifier. A baseline was formed by a simple look-up method that used these lexicons.

3.1 Corpora and lexicons

All classification experiments consisted of the task of sentence classification. That is, each training and testing sample consisted of a sentence and the models were trained to detect whether a sentence contained the category of interest or not. For exploring *negative* and *positive* sentiment, the previously mentioned corpus of 11,855 sentences (the Stanford Sentiment Treebank) that was annotated for sentiment was used (Socher et al., 2013). The annotations were collapsed into the three categories *positive*, *negative*, and *neutral*. These categories were then transformed into two binary text categorisation tasks: a) the detection of sentences that express *positive* sentiment in contrast to *negative* or *neutral*, and b) the detection of sentences that express *negative* sentiment in contrast to *positive* or *neutral*.

Data used for the other three categories consisted of the, above mentioned, corpora created by Konstantinova et al. (2012) and by Taboada and Hay (2008). Both of these two annotation projects were carried out on the 12,663 sentences included in the SFU Review corpus. The *speculation* category annotated by Konstantinova et al. and the *condition* category annotated by Taboada and Hay were used without modifications. The closely related categories *contrast* and *concession*, which were annotated by Taboada and Hay, were, however, merged into the one category that is here referred to as *contrast*. The annotations were transformed into three separate binary classifications tasks, i.e., the task to detect whether a sentence contained *speculation*, *contrast* and/or *condition*, respectively. The same procedure as used in the first of the CoNLL-2010 shared tasks (Farkas et al., 2010) for transforming the data into this format was applied. That is, if either the scope of a *speculation* cue or a segment annotated for *concession/contrast* or *condition* was present in a sentence, the sentence was categorised as belonging to this category (or categories, when several applied).

Limited-sized lexicons of marker words for the five categories were used. For positive and negative sentiment, SentiWordNet (Baccianella et al., 2010) was used to compile the lexicons. The 500 most positive and the 500 most negative words were extracted, and one annotator manually removed words from these lists that would not be considered as typically positive or negative in a movie review setting. Which words to extract as the most positive/negative was determined by ranking the words according to the difference between the positive and negative score of the SentiWordNet synset to which the word belonged. For words that belonged to several synsets, the score resulting in the best ranking on the the positive/negative list was used. The extraction and manual classification resulted in a final list of 373 markers for positive sentiment and 414 markers for negative sentiment.

The lexicons for *speculation* and *contrast* were based on marker words/constructions that have previously been listed by Konstantinova et al. (2012) and Velupillai et al. (2014), and by Reese et al. (2007), respectively. These markers were then used as seed words to expand the lists, by also adding their neighbours in a distributional semantics space to the lists (Sahlgren et al., 2016), as well as their synonyms from a traditional synonym lexicon (Oxford University Press, 2013). In the same fashion as for the sentiment words, the candidates on these expanded lists were then manually classified according to their suitability as marker words. This resulted in a list of 191 markers for *speculation*, and 39 for *contrast*. The *condition* category is a subset of what is defined as *speculation* by Konstantinova et al. (2012). The 26 markers used for this category were, therefore, compiled by manually extracting a subset of the *speculation* markers that were classified as suitable as markers for *condition*.

3.2 Machine learning and active learning methods used

Active learning is built on the idea to reduce the number of training samples required to train a machine learning classifier, by actively selecting useful samples from a pool of unlabelled data. Sample selection could, for instance, be based on the level of uncertainty for a classifier, on the level of disagreement among a number of different classifiers (Olsson, 2008, pp. 25–29), or on the expected model change when adding new data to the pool of labelled data (Tomanek, 2010). The sample selection method used in this study, *simple margin*, is based on expected model change. It is a computationally efficient approach for support vector machines, where the unlabelled sample closest to the separating hyperplane of the classifier is selected (Tong and Koller, 2002).

Support vector machines were used in all experiments, regardless of whether active or random selection of training samples was carried out. The Scikit-learn implementation of the SVC-class with a linear kernel was used (Pedregosa et al., 2011). For all approaches, except approach number four (see section 3.3, below), the machine learning features used were limited to unigrams and bigrams. For approach number four, the output of the lexicon-matching approach was also included as a feature. A minimum of two occurrences in the labelled data was used as a cut-off for including a bigram as a feature, and two occurrences in the entire data pool (labelled and unlabelled) was used as a cut-off for inclusion of unigrams. A corpus created through active selection instead of random selection is not representative of the true data distribution, and standard methods for parameter setting and feature selection do not give reliable results (Schohn and Cohn, 2000). Therefore, the default Scikit-learn SVC parameters were used, and the heuristics of limiting the number of features included to the *n* best was applied. An n equal to the number of samples was used, and, thereby the number of features was allowed to grow with an increasing number of training data samples. Which features were the best was, however, estimated by a χ^2-based feature selection.

3.3 Experiments

A total of five different approaches for detecting the categories investigated were compared, three methods based on active learning, one based on random sampling and one lexicon-matching approach:

(1) The *lexicon-matching* was the most basic approach. Sentences that contained a marker in any of the five compiled lexicons were classified as belonging to the category for which the lexicon was compiled.

(2) The second most basic approach was to use machine learning with random selection of data. (3) As the third approach, active learning based on *simple margin* for selecting a potentially useful training sample was used. An initial machine learning model was first trained on 30 randomly selected samples. Thereafter, the new training samples were chosen based on their distance to the separating hyperplane of the classifier. Two new training samples, i.e., the two samples closest to the separating hyperplane, were selected in each iteration. (4) The same active learning setup as for approach three was applied, but the output of the lexicon matching was used as one of the features for training the classifier. (5) The final approach was also identical to approach number three, but the initial seed set of 30 training samples was not randomly chosen. Instead, a set of 30 samples was selected, with the criterion of requiring each sample to contain a different marker from the lexicon compiled for this category. This follows previous work (Tomanek et al., 2007), in which results have been improved by the extraction of samples that contain known entities for forming the seed set. For the category *condition*, for which there were less than 30 items in the lexicon, the same lexicon item was used for selecting several seed samples.

3.4 Evaluation

A situation was simulated in which limited resources would be available to create an annotated corpus, and thereby a maximum of 500 annotated sentences would be available for training a classifier. Given a hypothetical annotation speed of 50 sentences per hour, it would be possible to construct such an annotated corpus in ten working hours. The five stance categories were evaluated separately, and separate binary classifiers were trained for each of the categories.

The work of the manual annotator was simulated by using the annotations in the corpora described above. Each corpus was split into two equally large sets: an evaluation set and a set to use as the pool of data from which training samples were to be selected. The pool of data from which samples were selected was thus used as simulated unlabelled data, and manual annotation of the selected samples was simulated by using the labelling available in the annotated corpus. The same randomly selected seed set of 30 training samples was used for all machine learning approaches, except for approach number five, for which the lexicons were used for selecting samples.

There is a large difference between the proportion of samples belonging to the minority category for the different categories. That is, a proportion of 24% for *speculation*, 8% for *contrast* and 4% for *condition*, compared to a proportion of 42% and 39% for *positive* and *negative* sentiment, respectively. In order to investigate whether potential differences between categories depend on these proportion differences, rather than on differences between how the categories are expressed, additional experiments were performed for modified versions of the *positive* and *negative* sentiment corpora. The original training data for the sentiment classifiers was modified to instead contain a 24% proportion of the minority category, i.e., the same proportion of minority category samples as the *speculation* category. This was achieved by removing a randomly selected set of instances that belonged to the minority category from the training data. That is, the instances classified as *positive* when investigating *positive* sentiment, and the instances classified as *negative* for *negative* sentiment.

For each of the seven data sets (five with original minority category frequencies and two with modified frequencies), the experiments were repeated 60 times, with a new random split into an evaluation set and into a pool of data from which to select training samples. For each of the 60 folds, a new randomly selected seed set (or a seed set selected based on the lexicon for approach number five) was used. Average precision, recall and F-score between the 60 folds were measured.

4 Results

Results for the five categories of stance are shown in Figures 1-4. The methods evaluated showed one trend for the two sentiment categories, and another trend for the three other categories.

For sentiment, results for active learning were very similar to those achieved when randomly sampling training data. When using 500 training samples, both methods achieved an average F-score of around 0.57 for detecting positive sentiment and an average F-score of around 0.52/0.53 for detecting negative sentiment. For the two versions of the sentiment corpora that had been modified to contain a lower

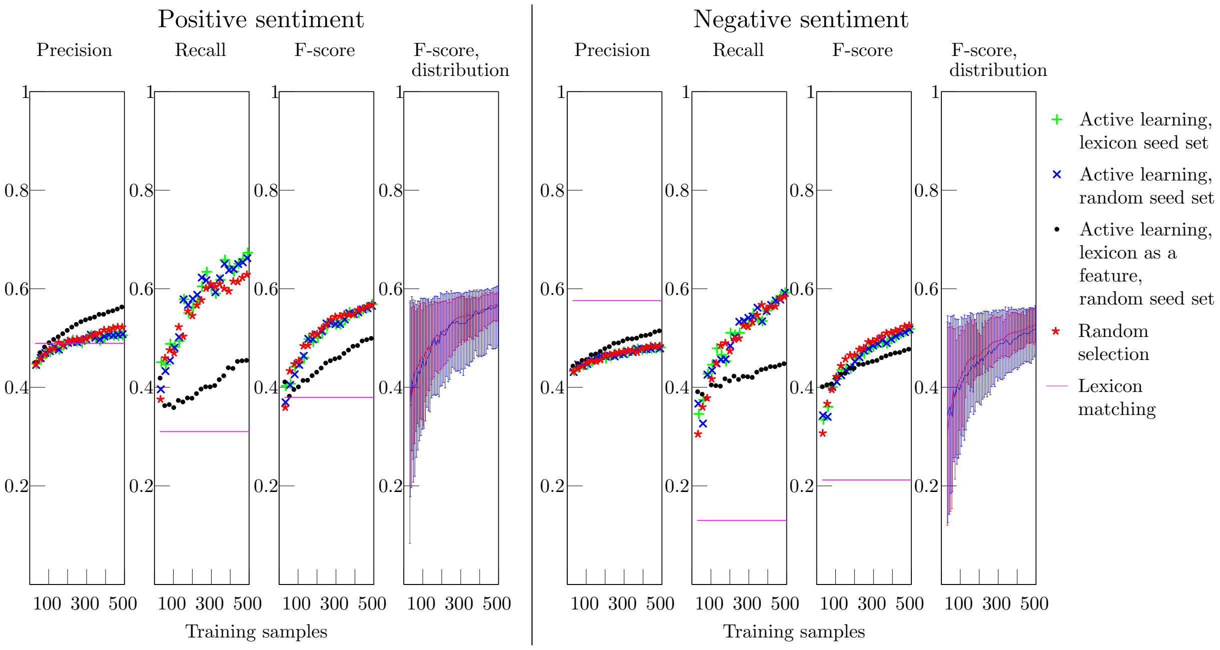

Figure 1: Results for the categories *positive* and *negative* sentiment. Average precision, recall and F-score for the 60 folds are shown. The bars for the distribution of F-score show the 5th to 95th percentile of the results for the 60 folds for *Active learning with a random seed set* and *Random selection*.

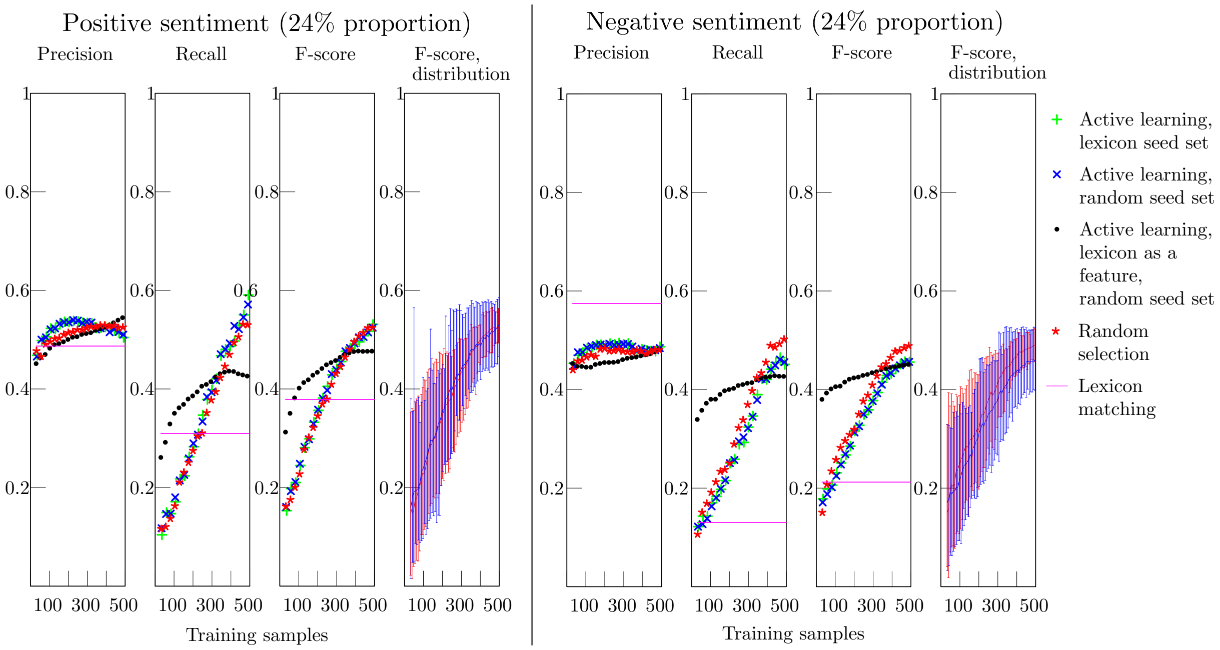

Figure 2: Results when the training data sets have been artificially modified to contain 24% of instances that belong to the minority categories, i.e., to the categories *positive* and *negative* sentiment, respectively. Average precision, recall and F-score for the 60 folds are shown. The bars for the distribution of F-score show the 5th to 95th percentile of the results for the 60 folds for *Active learning with a random seed set* and *Random selection*.

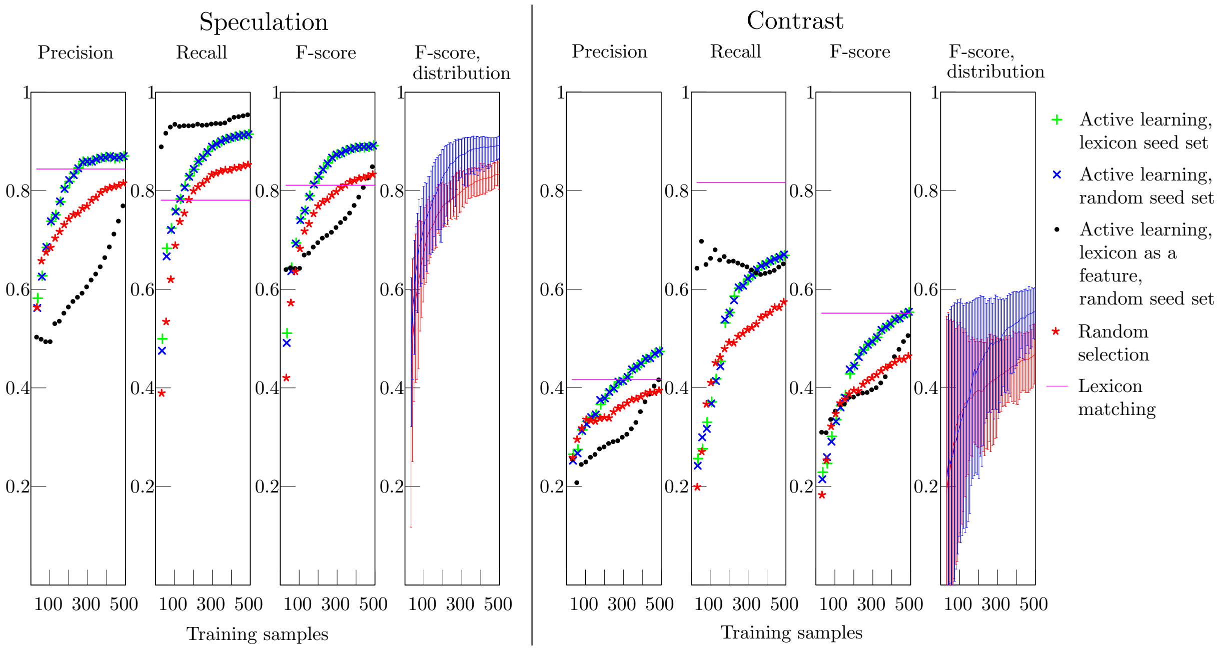

Figure 3: Results for the categories *speculation* and *contrast*. Average precision, recall and F-score for the 60 folds are shown. The bars for the distribution of F-score show the 5th to 95th percentile of the results for the 60 folds for *Active learning with a random seed set* and *Random selection*.

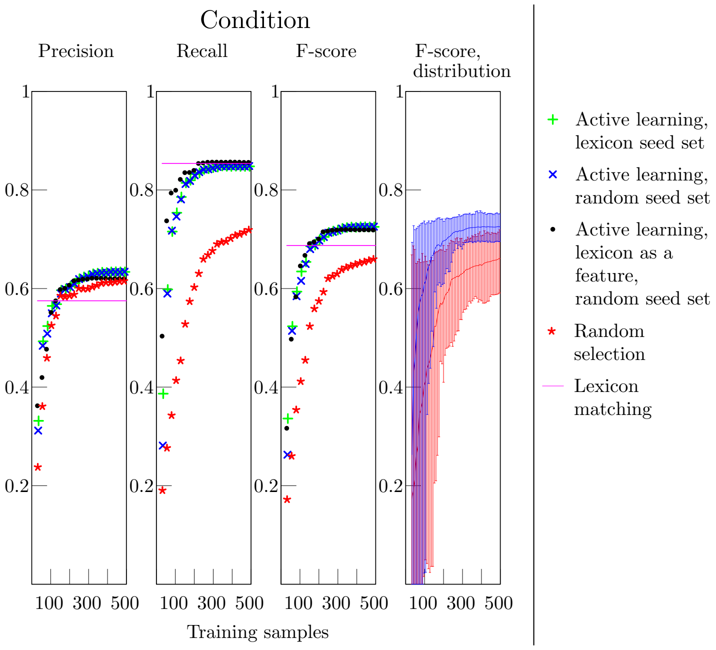

Figure 4: Results for the category *condition*. Average precision, recall and F-score for the 60 folds are shown. The bars for the distribution of F-score show the 5th to 95th percentile of the results for the 60 folds for *Active learning with a random seed set* and *Random selection*.

proportion of the minority categories, the same trend with similar results for active learning and random sampling was shown for *positive* sentiment, while random sample selection was slightly more successful than active learning for *negative* sentiment.

For *speculation*, *contrast* and *condition* on the other hand, active learning clearly outperformed random sampling. For *speculation*, an average F-score of 0.89 was achieved for active learning and an average F-score of 0.83 for random sampling, when using 500 training samples. The performance improvement for active learning started to level out already at 300 training samples for *speculation*, and at that point of measure, the difference between active learning and random selection of training samples was even larger. When using 500 training samples for *contrast*, an average F-score of 0.56 was achieved when using active learning and an average F-score of 0.47 for random sampling. The corresponding results for *condition* were an average F-score of 0.73 for active learning and 0.66 for random selection.

Another difference between these two groups of categories was the results of the lexicon-matching strategy. For *speculation* and *condition*, the lexicon matching performed in line with the classifier trained on randomly sampled data. For *contrast*, lexicon-matching outperformed the random sampling method and achieved results in line with the classifier trained on actively selected data using 500 training samples. However, the machine learning-based classifier showed a better balance between precision and recall.

The use of a seed set containing lexicon terms, instead of a randomly selected seed set, had almost no effect on the results. The use of the lexicon for generating features for training the classifier had a detrimental effect on the average F-score for all categories evaluated with a naturally occurring minority category proportion, except for *condition* for which it had no effect. For the sentiment categories, the use of the lexicon-matching feature led to a lower recall and a very limited increase in precision. The opposite results were observed for *speculation* and *condition*, with a much lower precision and an improvement in recall (for approximately the first 150 samples, a large improvement in recall). In contrast, when the sentiment corpora were modified to only contain 24% samples belonging to the minority category, the inclusion of lexicon-generated features led to a substantial improvement of results for very small data sets. Machine learning was, however, more successful when using 500 training samples.

5 Discussion

For this experiment, in which a very small training data set was used, active learning was successful for *speculation*, *contrast* and *condition*, but not for sentiment classification. This was observed regardless of which of the two minority category proportions for sentiment that was used. These differences are, therefore, likely to be due to a variation in how the different categories are expressed by the speakers.

Positive and *negative* sentiment are likely to be described with a larger set of words and constructions than the other three categories studied. This is indicated by the low recall results of the lexicon-matching approach for sentiment in comparison to the *speculation*/*contrast*/*condition* categories. Lower recall was achieved despite the fact that much larger lexical resources were used for detection of the sentiment categories. This larger variation in how the sentiment categories are expressed might be a reason why active learning did not have a positive effect for sentiment classification, since more variation in how a category is expressed results in a larger, natural, frequency of informative samples. This has the effect that a random selection of training samples for sentiment detection has a large probability of resulting in an informative sample being selected. For the three other categories, on the other hand, there is a lower probability that a randomly selected sample will be informative to the classifier, since there is a lower frequency of samples which contain information that is still unknown to the classifier.

The use of 500 training samples and an active learning approach gave results for *speculation* and *condition* that were in line with those previously achieved when more training data was used (Skeppstedt et al., 2015). Results for *contrast* were, however, slightly lower than those previously achieved. An active learning-based annotation effort of 500 samples, or possibly additional samples for *contrast*, would thus be the approach recommended for these three categories. For *contrast*, the same average F-score was achieved by lexicon-matching as for the classifier trained on 500 actively selected training samples. Although the compilation of a small lexicon of words and constructions signifying *contrast* is less time-consuming than the annotation of 500 sentences, the machine learning approach might be preferable, as it results in a better precision/recall balance.

In contrast, an F-score of 0.5 for detecting positive and negative sentiment is far from the results achieved in previous studies on the same classification task (Socher et al., 2013). A manual annotation of 11,855 sentences (including a more detailed annotation of 200,000 phrases in these sentences) was, however, required to achieve the average accuracy of above 0.8 that has been presented in previous studies. That is, an annotation effort that is far from reasonable in a project with limited resources. Whether to recommend such a limited-resource project to venture the construction of a sentence-level sentiment classifier, depends on the user requirements for this system. To be able to find around 60% of the sentences in which a negative or positive opinion is expressed, and to generate a list of such sentences of which around half are correctly categorised as positive/negative, is likely to be acceptable in some, but not all, circumstances. For sentiment, it is also less clear what training data selection method to recommended, since active and random sampling led to similar results.

It can be concluded that for the categories and text genre evaluated, i.e., the review genre, it is not worth the effort of compiling a limited lexical resource for selecting the seed set or for generating classifier features. However, the lexicons were useful for feature generation in the sentiment corpora with a smaller minority category proportion, when the data set contained up to around 300 training samples. It is, therefore, likely that lexical sentiment resources are more useful in a genre that lacks large resources of annotated data, and in which positive and negative sentiment is less frequently occurring.

5.1 Future work

Although the limited lexical resources compiled for this study did not contribute positively to the results for sentiment detection, it is still likely that an approach fully focusing on detection rules based on extensive and high-quality lexical resources could (i) either be a viable alternative to the machine learning models trained on limited data, which were explored here, or (ii) contribute positively when used as features for training a machine learning model. For instance, by compiling an extended version of SentiWordNet, and leveraging the sentiment scores of positive and negative terms in the resource, Dang et al. (2010) achieved precision and recall scores of around 80% for document level sentiment classification. Future work, therefore, includes the evaluation of such lexicon-based methods on sentence-level sentiment analysis, taking the resource-aware approach used in this study for evaluating its usefulness for projects with limited resources. In particular, there is previous research in which the active learning process has been improved by allowing the annotator to also rank features according to their importance to the category in question (Settles, 2011). Such an approach has the potential of being resource efficient, as it combines the process of compiling a sentiment lexicon with the process of creating labelled data that is useful for training a classifier.

6 Conclusion

Active learning was a successful strategy for three of the categories studied. When using 500 training samples and applying active learning, average F-scores of 0.89, 0.56, and 0.73 were achieved for detecting sentences containing *speculation*, *contrast* and *condition*, while the corresponding figures using random selection of training data were 0.83, 0.47, and 0.66.

For training classifiers to detect the categories *positive* and *negative* sentiment, however, similar results were achieved by active learning and random sampling of training data, an average F-score of 0.57 for detecting positive sentiment and an average F-score of around 0.52/0.53 for detecting negative sentiment. The reason for active learning not being successful for sentiment was not the high proportion of samples that belong to the minority categories in the sentiment corpora. Similar results were achieved when the training data set for sentiment was artificially modified to contain the same proportion of minority category samples as the corpus annotated for *speculation*. Instead, the difference is likely to be due to a larger variation in how *positive* and *negative* sentiment can be expressed, than in how speakers express *speculation*, *contrast* and *condition*. The larger variation in speakers' expressions for *positive* and *negative* sentiment is also indicated by the lower recall achieved by the lexicon-matching approach for sentiment than for the other three categories.

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

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2200–2204, Valletta, Malta, May. European Language Resources Association (ELRA).

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.

Yan Dang, Yulei Zhang, and HsinChun Chen. 2010. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25(4):46–53.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 1–12, Stroudsburg, PA. Association for Computational Linguistics.

Karën Fort, Gilles Adda, Benoît Sagot, and Joseph Mariani. 2011. Crowdsourcing for language resource development: Critical analysis of amazon mechanical turk overpowering use. In *LTC, 5th Language and Technology Conference*.

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 *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, pages 3190–3195, Istanbul, Turkey. European Language Resources Association (ELRA).

Janez Kranjc, Jasmina Smailović, Vid Podpečan, Miha Grčar, Martin Žnidaršič, and Nada Lavrač. 2015. Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the clowdflows platform. *Information Processing & Management*, 51(2):187 – 203.

Shoushan Li, Shengfeng Ju, Guodong Zhou, and Xiaojun Li. 2012. Active learning for imbalanced sentiment classification. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 139–148, Jeju Island, Korea.

William C. Mann and Maite Taboada. 2016. Rhetorical structure theory, relation definitions. http://www.sfu.ca/rst/01intro/definitions.html (Accessed 2016-09-19).

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2016. Stance and sentiment in tweets. *arXiv preprint arXiv:1605.01655*.

Fredrik Olsson. 2008. *Bootstrapping Named Entity Annotation by Means of Active Machine Learning*. Ph.D. thesis, University of Gothenburg. Faculty of Arts.

Oxford University Press. 2013. Oxford thesaurus of English. Digital Version 2.2.1 (156) on Mac OS X.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Brian Reese, Julie Hunter, Nicholas Asher, Pascal Denis, and Jason Baldridge. 2007. Reference manual for the analysis and annotation of rhetorical structure. timeml.org/jamesp/annotation_manual.pdf (accessed May 2015).

Magnus Sahlgren, Amaru Cuba Gyllensten, Fredrik Espinoza, Ola Hamfors, Jussi Karlgren, Fredrik Olsson, Per Persson, Akshay Viswanathan, and Anders Holst. 2016. The Gavagai living lexicon. In *Proceedings of the Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Greg Schohn and David Cohn. 2000. Less is More: Active Learning with Support Vector Machines. In *Proceedings of 17th International Conference on Machine Learning*, pages 839–846, San Francisco, CA, USA. Morgan Kaufmann.

Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1467–1478, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maria Skeppstedt, Teri Schamp-Bjerede, Magnus Sahlgren, Carita Paradis, and Andreas Kerren. 2015. Detecting speculations, contrasts and conditionals in consumer reviews. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 162–168, Stroudsburg, PA, USA, September. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA, October. Association for Computational Linguistics.

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 and Montana Hay. 2008. The SFU review corpus. www.sfu.ca/˜mtaboada/research/SFU_Review_Corpus.html (accessed May 2015).

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).

Oscar Täckström and Ryan McDonald. 2011. Discovering fine-grained sentiment with latent variable structured prediction models. In *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 368–374. Springer Berlin Heidelberg.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. Efficient annotation with the Jena ANnotation Environment (JANE). In *Proceedings of the Linguistic Annotation Workshop*, pages 9–16, Stroudsburg, PA, USA, June. Association for Computational Linguistics.

Katrin Tomanek. 2010. *Resource-Aware Annotation through Active Learning*. Ph.D. thesis, Technical University of Dortmund.

Simon Tong and Daphne Koller. 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, March.

Sumithra Velupillai, Maria Skeppstedt, Maria Kvist, Danielle Mowery, Brian E Chapman, Hercules Dalianis, and Wendy W Chapman. 2014. Cue-based assertion classification for swedish clinical text–developing a lexicon for pyConTextSwe. *Artif Intell Med*, 61(3):137–44, Jul.

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.

Veronika Vincze. 2010. Speculation and negation annotation in natural language texts: what the case of BioScope might (not) reveal. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 28–31, Stroudsburg, PA. Association for Computational Linguistics.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120.

Zhongyu Wei, Junwen Chen, Wei Gao, Binyang Li, Lanjun Zhou, Yulan He, and Kam-Fai Wong. 2013. An empirical study on uncertainty identification in social media context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL*, pages 58–62, Stroudsburg, PA. Association for Computational Linguistics.

Fei Xia and Meliha Yetisgen-Yildiz. 2012. Clinical corpus annotation: Challenges and strategies. In *The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), an LREC Workshop. Turkey.*