Detection of stance and sentiment modifiers in political blogs

Maria Skeppstedt[1], Vasiliki Simaki[1,2], Carita Paradis[2], and Andreas Kerren[1]

[1] Department of Computer Science, Linnaeus University, Växjö, Sweden
{maria.skeppstedt, vasiliki.simaki, andreas.kerren}@lnu.se
[2] Centre for Languages and Literature, Lund University, Lund, Sweden
carita.paradis@englund.lu.se

Abstract. The automatic detection of seven types of modifiers was studied: *Certainty*, *Uncertainty*, *Hypotheticality*, *Prediction*, *Recommendation*, *Concession/Contrast* and *Source*. A classifier aimed at detecting local cue words that signal the categories was the most successful method for five of the categories. For *Prediction* and *Hypotheticality*, however, better results were obtained with a classifier trained on tokens and bigrams present in the entire sentence. Unsupervised cluster features were shown useful for the categories *Source* and *Uncertainty*, when a subset of the training data available was used. However, when all of the 2,095 sentences that had been actively selected and manually annotated were used as training data, the cluster features had a very limited effect. Some of the classification errors made by the models would be possible to avoid by extending the training data set, while other features and feature representations, as well as the incorporation of pragmatic knowledge, would be required for other error types.

Keywords: stance modifiers, sentiment modifiers, active learning, unsupervised features, resource-aware natural language processing

1 Introduction

Stance detection and sentiment analysis are typically modelled as binary classification tasks within the field of natural language processing. That is, authors express stance by positioning themselves as *for* or *against* a given target or topic, or sentiment by giving a *positive* or *negative* opinion [11]. It has, however, been argued that this simple, binary model does not capture the full complexity of the language used for expressing stance and opinions [6]. Instead, authors employ a wide range of modifiers in their opinionated language. The *first aim* of this study is to investigate the automatic detection of seven types of such modifiers.

Many classification tasks within natural language processing rely on a large set of manually annotated training samples. However, in the cases when large training sets are not available, resource-aware methods must instead be relied upon. Active learning is one such resource-aware method that has previously been shown successful for detecting language modifiers. The method is built on

the idea to reduce the number of training samples required by actively selecting useful samples. Using a previously applied active learning approach as a baseline, the *second aim* of this study is to investigate two methods with a potential to improve this approach: (i) to provide annotations with a higher granularity and (ii) to incorporate machine learning features derived in an unsupervised fashion.

2 Background

Based on previous research [6], we study seven types of possible ways in which the categories *positive/negative* sentiment or stance *for/against* can be modified.[3]

Certainty and *Uncertainty* are epistemic modifiers, which in the context of sentiment and stance taking might also give information about the strength with which an opinion is expressed. The following three evaluations would all be classified as positive: "The arguments are irrefutable and readers will definitely enjoy the trip back in time", "The arguments are accurate and readers will enjoy the trip back in time", and "The arguments seem accurate and readers might enjoy the trip back in time". Still, given their different values on the *Certainty/Neutral/Uncertainty* scale, they all convey a slightly different message.

The modifiers *Hypotheticality*, *Recommendation* and *Prediction* indicate that an expression is not necessarily true at the moment at which it is expressed. For instance, the *Hypotheticality* in "If it had been less complicated, it would have been good" makes it into an expression of negative sentiment, despite containing "good". "A good film should never be too complicated" could, on the surface, be a positive or neutral expression. However, since it is a *Recommendation*, it is likely to rather have been primed by a negative opinion. Finally, "Her next film will be the best ever made", expresses positive sentiment without any indications of uncertainty, but since it is a *Prediction*, the author is likely to be less certain of this opinion, than had it been about a film that already had been made.

That an expression contains *Concession/Contrast* typically indicates that opinions of different polarities are expressed, e.g., "I enjoyed reading this book, but parts of it were boring". The occurrence of a contrast affects the overall opinion conveyed, i.e., the overall opinion is not likely to be unequivocally positive or negative. Finally, if an expression contains a statement of its *Source*, this also modifies how an opinion should be interpreted. That is, the existence of a source indicates that the opinion expressed is not necessarily the opinion of the author, e.g., "According to the guide book, this is the best restaurant in town".

We are only aware of one previous study in which resource-aware approaches have been applied for detecting stance and sentiment modifiers [14]. By simulating active learning, that study showed that active sampling outperformed random sampling of training data for categories that closely resemble *Uncertainty*, *Hypotheticality* and *Concession/Contrast*. We therefore here apply the same successful active learning method as our baseline method. In contrast to the previous study, we do not simulate active learning, but use it as the real

[3] Research funded by the StaViCTA project, framework grant "the Digitized Society Past, Present, and Future" with No. 2012-5659 from the Swedish Research Council.

sampling method for creating our training corpus. We also apply the method to a wider range of modifier categories than in the previous study, and we evaluate the effect of extending the method with additional resource-aware techniques.

3 Method

The baseline active learning method was compared to two extensions (i) to add more granular annotations and (ii) to use features from unlabelled data. A previously constructed gold standard corpus, in which the seven categories studied here had been (doubly) annotated [6] was used as evaluation data. This gold standard corpus consisted of opinionated texts in the form of political blogs on the topic of Brexit. The same procedure that had been used for creating this gold standard corpus was applied to create a large pool of data to use in the process of active selection of the training samples that were annotated in the present study. That is, documents from URL:s that started with the word *blog* and that contained expressions related to Brexit were downloaded. Boilerplate text, non-English text and HTML code were then removed using jusText, and the text was segmented into sentences with the standard sentence segmentation technique included in NLTK [2]. It was also ensured that no duplicates from the gold standard were included in the pool of unlabelled data. The annotation of the actively selected sentences was performed with an annotation tool [9] specifically designed for this task. Sentences selected for annotation were presented to the annotator, who classified them according to the seven modifying categories included in the study. Annotation was conducted on the basis of a sentence, with respect to whether the sentence included a modifier (one or more) or not. One sentence at a time, without context, was presented for annotation. The annotator could also mark a sentence as irrelevant if it was a result of a pre-processing error (e.g., boiler plate texts or incomplete sentences). All annotations were performed in an entirely topic-independent fashion. That is, a sentence was, for instance, classified as *Uncertain* if it contained uncertainty in general, regardless of whether this uncertainty was targeted towards a statement related to Brexit.

The active sampling of training data was performed according to the active learning method that had previously been shown successful for modifiers [14]. That is, the training samples that were estimated to be most useful for a support vector machine classifier were actively selected for annotation from the pool of unlabelled data. The estimation was based on the standard method of selecting the unlabelled sample closest to the separating hyperplane of the classifier [16]. The Scikit-learn [12] SVC class with a linear kernel was used for implementing the data selection. A separate binary SVM model was trained for each of the seven categories, using unigrams and bigrams as features and the same classifier settings that had previously been shown successful for detecting modifiers [14]. A previously constructed vocabulary with 20 terms signalling each of the modifying categories studied was used for creating the seed set required to start the active learning process. Three corpus sentences containing each one of these vocabulary terms were selected from the pool of unlabelled data to form a seed set.

The annotator first annotated the sentences in the seed set. Thereafter, active learning was applied to select the five most useful unlabelled sentences for each one of the seven categories. These sentences were then manually annotated, and, thereafter, the models were retrained, also including these newly annotated sentences. This process was then repeated, until 2,095 actively selected sentences had been annotated to form the training set. Results achieved when evaluating the SVM classifier on the gold standard corpus were used as the baseline results.

3.1 Adding annotations with a higher granularity

We hypothesised that the categories studied are mainly expressed by local cue words, and that it would therefore not be optimal to model the task as the text classification task based on sentence-level occurrences of unigrams and bigrams that was used in the training data selection process. Instead, we hypothesised that the task would be more suitable to model as a chunk detection task, with the aim of detecting chunks that function as cue words for the categories studied. A second round of annotation was therefore performed on the training data, in which annotations on a more granular level were provided, on a token-level instead of on a sentence-level. That is, the tokens signalling the modifying categories were marked, using the Brat annotation tool [15].

A classifier was, thereafter, trained to detect tokens/chunks signalling the modifying category in question. For evaluation, the detected chunks were, however, transformed back to a text classification format, in order to match the format of the sentence-level annotations of the gold standard. That is, if the classifier marked a token/text chunk as signalling a modifier category, the sentence containing the token/chunk was classified as belonging to that category. As classifier, Scikit-learn's LogisticRegression classifier was used. The choice was based on an external requirement to provide classifications with easily interpretable confidence estimates, which would not be provided by, e.g., an SVM or a rule/lexicon-based classifier. The token to be classified, as well as the two tokens immediately preceding and the one following it were used as features. To limit the dimensionality of the feature vectors created, a minimum of three occurrences in the training data was required for a neighbouring token to be included as a feature, while a minimum of three occurrences in the entire pool of unlabelled data was required for the current token. A suitable value for the logistic regression regularisation parameter was determined by 30-fold cross-validation on the training data.

Similar to previous studies [10], information derived from a large, unlabelled corpus was also incorporated as features. This was achieved by applying the Gensim library [13], through which semantic vectors in the form of word embeddings from an out-of-the-box word2vec model trained on Google news[4] are provided. Semantic vectors corresponding to the words in the training corpus were clustered using dbscan clustering [5], and each of the n clusters created were given a unique representation in the form of a *cluster representation vector* of length n.

[4] https://code.google.com/archive/p/word2vec/

That is, all vector elements were set to 0, except the one element that represented the cluster, which was set to 1. The cluster information was incorporated in the feature representation for a token by (i) determining which cluster was closest to the semantic vector that corresponds to the token (measured through the Euclidean distance between the semantic vector and the cluster centroids), and (ii) concatenating the *cluster representation vector* of this cluster to the feature representation for the token. A maximum Euclidean distance of 0.8 between two semantic vectors in the same cluster was allowed when performing the dbscan clustering. This distance was determined by manually inspecting the semantic coherence in a subset of created clusters, for different distances.

4 Results and discussion

Models using the three methods investigated were trained on two versions of the training corpus, after 1,525 actively selected sentences had been annotated and after 2,095 actively selected sentences had been annotated. F-scores obtained when evaluating these models against the gold standard are presented in Table 1, and precision and recall are presented in Figure 1.

The hypothesis that a chunk detection model would be most suitable seems to hold for five of the categories. For *Hypotheticality* and *Prediction*, however, the sentence-level classifiers performed better. The general trend (with the exception of *Certainty*) was that the baseline method resulted in a better recall, while better precision was shown by the other two methods investigated. When only 1,525 training data samples were used, the cluster features had (i) a relatively large positive effect on the *Source* category and (ii) a small positive effect on the *Uncertainty* category. The effects of incorporating cluster features were, however, very limited when all data available was used.

Regardless of which method was used, only the results for the best-performing classifier, *Concession/Contrast*, were close to the annotator agreement. Results for the best performing chunk-level models were, therefore, analysed to identify frequent reasons for false negatives, i.e., when the classifier failed to detect the category in question, and false positives, i.e., sentences incorrectly classified as belonging to the category in question. Table 2 lists typical challenges to the classifiers (referred to as Ex. 1.1–16.5 in the following paragraphs).

In sentences annotated according to the category *Uncertainty*, there were eight frequently used words, including "think", "should" and "would" that caused classification problems (Ex. 1.1–8.2). These words occurred in 76%/75%/59% of the true positives/false negatives/false positives, respectively. Whether these words function as cues for uncertainty can, sometimes, be determined by the words in their context. As many of the examples illustrate, however, pragmatic knowledge is often required to determine what they indicate, i.e., knowledge that is not possible to capture without using vast resources of annotated data. Some of these words were also frequently used in sentences classified into *Hypotheticality* ("could", "would", "should" or "might" occurred in 89% of the false negatives and 66% of the false positives for *Hypotheticality*). This might have

Fig. 1. Precision and recall when using a training set of 2,095 sentences. The error bars show the 95% confidence interval for the results [3, pp. 91–92, 94–96].

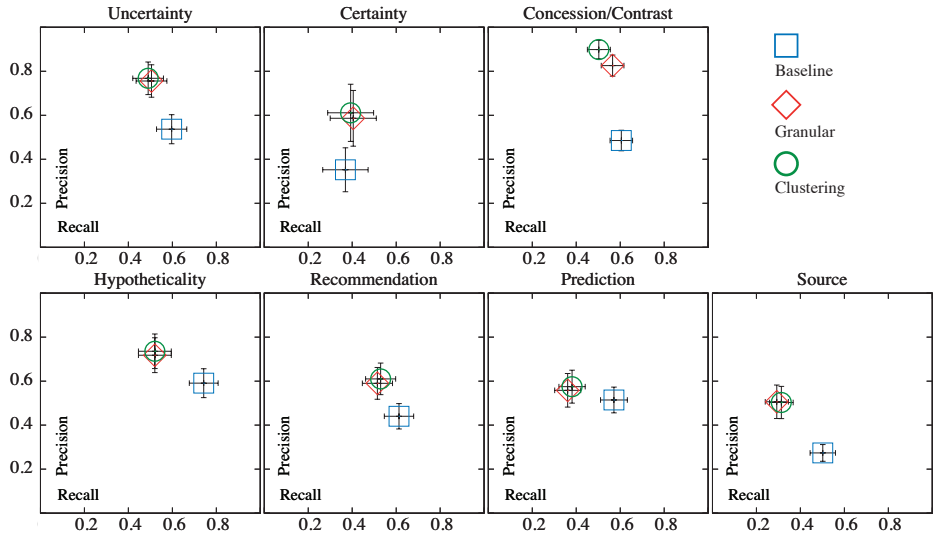

Table 1. F-scores for the three methods investigated, when using 1,525 sentences and 2,095 sentences to train the classifiers, respectively. Best results for each data size are shown in bold. F-scores for the intra-annotator agreement are provided with a relatively low confidence (as they were calculated on half of the gold standard corpus) and no point estimates are, therefore, given. Instead, confidence intervals (95%) were computed with a bootstrap resampling approach [7], using the 2.5/97.5 percentiles of 10,000 bootstrapping folds. *Category frequency* is the percentage of sentences that contain the category in question (in the gold standard/after 1,525 training sentences had been annotated/after 2,095 training sentences had been annotated).

	F-score 1,525 training instances			F-score 2,095 training instances			F-score intra-annotator	Category frequency
	Baseline	Granular	Cluster	Baseline	Granular	Cluster	(Min – Max)	(%)
Uncertainty	0.53	0.59	**0.63**	0.56	**0.61**	0.60	0.74 – 0.87	10/14/14
Certainty	0.29	0.50	**0.52**	0.36	**0.48**	**0.48**	0.55 – 0.78	4/7/6
Conc./Contrast	0.52	**0.66**	0.65	0.54	**0.67**	0.65	0.71 – 0.81	17/21/20
Hypotheticality	**0.65**	0.60	0.59	**0.66**	0.60	0.61	0.72 – 0.86	8/13/13
Recommend.	0.48	**0.60**	0.57	0.51	0.55	**0.57**	0.72 – 0.85	10/14/13
Prediction	**0.50**	0.44	0.48	**0.54**	0.44	0.46	0.73 – 0.84	12/14/15
Source	0.35	0.34	**0.42**	0.35	0.37	**0.39**	0.66 – 0.79	14/14/17

	F-score, previous studies	
Speculation	0.92	(10-fold cross validation on 17,263 sentences [4])
Speculation	0.89	(500 actively selected sentences [14])
Conc./Contrast	0.56	(500 actively selected sentences [14])
Hypotheticality	0.73	(500 actively selected sentences [14])

(Speculation corresponds to *Uncertainty* ∪ *Hypotheticality*)

Table 2. Examples of sentences that might be challenging to a classifier. **B** (for **B**rexit) means annotated in this study and *K* means annotated by *K*onstantinova et al. [8].

	Sentences containing difficult expressions (and their classifications)
B 1.1	"I **think** you are from a well to do family ..." (*Uncertainty*)
B 1.2	"I **think** it's a slap in the face to anyone who has experienced ..." (opinion)
K 1.3	"I don't **think** it's too bad for a cordless phone."(Speculation in [8], but opinion here)
K 1.4	"I **think** it makes the phone look less modern."(Speculation in [8], but opinion here)
B 2.1	"... events **should** make it difficult for Camerlot ..." (*Uncertainty/Prediction*)
B 2.2	"**Should** they not win a constituency vote I don't want to risk ... " (*Hypotheticality*)
B 2.3	"People **should** vote on the basis of a citizen's duty ... " (*Recommendation*)
K 2.4	"P. **should** really consider adding this ... " (Speculation in [8], *Recommendation* here)
B 3.1	"This **could** provide Washington with more flexibility ... " (*Uncertainty*)
B 3.2	"If an officer was on parade such language **could** not be used. (*Hypotheticality*)
B 3.3	"... because it was during a campaign they **could** not ignore it. " (past tense of can)
K 3.4	"It went by so fast you **could** barely tell what he was saying." (Speculation in [8])
B 4.1	"... even these pro - EU industries **might** see benefits from exit ... " (*Uncertainty*)
B 4.2	"Brexit **might** be another turning point ..." (*Prediction/Uncertainty*)
B 4.3	"... If the dates are extended sufficiently then it **might** be worth it." (*Hypotheticality*)
B 5.1	"... it looked like Iain Gray **would** be FM." (*Uncertainty*)
B 5.2	"... more integrated capital markets **would** tie things together better." (*Hypotheticality*)
B 6.1	"Be that as it **may**, we finally did join the European Union in January 1973."
B 6.2	"Granted, his party **may** commit regicide in the process." (*Uncertainty*)
B 7.1	" ... being elected **seems** to be about reconciling the unreconcilable. (*Source*)
B 7.2	" ... international cooperation **seems** to have lost its way ... " (*Uncertainty*)
B 8.1	"In this fall he **appeared** to hurt his leg ... (*Source*)
B 8.2	"Capitalism doesn't **appear** to work without someone losing out." (*Uncertainty*)
B 9.1	"It is **vanishingly unlikely** ... "/"It is **inconceivable** that ..." (*Certainty*)
B 10.1	"I asked a man **if** he knew the way." (indirect question)
B 10.2	"There will be no going back **if** we decide to leave" (*Hypotheticality*)
B 10.3	"There'd be no residue of benefit – **even if** that were possible." (*Hypotheticality*)
B 10.4	"I listen carefully to what is being said **even if** I don't agree." (*Conc./Contrast*)
B 11.1	"The referendum has triggered the eurogroups **need** for additional safeguards"
B 11.2	"What we actually **need** is a manifesto that provides detail" (*Recommendation*)
B 12.1	"I can't **see** him winning by putting his foot in his mouth." (*Prediction*)
B 12.2	"I could **see** that many of the trees in his orchard bore the scars ..." (*Source*)
B 13.1	"**But** it is not the funding side of Greek banks that is the real problem." (no contrast)
	Sentences containing antithesis without a contrast marker
B 14.1	"Public schools in Barcelona teach in Catalan, not Spanish." (*Conc./Contrast*)
B 14.2	"Germany has not searched our mails as have the British." (*Conc./Contrast*)
B 14.3	"... any other author, alive or deceased" (*Conc./Contrast*)
B 14.4	"Having heavily lost the referendum, their vote soared to over 49% ..." (*Conc./Contrast*)
	Sentences categorised as *Source* (the source in italics and the marker in bold)
B 15.1	"*Nigel Farage*, the 'Saviour of British Sovereignty', whilst knowing this, **insists** that .."
B 15.2	"*Statistics* also **show**, that despite or because of the NHS, no one gets out of here alive!"
B 15.3	"Most *opinion polls* have Ukip, which has 11 MEPs .."
	Sentences using a number of isolated markers to express *Prediction* (markers in bold)
B 16.1	"... the **possibility** that the Government's policies **could harm growth** ..."
B 16.2	"The situation is **moving** too fast now for a controllable **outcome**."
B 16.3	"Greece **may** experience rapidly **accelerating** inflation."
B 16.4	"I **think** the early **2020s** are the **best bet**."
B 16.5	"I don't **think** this **can last**."

caused additional difficulties for the classifiers, and might be the reason why the baseline classifier, which used cues from the entire sentence, outperformed the chunk-based one for the category *Hypotheticality*. Many previous studies (for instance [4, 8, 14], but not [17]) have (i) grouped *Hypotheticality* and *Uncertainty* into one category, and (ii) treated, e.g., "think"/"should"/"could" as markers for *Uncertainty* regardless of the pragmatics (Ex. 1.3/1.4/2.4/3.4), which might explain why lower results were achieved here than in previous studies (Table 1).

For the category *Certainty*, 24% of the false negatives included either one of the words "clear" or "sure", and there were also other expressions, for which it is dependent on the context whether they signal *Certainty*, e.g., "of course". The classifier, however, also failed to detect many evident cues for *Certainty*, e.g., "definitely" and some (but not all) modified uncertainty cues, e.g., "without doubt" and "too plausible". There were also confusions between *Certainty* and *Uncertainty*, for cases that might be equally challenging for a human (Ex. 9.1).

For 13% of the false negatives for *Concession/Contrast*, an expression of contrast started the sentence. This might be explained by that contrast markers often start a sentence without signalling contrast (Ex. 13.1). The expressions "even if", "yet", "and then" caused another 10% of the false negatives (Ex. 10.3, 10.4), while 20% contained more univocal contrast markers, e.g., "compared with". Most false negatives did, however, not contain an explicit contrast marker, but expressed *antithesis* [1], e.g., by applying negations or antonyms (Ex. 14.1–14.4). These constructions are impossible to detect by the models applied here, but more complex approaches would be required, e.g., approaches built on external semantic resources that model semantic relations between words.

For *Source*, 17% of the false negatives and 26% of the false positives contained versions of the ambiguous expressions "appear", "seem" or "see" (Ex. 7.1-8.2). 75% of the remaining false negatives contained a clear cue that indicated the existence of a source, e.g., "show" or "insists" (Ex. 15.1,15.2). In only a few cases was the source mentioned without a marker (Ex. 15.3). There was, however, a large variation in what cues were used, which might explain the low results. Sometimes there was also a distance of many tokens between the cue and the actual source (Ex. 15.1), which indicates that information from a parser might be useful for constructing features for this category. As the source of information often consists of names, the output from a named entity recogniser might also be useful. Among false positives, there were many examples where the model had learnt to detect typical cues for a source, but had not learnt in which contexts it functions as a cue for *Source*. For instance, "report" in "The IMF is leaking a report ...". This indicates that more training data, which would allow more examples of context, would be required in order to improve the *Source*-classifier.

Modifiers expressed by someone else than the author were not counted as belonging to that modifying category. E.g., "The US and some of its partners suspect Iran of ..." should not be classified as *Uncertainty*, since the uncertainty is expressed by someone else than the author. This was a general source of false positives, which shows the need for a high-performing *Source*-classifier for improving the other classifiers.

The model learnt for *Recommendation* did not reflect the complexity with which the category was expressed. 79% of the true positives and 61% of the false positives contained versions of "should"/"must"/"need"/"have to" (Ex. 2.3, 11.1–11.2). These expressions were, however, only present in 30% of the false negatives. Among the rest of the false negatives, around half contained specific expressions that mark recommendation, e.g., "let's", "I urge" and "I suggest", while the rest were recommendations expressed by an imperative verb, e.g., "Stop using it" and "Count me out". Using the same features and a larger training data set might lead to that more of the recommendation-specific expressions will be detected. Features that include part-of-speech tagging might, however, be required for detecting recommendations expressed by an imperative verb.

The model learnt for *Prediction* seems to be even less complex, with 86% of the true positives and 56% of false positives that contained versions of "will"/ "going to". There is, however, a potential for a large complexity of a good model, since the same frequency among false negatives was 33%, and there was a large variety in how predictions were expressed. Contrary to the other categories that were typically expressed by isolated chunks in the sentence, *Prediction* was often expressed using several different cues that would all be required for the reader (and thereby the model) to understand that the sentence contained a prediction. This difference (Ex. 16.1-16.5) is likely to be the reason why the sentence-level baseline classifier outperformed the chunk-based one for *Prediction*.

Future work includes an incorporation of some of the types of features suggested here, as well as a further expansion of the training data. It would also be possible to combine the two approaches evaluated here, by applying the output of a chunk-based classifier as features for training the sentence-level classifier.

5 Conclusion

We hypothesised that stance and sentiment modifiers are mainly expressed by local cue words, and that detection of such modifiers therefore is most suitable to model as a chunk detection task, with the aim of detecting these cue words. This hypothesis held true for five of the categories studied, but for *Prediction* and *Hypotheticality*, better results were obtained with a sentence-level classifier trained on tokens and bigrams present in the entire sentence. Cluster features derived in an unsupervised fashion were useful for the categories *Source* and *Uncertainty* when a subset of the training data available was used. When all data available (2,095 actively selected sentences) was used, however, the effects of incorporating cluster features were very limited. The analysis showed that some types of classifier errors might be avoided by providing more training data, and thereby more examples of cue words and contexts that could determine whether potential cue words signal the categories investigated. For other types of errors, however, other features and feature representations than the ones used here might be required. Yet other types of errors would only be possible to avoid by taking on the difficult task of incorporating pragmatic knowledge.

References

1. Azar, M.: Argumentative text as rhetorical structure: An Application of Rhetorical Structure Theory. Argumentation 13(1), 97–114 (1999)
2. Bird, S.: NLTK: The natural language toolkit. In: Proceedings of the Workshop on Effective Tools and Methodologies for Teaching NLP and Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
3. Campbell, M.J., Machin, D., Walters, S.J.: Medical statistics : A textbook for the health sciences. Wiley, Chichester, 4. ed. edn. (2007)
4. Cruz, N.P., Taboada, M., Mitkov, R.: A machine-learning approach to negation and speculation detection for sentiment analysis. Journal of the Association for Information Science and Technology pp. 526–558 (2015)
5. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining. pp. 226–231. AAAI Press (1996)
6. Forthcoming: Annotating speaker stance in discourse: the Brexit Blog Corpus (2017)
7. Kaplan, D.: Resampling stats in MATLAB. http://www.macalester.edu/ ˜kaplan/Resampling/ (accessed August 2015, 1999)
8. Konstantinova, N., de Sousa, S.C., Cruz, N.P., Maña, M.J., Taboada, M., Mitkov, R.: A review corpus annotated for negation, speculation and their scope. In: Proceedings of the Conference on Language Resources and Evaluation. pp. 3190–3195. European Language Resources Association, Paris, France (2012)
9. Kucher, K., Kerren, A., Paradis, C., Sahlgren, M.: Visual Analysis of Text Annotations for Stance Classification with ALVA. In: EuroVis 2016 - Posters. pp. 49–51. The Eurographics Association, Geneva, Switzerland (2016)
10. Miller, S., Guinness, J., Zamanian, A.: Name tagging with word clusters and discriminative training. In: Proceedings of NAACL HLT. pp. 337–342. Association for Computational Linguistics, Stroudsburg, PA, USA (2004)
11. Mohammad, S.M., Sobhani, P., Kiritchenko, S.: Stance and sentiment in tweets. arXiv preprint arXiv:1605.01655 (2016)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
13. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of Workshop on New Challenges for NLP Frameworks. pp. 45–50. European Language Resources Association, Paris, France (May 2010)
14. Skeppstedt, M., Sahlgren, M., Paradis, C., Kerren, A.: Active learning for detection of stance components. In: Proceedings of the PEOPLES workshop. pp. 50–59. Association for Computational Linguistics, Stroudsburg, PA, USA (December 2016)
15. Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., Tsujii, J.: BRAT: a web-based tool for NLP-assisted text annotation. In: Proceedings of EACL. pp. 102–107. Association for Computational Linguistics, Stroudsburg, PA, USA (2012)
16. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. 2, 45–66 (Mar 2002)
17. Velupillai, S.: Shades of Certainty – Annotation and Classification of Swedish Medical Records. Doctoral thesis, Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden (April 2012)