Unshared task: (Dis)agreement in online debates

Maria Skeppstedt[1,2] **Magnus Sahlgren**[1] **Carita Paradis**[3] **Andreas Kerren**[2]
[1]Gavagai AB, Stockholm, Sweden
{maria, mange}@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

Topic-independent expressions for conveying agreement and disagreement were annotated in a corpus of web forum debates, in order to evaluate a classifier trained to detect these two categories. Among the 175 expressions annotated in the evaluation set, 163 were unique, which shows that there is large variation in expressions used. This variation might be one of the reasons why the task of automatically detecting the categories was difficult. F-scores of 0.44 and 0.37 were achieved by a classifier trained on 2,000 debate sentences for detecting sentence-level agreement and disagreement.

1 Introduction

Argumentation mining involves the task of automatically extracting an author's argumentation for taking a specific stance. This includes, e.g., to extract premises and conclusion, or the relationship between arguments, such as argument and counter-argument (Green et al., 2014; Habernal and Gurevych, 2015). In a corpus containing dialog, e.g., different types of web fora or discussion pages, the argumentation often involves a reaction to arguments given by previous authors in the discussion thread. The author might, for instance, give a counter-argument to an argument appearing earlier in the thread, or an argument supporting the stance of a previous author. A sub-task of detecting the argument structure of a dialogic corpus is, therefore, to detect when the author conveys agreement or disagreement with other authors.

The aim of this study was to investigate this sub-task, i.e., to automatically detect posts in a dialogic corpus that contain agreement or disagreement.

2 Previous research

Dis/agreement has been the focus of conversational analysis (Mori, 1999), and is linked to Speech Act Theory (Searle, 1976). The categories have been annotated and detected in transcribed speech, e.g., in meeting discussions (Hillard et al., 2003; Galley et al., 2004; Hahn et al., 2006), congressional floor-debates (Thomas et al., 2006), and broadcast conversations (Germesin and Wilson, 2009).

Online discussions in form of Wikipedia Talk have been annotated for *dis/agreement* (Andreas et al., 2012), for *positive/negative* attitude towards other contributors (Ferschke, 2014), and for subclasses of *positive/negative* alignment, e.g. explicit agreement/disagreement, praise/thanking, and critic/insult (Bender et al., 2011).

For online debate fora, there is a corpus of posts with a scalar judgment for their level of *dis/agreement* with a previous post (Walker et al., 2012). Misra et al. (2013) used frequently occurring uni/bi/trigrams from the non-neutral posts in this corpus for creating a lexicon of topic-independent expressions for *dis/agreement*. This lexicon was then used for selecting features for training a topic-independent classifier. The approach resulted in an accuracy of 0.66 (an improvement of 0.6 points compared to standard feature selection) for distinguishing the classes *agreement/disagreement*, when evaluating the classifier on debate topics not included in the training data.

Despite this usefulness of the lexicon for creating a topic-independent dis/agreement classifier, there are, to the best of our knowledge, no debate forum corpora annotated with the focus of topic-independent expressions of dis/agreement. Here, the first step towards creating such a resource was, therefore, taken.

Proceedings of the 3rd Workshop on Argument Mining, pages 154–159,
Berlin, Germany, August 7-12, 2016.

3 Method

The study was conducted on discussions from a debate forum. The data originates from createdebate.com, which is a debate forum that hosts debates on a variety of topics. The data used as evaluation set was provided for task *Variant A* in the 3rd Workshop on Argument Mining, and consists of 27 manually collected discussion threads.[1] The debates start with a question, e.g., "Should the age for drinking be lowered?", which users then debate, either by posting an independent post, or by supporting/disputing/clarifying a previous post.

The same division into topic-specific/topic-independent means for conveying dis/agreement as previously used by Misra et al. (2013) was adopted. Instead of using it for creating a lexical resource, it was, however, used as a guideline for annotation. A preliminary analysis of posts tagged as *support/dispute* in 8 discussion threads showed that typical topic-specific strategies for conveying dis/agreement were reformulations/expansions/elaborations of what was stated in a previous post. A new argument for or against the initial debate question could, however, also be given, without references to the content of the previous post. Topic-independent means for conveying dis/agreement were typically either explicit statements such as "I (dis)agree", "NO way!", or critical follow-up questions, "A: Alcohol should be forbidden. B: *Should it then* also be illegal with cell phones?". All means of conveying dis/agreement independent of debate topic were, however, included in the task, e.g., as exemplified by Bender et al. (2011), topic-independent explicit dis/agreement, (sarcastic) praise/thanking, positive reference, doubt, criticism/insult, dismissing.

The preliminary analysis also showed that the *support/dispute* tagging provided in the unshared task data would not suffice for distinguishing agreement from disagreement, as there were posts tagged as *support* that consisted mainly of expressions of disagreement.

3.1 Annotation of task data (evaluation set)

All instances in which agreement or disagreement were conveyed using topic-independent expressions were annotated in the unshared task data set. The annotation was performed by marking a relevant scope of text, in the form of the longest possible chunk that was still a topic-independent expression conveying dis/agreement. For instance, in Figure 1, "fighting a war" is specific to the topic of the debate, whereas the annotated chunk, "is a good thing?", is topic-independent and could be used for expressing disagreement in other cases.

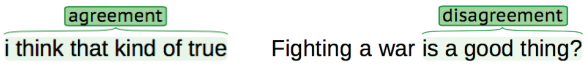

Figure 1: Two of the chunks in the unshared task data that were annotated as dis/agreement.

The annotation was performed by one annotator, with Brat as the annotation tool (Stenetorp et al., 2012).

3.2 Annotation/classification of training set

Identifying and annotating relevant chunks in running text is a time-consuming task, which also requires a large amount of attention from the annotator. Classifications of individual sentences is, however, an easier task, and to classify a limited corpus of 2,000 sentences is feasible in a relatively short amount of time. For creating a larger (but still relatively limited) training set of discussion sentences conveying dis/agreement, the chunk annotation task was reformulated as a text classification task, and individual sentences were manually classified according to the categories *agreement*, *disagreement* or *neutral*. As for the previous annotation set-up, sentences containing topic-independent expressions for conveying the two categories of interest were classified as containing *agreement* or *disagreement*.

The 2,300 most popular threads, i.e., those containing the largest number of posts, were downloaded from the createdebate.com website (excluding threads present in the evaluation data). The posts are provided with author tagging that states what posts are *disputing* or *clarifying* previous posts. Among posts for which no such tag was attached (the *other* posts), and among posts tagged as *disputing* a previous post, 2,000 first-sentences were randomly selected for manual classification. Only first-sentences of posts were included to make it possible to classify each individual sentence without context, since it is likely that their agreement/disagreement classification is less dependent on the context of the post. For sentence segmentation, the standard functionality in NLTK (Bird, 2002) was used.

[1] https://github.com/UKPLab/argmin2016-unshared-task.

3.3 Training a classifier

As the final step, linear support vector machines were trained to perform the binary text classification tasks of detecting sentences containing *agreement* and *disagreement*. The LinearSVC class included in Scikit learn (Pedregosa et al., 2011) was trained with uni/bigrams/trig-rams as features, with the requirement of a uni/bigram/trigram to having occurred at least twice in the training data to be included. The *n* best features were selected by the built-in χ^2-based feature selection, and suitable values of *n* and the support vector machine penalty parameter *C* were determined by 10-fold cross-validation. The text was not transformed into lower-case, as the use of case is one possible way of expressing or emphasising dis/agreement, e.g., 'NO way!'. The settings that achieved the best results were used for training a model on the entire training data set, which was then evaluated on the data provided for the unshared task. The annotations in the unshared task data were transformed into an evaluation set by transforming the text chunk annotations into sentence-level classifications of whether a sentence contained *agreement* or *disagreement*.

Two versions of the classifiers were trained, one in which neutral sentences were included and one with the same set-up as used by Misra et al. (2013), i.e., to train a classifier to distinguish *agreement* from *disagreement* and thereby not including neutral sentences.

4 Results and discussion

# of chunks annotated in total: 175 (163 unique)	
# *agreement:* 43	# *disagreement:* 132

Table 1: Statistics of unshared task annotated data.

Statistics of the annotated data (Tables 1, 2) shows that expressions for disagreement are more frequently occurring than expressions for agreement. This is most likely explained by the typical style used in debate fora, in which debating often is conducted by disputing other debaters, but it could also be due to a more frequent use of topic-independent expressions for this category.

A large variation in the expressions used was observed during annotation. This observation is supported by the data, as 163 unique expressions were annotated. This shows that the approach used by Misra et al. (2013), i.e., to classify frequently occurring n-grams, is not sufficient for creating a high-coverage lexicon of expressions, and it also indicates that automatic detection of these expressions might be a difficult task.

	Disputed	Other	Total
# *agreement*	36	73	109
# *disagreement*	420	92	512
# *sentences in total*	1,000	1,000	2,000

Table 2: The training data statistics shows the number of sentences annotated as *agreement* and *disagreement*, extracted from posts tagged as *disputing* a previous post or as *other*. *# sentences in total* is the total number of annotated sentences. The corpus also included 57 sentences, for which it could not be determined without context whether disagreement or agreement was expressed. These were classified as *neutral*. The 25 sentences that contained both agreement and disagreement were classified as belonging to the *agreement* category.

The most important features used by the classifiers (Figure 2) are topic-independent, which indicates that the aim to create topic-independent classifiers was reached. Among less important features, there were, however, also topic-specific expressions, which shows that the trained classifiers were not entirely topic-independent.

The classifier results are shown in Table 3. For the training set, an F-score of around 0.47 was obtained for *agreement* and around 0.55 for *disagreement*. Results were, however, substantially lower for *disagreement* on the evaluation set. This decrease in results could be explained by overfitting to the training data, and by uncertainty of the results due to the small evaluation set. There might, however, also be a difference between what is considered as an expression of disagreement when it occurs in the first sentence of a post (which was the case for the training data) and when it occurs somewhere else in the text (which was the case for many sentences in the evaluation data).

To distinguish agreement from disagreement was an easier task, resulting in F-scores of 0.60 for *agreement* and 0.92 for *disagreement* on the training set and F-scores of 0.55 and 0.81, respectively on the evaluation set. The recall for *agreement* was, however, low also for this task, proba-

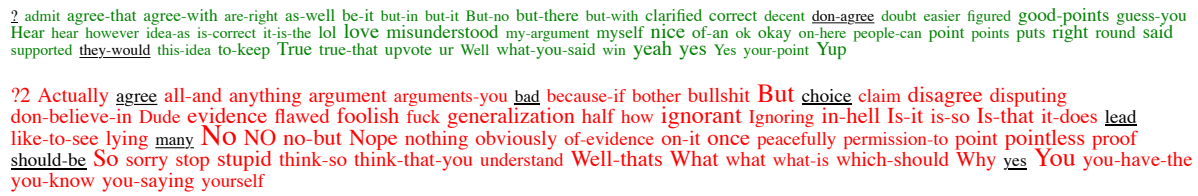

Figure 2: The most important features for detecting *agreement* (green) and *disagreement* (red). Font size corresponds to the importance of the feature, and negative features (in black) are underlined.

	Including neutral sentences			Agreement vs. disagreement (no neutral sent.)	
		Precision	Recall	Precision	Recall
Training-set (10-fold)	*agreement*	0.46	0.47	0.64	0.56
	disagreement	0.54	0.56	0.91	0.93
Evaluation-set	*agreement*	0.45±0.15	0.44±0.15	0.70±0.17	0.46±0.15
	disagreement	0.29±0.06	0.50±0.09	0.84±0.06	0.93±0.04

Table 3: Machine learning results obtained on the corpus annotated in this study.

bly due to the few occurrence of this class in the training data.

Previous machine learning approaches were generally more successful. In Wikipedia Talk, F-scores of 0.69 and 0.53 were achieved for detecting *positive* and *negative* attitudes (Ferschke, 2014), and F-scores of 0.61 and 0.84 for detecting *explicit agreement/disagreement* (Opitz and Zirn, 2013). In other types of online debates, F-scores of 0.65 and 0.77 have been achieved for detecting dis/agreement (Yin et al., 2012), and an F-score of 0.75 for detecting disagreement (Allen et al., 2014). Including a neutral category, however, has resulted in agreement/disagreement F-scores of 0.23/0.46 for Wikipedia Talk and 0.26/0.57 for debate forums (Rosenthal and McKeown, 2015). Not all of these previous studies are, however, directly comparable, e.g., since more narrowly or broadly defined categories were used and/or larger training data sets or external lexical resources.

The next step includes an expansion of the training and evaluation sets, as well as to involve a second annotator to measure inter-annotator agreement and to create a gold standard. Without this measure of reliability, the annotated corpus cannot be considered complete. However, as a snapshot of its current status, the annotations have been made publicly available.[2] Future work also includes studying to what extent a topic-independent classifier detects dis/agreement in general. If dis/agreement is frequently conveyed by means specific to the topic of the debate, relations between the content of the debate posts need to be modelled, to be able to analyse reformulations/expansions/elaborations of previous posts.

5 Conclusion

To be able to train a topic-independent classifier for detecting dis/agreement in online debate fora, a corpus annotated for topic-independent expressions of dis/agreement is a useful resource. Here, the first step towards creating such a resource was taken. A debate forum corpus consisting of 27 discussion threads was annotated for topic-independent expressions conveying dis/agreement. Among the 175 annotated expressions (43 for *agreement* and 132 for *disagreement*), 163 were unique, which shows that there is a large variation in expressions used.

This variation might be one of the reasons why the task of detecting *dis/agreement* was difficult. 10-fold cross-validation on an additional set of 2,000 randomly selected sentences annotated for sentence-level *dis/agreement* resulted in a precision of 0.46 and a recall of 0.47 for *agreement* and a precision of 0.54 and a recall 0.56 for *disagreement*. Results for *disagreement*, however, decreased when the model was applied on held-out data (precision 0.29, recall 0.50). Better results were achieved for the task of distinguishing agreement from disagreement, i.e., not including neutral sentences, but recall for the more infrequently occurring category *agreement* was still low.

[2] http://bit.ly/1Ux8o7q

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

Kelsey Allen, Giuseppe Carenini, and Raymond T. Ng. 2014. Detecting disagreement in conversations using pseudo-monologic rhetorical structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1169–1180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jacob Andreas, Sara Rosenthal, and Kathleen McKeown. 2012. Annotating agreement and disagreement in threaded discussion. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 818–822.

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 48–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Oliver Ferschke. 2014. *The Quality of Content in Open Online Collaboration Platforms: Approaches to NLP-supported Information Quality Management in Wikipedia*. Dissertation, Technische Universität Darmstadt, July.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sebastian Germesin and Theresa Wilson. 2009. Agreement detection in multiparty conversation. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09, pages 7–14, New York, NY, USA. ACM.

Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker, editors. 2014. *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Stroudsburg, PA, USA.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137, Stroudsburg, PA, September. Association for Computational Linguistics.

Sangyun Hahn, Richard Ladner, and Mari Ostendorf. 2006. Agreement/disagreement classification: exploiting unlabeled data using contrast classifiers. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 53–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dusting Hillard, Mari Ostendorf, and Elisabeth Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Amita Misra and Marilyn Walker. 2013. Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of the SIGDIAL 2013 Conference*, pages 41–50, Stroudsburg, PA, USA, August. Association for Computational Linguistics.

Junko Mori. 1999. *Negotiating agreement and disagreement in Japanese : connective expressions and turn construction*. J. Benjamins Pub. Co, Amsterdam.

Bernd Opitz and Cäcilia Zirn. 2013. Bootstrapping an unsupervised approach for classifying agreement and disagreement. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*. Linköping Univ. Electronic Press, Linköping.

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.

Sara Rosenthal and Kathleen McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *SIGDIAL 2015 Conference*, pages 168–177, Stroudsburg, PA, USA. Association for Computational Linguistics.

John R. Searle. 1976. A Classification of Illocutionary Acts. *Language in Society*, 5(1):1–23.

Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for nlp-assisted text annotation. In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marilyn A. Walker, Pranav An, Jean E. Fox Tree, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 23–25.

Jie Yin, Paul Thomas, Nalin Narang, and Cecile Paris. 2012. Unifying local and global agreement and disagreement classification in online debates. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 61–69, Stroudsburg, PA, USA. Association for Computational Linguistics.