Bootstrapping an Unsupervised Approach for Classifying Agreement and Disagreement

Bernd Opitz, Cäcilia Zirn

University of Mannheim

jopitz@mail.uni-mannheim.de, caecilia@informatik.uni-mannheim.de

ABSTRACT

People tend to have various opinions about topics. In discussions, they can either agree or disagree with another person. The recognition of agreement and disagreement is a useful prerequisite for many applications. It could be used by political scientists to measure how controversial political issues are, or help a company to analyze how well people like their new products. In this work, we develop an approach for recognizing agreement and disagreement. However, this is a challenging task. While keyword-based approaches are only able to cover a limited set of phrases, machine learning approaches require a large amount of training data. We therefore combine advantages of both methods by using a bootstrapping approach. With our completely unsupervised technique, we achieve an accuracy of 72.85%. Besides, we investigate the limitations of a keyword based approach and a machine learning approach in addition to comparing various sets of features.

KEYWORDS: Text Classification, Agreement, Disagreement, Opinion Mining.

1 Introduction

It is commonly known that humans do not always agree with each other. Where people disagree, discussions emerge. Discussions are an important manner of communication. For many purposes such as analyses in political sciences, it is relevant to have knowledge about agreement and disagreement. For instance, by measuring the degree of disagreement in ongoing discussions it can be compared how controversial different political issues are. To give another example, detecting heated discussions in online conversations could signal the need for a moderator or mediator to get involved. Knowledge about agreement and disagreement might further be useful for identifying groups that support or oppose a given opinion. In the context of business meetings, it might help to identify decisions and controversy. Furthermore, detecting agreement and disagreement helps understanding social dynamics (Hillard et al., 2003) and can generally aid summarizing conversations or discussions (Janin et al., 2003).

We intend to develop a method to recognize agreement and disagreement (referred to as A/D from now on) automatically, which turns out to be a challenging task. The most intuitive approach might be to watch out for indicating phrases like "I disagree" or "that's wrong". We compiled lists of expressions for both agreement and disagreement based on existing research and classified text snippets based on their occurrences. While this simple approach achieves a high precision, it is not able to make informed choices on how to classify instances that either do not contain any key phrase or even contain indicators for both classes. Another straightforward approach is using machine learning. While this works quite well, the drawback of this approach is that it requires a large amount of manually annotated training data. In order to work around these problems we developed a bootstrapping approach combining the advantages of both previous approaches, which is the main contribution of this paper. In the bootstrapping approach, we first classify a part of the data set by using keywords only. We then take these classified instances as the training data for machine learning and classify the remaining instances with the resulting classifier. For comparison, we describe and evaluate keyword based and machine learning approaches as well.

2 Related Work

Detection of A/D is quite a new research field and as such there is only limited work about it. However, the interest seems to increase, as most of the publications we are aware of were published in 2012.

The research of (Galley et al., 2004) is very close to our work. They applied machine learning techniques, more precisely Bayesian Networks, to the ICSI Meeting corpus (Janin et al., 2003), a collection of human-to-human multi-party conversations, with the preliminary goal of identifying A/D. To do so they used a wide-ranging set of global contextual features as well as local features consisting of structural, durational and lexical features. Among the lexical features, they used positive and negative polarity adjectives as known from sentiment analysis. Their approach achieved an accuracy of 86.9%. Adjectives expressing polarity are commonly used for sentiment analysis, as in the research of (Fahrni and Klenner, 2008) and (Moghaddam and Popowich, 2010). Moghaddam et al. made use of adjective polarity to classify the opinion expressed in a product review as positive, negative or neutral. Their approach reached an accuracy of 73% and thus outperforms Naïve Bayes classifiers which usually achieve accuracies of 58-64% in the same area.

Consistent with the insights outlined in the papers listed above, (Hillard et al., 2003) and (Galley

et al., 2004) state that "lexical information make the most helpful local features", so these were also applied in analyses for our work. A wide range of the features we used, especially discourse markers and repeated punctuation, were chosen based on the findings in (Abbott et al., 2011). In their work they made use of word-based features, among others, to analyze online dialogic discussions concerning subjects such as politics, society or religion in an attempt to recognize A/D. Their approach was carried out using two different Machine Learning algorithms, namely the NaiveBayes and JRip implementations in the Weka toolkit. The success of their approach was an accuracy of 68% which is 5% higher than a simple unigram baseline.

(Yin et al., 2012) analyze online discussions for agreement and disagreement using machine learning methods, among them SVMs which perform quite well. The features they use are similar to our, in so far that they too use e.g. sentiment, emotional and durational features. These features include keywords such as discourse markers as well as occurrences of special characters such as question or exclamation marks, the sentiment polarity of posts, the length of a post and foul words. They developed a multistage process for agreement and disagreement detection, which first decides whether two posts in a discussion agree or disagree with each other. In the second step, these results are aggregated and then in the third step the position towards the initial post in the discussion is determined as the "global" position. The experiments were carried out on two corpora and in the best case achieved an accuracy of 64% and an F-measure of 77%.

(Mukherjee and Liu, 2012) analyze product reviews and comments on them. Agreement and disagreement detection plays a role in their analyses because reviews that people agree with may be more useful than those that a lot of people disagree with. One reason for that is that fake reviews may receive a lot of disagreeing responses as fake reviews are often authored by people who are not users of the product which commenters (as actual users) can usually detect.

Besides approaches to automatically identify agreement and disagreement, new corpora with manual agreement/disagreement annotations were released. (Walker et al., 2012) extracted discussions from 4forums.com and annotate the topic of the discussion as well as agreement of the posts including additional information about the type of agreement (e.g. sarcasm) and its degree. (Andreas et al., 2012) contains data of livejournal blogs and Wikipedia discussion forums. Maintaining the discussion thread structure, they annotate agreement and disagreement and the mode, i.e. either a direct/indirect response or a direct/indirect paraphrase.

3 Agreement and Disagreement

Agreement and disagreement as commonly understood can occur in various forms. Generally, agreement is defined as "harmony of opinion, action, or character" (Merriam-Webster.com, 2012a) whereas disagreement is defined as "the state of being at variance" (Merriam-Webster.com, 2012b). In a wider sense, agreement or disagreement are also present if a person positively respectively negatively refers to another person's statement. Beyond that, agreement and disagreement can either be direct or indirect. Direct agreement or disagreement is present if a speaker explicitly aligns with one or more other speakers (see Example 1 and Example 2).

Example 1*: Example of explicit and direct positive alignment.*
"I agree.", "Great idea", "As you mentioned earlier"

Indirect alignment, such as a shared or contradictory opinion between two speakers can be

Example 2: Example of explicit and direct negative alignment.
"That's wrong.", "That's questionable." , "That's a terrible idea."

seen in Example 3 and Example 4.

Example 3: Example of indirect alignment move: shared opinion.
Speaker1: *I think we should elect John Kerry.*
Speaker2: *John Kerry should be elected as President of the United States.*

Example 4: Example of indirect alignment move: Contradicting opinion (disagreement that should not be coded).
Speaker1: *So you don't think we should use the [. . .] Iraqi government casualty number in this section?*
Speaker2: *The number doesn't come from the Iraqi government, it comes from the LA Times.*

In this work, we restrict our experiments to recognizing direct and explicit A/D only, as the corpus we are using is limited to these annotations.

3.1 Classifying Agreement and Disagreement

Our goal was to come up with an approach for identifying A/D that is independent of manually annotated training data, yet able to identify sentences that do not contain an explicit, predefined keyword. We hypothesized that classification based on keywords only would yield a high precision but low recall, i.e. that it would be able to correctly classify instances containing keywords, but failing to classify any of the others. We will show later that this intuition turned out to be true (see subsection 4.4). A machine learning approach, on the other hand, requires a large amount of training data, which is time and labor intensive to create. We therefore developed a bootstrapping approach combining the keyword approach with machine learning. In brief, the idea is the following: First, we pick those sentences from a corpus that contain predefined key expressions that indicate either agreement or disagreement and classify them accordingly. We then use the result as training data for a machine learning approach.

For the first part of the bootstrapping approach, the keyword based labeling process, we begin by collecting keywords that indicate utterances of agreement and disagreement. In the following, by keyword we refer to single terms as well as expressions or phrases consisting of more than one word,, e.g. "conversely" and "on the contrary". We then check the instances within the data set for the occurrence of those keywords. The range of an instance, i.e. the unit that is classified, depends on the data set, it could be defined as a sentence or a coherent sequence of sentences that comprise a complete utterance of a speaker. In our case, an instance consists of a speaker's turn, which will be described in detail in subsection 4.1. If the instance contains more keywords for agreement, we label it as agreement; in case we find more keywords for disagreement, we label it as disagreement. For a more precise classification, we also consider negations. A negation is an expression that negates the meaning of a phrase. We consider keywords directly following a negation as an indicator for the opposite label. If an instance contains an even number of cues for agreement and disagreement or if it does not contain any keywords at all, it is not labeled in this step. It is important to point out that neither the handling of negations nor the classification based on the key expressions for the instances are necessarily correct.

However, as we will show later in Table 1, the precision of this simple classifier achieves over 61% for Agreement and over 90% for Disagreement.

The resulting annotations derived by this keyword based approach will be used as training data for machine learning. The remaining part of the corpus that could not be classified in the first step - either because it did not contain any keywords or because it contained an equal number of agreeing and disagreeing keywords - will then be classified by trained the machine learning classifier. The features we used for machine learning are described below in 4.2.

After these two steps, all instances are classified as either Agreement or Disagreement. The salient advantage of this approach is that it does not require any manual annotations and thus represents an unsupervised approach. To evaluate the potential of this approach while discovering its limitations, we compared it to two different baselines. On the one hand, we will evaluate the results of using the keyword-based approach only. On the other hand, we explore the performance of a standard machine learning scenario training on manually annotated labels. Furthermore, we will experiment with different combination of features to find out how helpful they are for the classification task of distinguishing between agreement and disagreement.

4 Evaluation

In the following, we will explain how we evaluate the different approaches for classifying agreement and disagreement and which features we use. Furthermore, we describe the data in more detail.

4.1 Corpus

For the experiments in this work we used the Authority and Alignment in Wikipedia Discussions (AAWD) corpus by (Bender et al., 2011), which is a set of 365 discussions taken from 47 Wikipedia talk pages. The corpus contains annotations for alignment and authority claims by three independent annotators and is available in multiple versions and languages. Experiments are solely based on the merged alignment annotations – i.e. the combination of the three annotators – from the English version. From this corpus, we take 3390 turns from 211 discussions that are annotated for alignment moves. Out of these 3390 instances, we take all the sentences that have been annotated for either agreement or disagreement but not both. This leaves us with 2302 sentences out of which 478 (roughly 21%) are annotated for agreement, the remaining 1824 sentences are labeled as disagreement. All data files of the corpus are supplied in extended tab-delimited format (xtdf).

4.2 Features

To distinguish between agreement and disagreement, we use various features that we selected based on findings in existing research. All features' values were calculated at the level of turns, where a turn is defined as a "contiguous body of text on the [...] page that was modified as part of a single revision" (Bender et al., 2011).

1. *Keywords:* Based on the findings of related work(Schourup, 1999; Anand et al., 2011)), we made use of specific keywords to identify A/D. They serve as well for the keyword based approach as for the machine learning. We compiled separate keyword lists for different types of keywords. The keywords were collected by introspection and looking at related work as well as the annotation guidelines of the corpus. It must be noted that

expressions taken from the corpus itself would weaken the generality of our approach. For that reason, we included only few expression from the annotation guidelines of the corpus and only those, which we considered sufficiently generic.

As an indicator for agreement, we created distinctive lists:

- positive adjectives, such as “excellent” or “perfect”
- positive alignment cues, such as “i agree” or “that’s right” (Ali, 2011)
- positive discourse markers, such as “I think” or “I know” (Abbott et al., 2011)
- various additional agreement keywords, such as “agreement”

For disagreement, we compiled the following lists:

- negative adjectives, such as “questionable” or “unsatisfactory”
- negative alignment cues, such as “i doubt that” or “that’s irrelevant” (Ali, 2011)
- negative discourse markers, such as “actually” or “but” (Abbott et al., 2011)
- insults/swear words, such as “idiot” or “narrow-minded”
- various additional disagreement keywords, such as “disagreement”

For the machine learning approaches, we consider each of these lists as one separate feature. If a term of the particular list is located in the text, the value of the feature is increased by one. Please note that regarding the machine learning, we do not explicitly connote the lists with agreement or disagreement, but the algorithm is supposed to learn from the training data which label is indicated by a particular list. In subsection 4.3, we will give more details about how we use the lists for the keyword based approach.

2. *Unigrams:* Most common in text analysis, the unigrams of the text - i.e. its words - are used as features modeling a bag of words approach.

3. *Word count:* For this feature, the total amount of words in a turn is counted. According to (Cohen, 2002), disagreeing statements are usually longer than agreeing statements.

4. *Pronouns:* The hypothesis is that there is a difference in the amount of usage in agreement and disagreement. This is supported by (Anand et al., 2011) , who found that pronouns such as you occurred more frequently in rebuttals. We treat them as distinct features.

5. *Negations:* In general, negations describe the concept of reversing the value of a statement. For instance, “This is not true” obviously is the opposite of “This is true”. As for disagreement, people tend to negate the utterances of the previous speaker, we expect them to contain a higher amount of negations. For the experiments, we collect various keywords for negations and use the sum of their negations as one feature.

6. *Special characters:* We counted special characters such as *? ! , . ;* - _ each individually, resulting in 7 distinct features.

7. *Repeated punctuation:* Repeated punctuation such as **!!** has a different meaning than simple punctuation (Abbott et al., 2011). For example, if a person questions the utterance of a previous speaker, he might indicate this by using “?!” marking a rhetorical question. We implement the repeated punctuation feature by counting occurrences of more than one question or exclamation mark in a row, i.e. any combination of question and exclamation marks.

8. *Formatting:* Participants of written discussions tend to highlight certain points, such as the strong points of their argument, using markup text to draw special attention to it. It can be compared to raising the voice or emphasizing something using intonation in spoken language. We capture this by counting the number of formatting instructions for **bold** and *italic* markup each as an individual feature.

4.3 Experiments

As mentioned before, we compare our bootstrapping approach to straightforward baseline approaches. In each case, the entities that we classify - thus the instances - consist of a speaker's turn as described in subsection 4.1. More precisely, we compare the following approaches:

1. Keywords baseline:
   Keyword-based classification for a single instance works as follows: We first initialize a score with 0. For each occurrence of a keyword indicating agreement, we add 1. For each occurrence of a keyword for disagreement, we subtract 1. For a more precise classification, we also consider negations, utilizing the same list we described in subsection 4.2: If a negation, i.e. a keyword from the list of negations directly precedes an agreement keyword, instead of adding 1 to the score, 1 is subtracted. The same applies analogously for negated keywords for disagreement. If the final score is greater than 0, the instance is considered to be agreement, if it is less than 0 it is considered to be disagreement. A score of 0 indicates that this instance cannot be classified which may have two reasons: either no keyword at all occurred in the instance or the amount of keywords for agreement and disagreement was even. Therefore, we calculate the performance of the keyword baseline in two ways:

   - *Containing keywords:* We regard those instances only that could be classified and calculate precision and recall relative to this number.
   - *Whole dataset:* We calculate precision and recall for the whole dataset. Instances that have a score of 0 are regarded as misclassified for agreement as well as for disagreement.

2. Machine learning baseline
   For the machine learning, we extract all features described in subsection 4.2 for each instance. We compare different combinations of those features.

   - *Unigrams:* To explore how efficient this task can be solved using a bag-of-words approach, we consider unigrams only.
   - *Others:* In this scenario, all features except for unigrams are used.
   - *Unigrams + others:* Finally, all features are used. We want to point out that there might be an overlap between some of the features, concerning unigrams, keywords, pronouns and the number of negations.

   Please note: For machine learning we only use the number of negation keywords and do not consider their effect on any other features.

3. Bootstrapping:
   The bootstrapping approach is implemented as described in 3.1. We first apply the

keyword approach to the data. In a second step, we train an SVM on the data that could be classified, and use the trained SVM to classify the data that had not been assigned a class by the keyword approach. As for the machine learning baseline, we compare the three combination of features:

- *Unigrams*
- *Others*
- *Unigrams + others*

For each of these three configurations, we separately calculate the performance for the instances that were classified in the second step, i.e. by the SVM. In contrast to the baseline, we do not apply cross-validation in the Bootstrapping approaches. The reason for that is this is that the number of seeds is already quite low and cross-validating would provide an even lower number of training instances for each run.

4. Bootstrapping upper bound:
For better comparison of the bootstrapping approach, we evaluate its upper bound: If the label assigned by the keyword approach is wrong, we correct it for those instances. In this way, we can evaluate how the performance of the bootstrapping approach is influenced by the mistakes of the keyword approach.

All machine learning experiments were conducted using the tool Weka[1]. We used the SMO implementation of a Support Vector Machine (SVM) which is available in Weka and evaluated its performance using 10-fold cross-validation with stratified sampling.

4.4 Results

Table 1 shows the results for the keyword baseline. Out of 2302 instances, only 954 could be classified (41% of the data). However, 80.29% of these classifications were correct. As can be seen in the first line of the table ("containing keywords"), the approach achieves good results for both classes with an F-measure of 68.49% for agreement and 85.67% for disagreement. It even achieves a precision of 90.50% for disagreement. Due to the fact that more than half of the data was not classified at all, however, the performance for the whole corpus is rather low, with an F-measure of 20.99% and 42.26% respectively. As the input data was unbalanced, the following tables always show measures for two cases: the Agreement part shows the measures under the assumption that agreement was considered the positive class and disagreement was considered the negative class, analogously for the Disagreement part.

Keyword baseline						
	Agreement			Disagreement		
	P	**R**	**F_1**	**P**	**R**	**F_1**
containing keywords	61.26	77.57	**68.46**	**90.50**	81.33	**85.67**
whole dataset	13.92	42.68	20.99	67.22	30.81	42.26

Table 1: **P**recision, **R**ecall and **F**-Measure for the keyword based approach.

[1] `http://www.cs.waikato.ac.nz/ml/weka/`

Machine learning baseline						
	Agreement			Disagreement		
	P	**R**	**F_1**	**P**	**R**	**F_1**
unigrams	64.19	52.51	57.77	88.12	92.32	90.17
others	74.16	32.43	45.12	84.57	97.04	90.38
unigrams + others	66.92	56.28	**61.14**	89.00	92.71	**90.82**

Table 2: **P**recision, **R**ecall and **F**-Measure for the machine learning approach.

The machine learning baseline surprised with comparably high results (table 2). Among the three combinations of features, the best result was achieved using all features (unigrams + others). Especially for the classification of disagreement, it shows excellent results with precision, recall and F-measure around 90%. For agreement, due to a low recall of 56.28%, it results in an F-measure of 61.14%, which is still significantly better than the keyword baseline on the complete data set. We assume that the reason for the higher results in the class of disagreement is caused by the imbalanced data set of which about 80% of the instances are labeled as disagreement. Comparing the three feature combinations, it is interesting to see that for disagreement, both the unigrams and the other features alone achieve good results. For agreement, which seems to be more difficult to classify, they supplement each other and their combination boosts the results.

Bootstrapping						
	Agreement			Disagreement		
	P	**R**	**F_1**	**P**	**R**	**F_1**
whole dataset	43.40	52.30	**47.44**	86.79	82.13	**84.39**
(2. step)	18.93	21.40	20.09	84.76	82.61	83.65
Bootstrapping - upper bound						
(2. step)	37.84	19.53	**25.77**	86.01	93.91	**89.79**

Table 3: **P**recision, **R**ecall and **F**-Measure for the bootstrapping approach using all features. (2. step) gives the results for the subset of the data that was labeled in the second step, i.e. by the SVM trained on the data labeled by keywords.

The bootstrapping approach was performed using the same sets of features for the machine learning part as for the machine learning baseline. Again the combination of all features showed better results than using unigrams and the other features separately, we therefore omit the latter ones in the table for better readability. The results are shown in table 3. For the whole data set, the bootstrapping approach yielded in an accuracy of 72.85% with an F-measure of 47.44% for agreement and 84.39% for disagreement. As expected, the results are not as high as the machine learning baseline, but one has to keep in mind the fact that this approach is completely unsupervised and does not require any labeled data. To explore whether the correct classifications are mainly due to the labeling process in the first step, namely the keyword based approach, and to get a better insight in the performance of the classifier that was trained on the automatically acquired training data, we separately measure the correctness of the instances labeled in the second step. We find that the results for disagreement are quite high (precision and recall both above 80%), but very low for agreement (precision and recall around 20%).

This might be partly caused by the fact that the data set that is labeled in the second step contains only 215 instances of agreement in contrast to 1133 instances of disagreement.

Given that the machine learning approach itself does not seem to be inefficient, as demonstrated by the baseline, we further investigated whether the performance of the trained classifier in the bootstrapping approach suffers from the misclassifications of the keyword based approach in the first step, or whether some of the agreement instances are just very difficult to be classified at all. This gives us an upper bound for the bootstrapping approach on the data set. For this experiment, we correct all misclassifications produced by the keyword approach in the first step and train the classifier on the true labels. The results are displayed in the last row of table 3. Although the performance is increased, it still achieves an F-measure of 25.77% only for classifying agreement.

4.5 Error Analysis

We analyze some of the misclassified instances to figure out the problems of the particular approaches. The following list shows some examples of misclassified instances. We come up with explanations for the errors, though it cannot always be determined with certainty.

- *Keyword*: Good luck putting your pov [point of view] in the article.
  This instance is misinterpreted as Agreement due to the occurrence of the word "good". The obvious deficiency of the keyword approach is its inability to detect irony.

- *Bootstrapping(others)*:

  - It's obvious and disgusting POV. → 6 words, 1 period
  - I actually agree to an extent, in that at an article addressing past powerful Hurricanes that have struck the U.S. could be relevant for perspective. → 26 words, 1 negative discourse marker, 1 other positive keyword, 3 periods, 1 comma, 1 "I"

  The first instance is misclassified as agreement which seems to be due to a lack of useful information. Note that this instance also shows that the lists of adjectives used are far from complete. The second instance is misclassified as disagreement which may be due to the negative discourse marker, word count and punctuation, which the model may have trained to be connected with disagreement.

- *Bootstrapping(unigrams + others)*: Your edit is cool with me., 6 words, 1 period
  This instance is misclassified as disagreement. The same instance is also misclassified by the bootstrapping approach using unigrams. What is surprising, is that bootstrapping using all features except for unigrams would label this instance as agreement (since its non-unigram properties are: 6 words, 1 period).

These examples show that the combination of all features fixes some misclassifications but also introduces errors. Thus, finding the right feature set is a balancing act.

5 Conclusion and Future Work

In a supervised scenario, we were able to achieve F-measures of 61.14% for classifying agreement and 84.39% for disagreement. Combining a keyword based approach with machine

learning, we implemented a completely unsupervised approach that is able to classify agreement and disagreement with an accuracy of 72.85%, achieving an F-measure of 47.44% and 84.39% for the particular classes. In all experiments, we merely used local features, i.e. only features that were extracted from the particular text unit that is assigned a label. Our experiments show that while unigrams alone already seem to be able to distinguish between agreement and disagreement statements, features like the length of an utterance, punctuation or information about negations add to their performance. We see types of additional information that might further improve the classification of agreement and disagreement. As shown in (Galley et al., 2004), knowledge about the context could be included by extracting information about surrounding pieces of the discussion, especially the utterances of previous speakers. On the other hand, external sources could be used to derive knowledge about opposing opinions for a topic. Furthermore, we assume that methods for detecting sarcasm and irony would increase the performance of this task. Up to this point, we only considered explicit agreement and disagreement. In the future, we also plan to investigate the limitations of recognizing implicit agreement and disagreement. Another point for potential improvement is the way the Bootstrapping approach was applied. In our experiments we did not modify the classifier once it was trained. One could carry out classification iteratively, i.e. carry out classify instances based on previous classifications.

References

(2011). Alignment annotation guidelines (version 14). [Online; accessed 05 March 2012].

Abbott, R., Walker, M., Anand, P., Fox Tree, J. E., Bowmani, R., and King, J. (2011). How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 2–11, Portland, Oregon. Association for Computational Linguistics.

Anand, P., Walker, M., Abbott, R., Fox Tree, J. E., Bowmani, R., and Minor, M. (2011). Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 1–9, Portland, Oregon. Association for Computational Linguistics.

Andreas, J., Rosenthal, S., and McKeown, K. (2012). Annotating agreement and disagreement in threaded discussion. In *Proceedings of the 8th International Conference on Language Resources and Computation (LREC), Istanbul, Turkey, May*.

Bender, E. M., Morgan, J. T., Oxley, M., Zachry, M., Hutchinson, B., Marin, A., Zhang, B., and Ostendorf, M. (2011). Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 48–57, Portland, Oregon. Association for Computational Linguistics.

Cohen, S. (2002). A computerized scale for monitoring levels of agreement during a conversation. In *Proc. of the 26th Penn Linguistics Colloquium*.

Fahrni, A. and Klenner, M. (2008). Old wine or warm beer: Target-specific sentiment analysis of adjectives. In *Proc. of the Symposium on Affective Language in Human and Machine, AISB*, pages 60–63.

Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E. (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.

Hillard, D., Ostendorf, M., and Shriberg, E. (2003). Detection of agreement vs. disagreement in meetings: training with unlabeled data. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers - Volume 2*, NAACL-Short '03, pages 34–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al. (2003). The icsi meeting corpus. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–364. IEEE.

Merriam-Webster.com (2012a). "agreement". [Online; accessed 6 April 2012].

Merriam-Webster.com (2012b). "disagreement". [Online; accessed 6 April 2012].

Moghaddam, S. and Popowich, F. (2010). Opinion polarity identification through adjectives. *CoRR*, abs/1011.4623.

Mukherjee, A. and Liu, B. (2012). Modeling review comments. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 320–329, Stroudsburg, PA, USA. Association for Computational Linguistics.

Schourup, L. (1999). Discourse markers. *Lingua*, 107(3-4):227 – 265.

Walker, M., Anand, P., Tree, J. F., Abbott, R., and King, J. (2012). A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC*, pages 23–25.

Yin, J., Thomas, P., Narang, N., and Paris, C. (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.