I Couldn't Agree More: The Role of Conversational Structure in Agreement and Disagreement Detection in Online Discussions

Sara Rosenthal
Columbia University
Computer Science Department
NY, NY, USA
sara@cs.columbia.edu

Kathleen McKeown
Columbia University
Computer Science Department
NY, NY, USA
kathy@cs.columbia.edu

Abstract

Determining when conversational participants agree or disagree is instrumental for broader conversational analysis; it is necessary, for example, in deciding when a group has reached consensus. In this paper, we describe three main contributions. We show how different aspects of conversational structure can be used to detect agreement and disagreement in discussion forums. In particular, we exploit information about meta-thread structure and accommodation between participants. Second, we demonstrate the impact of the features using 3-way classification, including sentences expressing disagreement, agreement or neither. Finally, we show how to use a naturally occurring data set with labels derived from the sides that participants choose in debates on createdebate.com. The resulting new agreement corpus, Agreement by Create Debaters (ABCD) is 25 times larger than any prior corpus. We demonstrate that using this data enables us to outperform the same system trained on prior existing in-domain smaller annotated datasets.

1 Introduction

Any time people have a discussion, whether it be to solve a problem, discuss politics, products, or more casually, gossip, they will express their opinions. As a conversation evolves, the participants of the discussion will agree or disagree with the views of others. The ability to automatically detect agreement and disagreement (henceforth referred to as (dis)agreement) in the discussion is useful for understanding how conflicts arise and are resolved, and the role of each person in the conversation. Furthermore, detecting (dis)agreement has been found to be useful for other tasks, such as detecting subgroups (Hassan et al. 2012), stance (Lin et al., 2006; Thomas et al., 2006), power (Danescu-Niculescu-Mizil et al., 2012; Biran et al., 2012), and interactions (Mukherjee and Liu, 2013).

In this paper, we explore a rich suite of features to detect (dis)agreement between two posts, the *quote* and the *response* (Q-R pairs (Walker et al., 2012)), in online discussions where the response post directly succeeds the quote post. We analyze the impact of features including meta-thread structure, lexical and stylistic features, Linguistic Inquiry Word Count categories, sentiment, sentence similarity and accommodation. Our research indicates that conversational structure, as indicated by meta-thread information as well as accommodation between participants, plays an important role. *Accommodation* (Giles et al., 1991), is a phenomenon where conversational participants adopt the conversational characteristics of the other participants as conversation progresses. Our approach represents accommodation as a complex interplay of semantic and syntactic shared information between the Q-R posts. Both meta-thread structure and accommodation use information drawn from both the quote and response; these features provide significant improvements over information from the response alone.

We detect (dis)agreement in a supervised machine learning setting using 3-way classification (agreement/disagreement/none) between Q-R posts in several datasets annotated for agreement, whereas most prior work uses 2-way classification. In many online discussions, none (i.e., the lack of (dis)agreement) is the majority category so leaving it out makes it impossible to accurately classify the majority of the sentences in an online discussion with a binary classification model.

We also present a new naturally occurring agreement corpus, Agreement by Create Debaters (ABCD), derived from a discussion forum web-

Proceedings of the SIGDIAL 2015 Conference, pages 168–177,
Prague, Czech Republic, 2-4 September 2015. ©2015 Association for Computational Linguistics

Example of disagreement in an ABCD discussion indicated by different sides (Against and For).
Abortion is WRONG! God created that person for a reason. If your not ready to raise a kid then put it up for adoption so it can be with a good family. Dont murder it! Its wrong. It has a life. If you can have sex then you should be ready for the consequences tht come with it! **Side:** ***Against***
Those who were raped through the multiple varieties of means, are expected to birth this child although it was coerced rape. I don't think so. Taking a woman's right to choice is wrong regardless what a church or the government suggests. **Side:** ***For***
Example of agreement in an ABCD discussion indicated by the same side (Against).
HELL NO! ... KILLING A INNOCENT BABY ISN'T GONNA JUST GO AWAY YOU WILL HAVE TO LIVE WITH THE GUILT FOREVER!!!!!!! **Side:** ***Against***
——————————————> That is soo true living with the guilt forever know you murder you child it would have been even better if the murder hadn't been born. **Side:** ***Against***
Example of no (dis)agreement in an ABCD discussion between the original post and a response.
Coke or Pepsi?
They taste the same no big difference between them for me

Table 1: Examples of Agreement, Disagreement, and None in ABCD discussions

site, createdebate.com, where the participants are required to provide which side of the debate they are on. This enabled us to easily gather over 10,000 discussions in which there are over 200,000 posts containing (dis)agreement or the lack of, *25 times larger* than any pre-existing agreement dataset. We show that this large dataset can be used to successfully detect (dis)agreement in other forums (e.g. 4forums.com and Wikipedia Talk Pages) where the labels cannot be mined, thereby avoiding the time consuming and difficult annotation process.

In the following sections, we first discuss related work in spoken conversations and discussion forums. We then turn to describe our new dataset, ABCD, as well as two other manually annotated corpora, Internet Argument Corpus (IAC), and Agreement in Wikipedia Talk Pages (AWTP). We explain the features used in our system and describe our experiments and results. We conclude with a discussion containing an error analysis of the hard cases of (dis)agreement detection.

2 Related Work

Early prior work on detecting (dis)agreement has focused on spoken dialogue (Galley et al., 2004; Hillard et al., 2003; Hahn et al., 2006) using the ICSI meeting corpus (Janin et al., 2003). Germesin and Wilson (2009) detect (dis)agreement on dialog acts in the AMI meeting corpus (Mccowan et al., 2005) and Wang et al (2011a, 2011b) detect (dis)agreement in broadcast conversation in English and Arabic. Prior work in spoken dialog has motivated some of our features (e.g., lists of agreement and disagreement terms, sentiment and n-grams).

Recent work has turned to (dis)agreement detection in online discussions (Yin et al., 2012; Abbott et al., 2011; Misra and Walker, 2013; Mukherjee and Liu, 2012). The prior work performs 2-way classification between agreement and disagreement using features that are lexical (e.g. n-grams), basic meta-thread structure (e.g. post length), social media features (e.g. emoticons), and polarity using dictionaries (e.g. SentiWordNet). Yin et al (2012), detect local and global (dis)agreement in discussion forums where people debate topics. Their focus is global (dis)agreement, which occurs between a post and the root post of the discussion. They manually annotated posts from US Message Board (818 posts) and Political Forum (170 posts) for global agreement. This approach ignores off-topic posts in the discussion which can indicate incorrect labeling and the small size makes it difficult to determine how consistent their results would be in unseen datasets. Abbott et al (2011), look at (dis)agreement using 2,800 annotated posts from the Internet Argument Corpus (IAC) (Walker et al., 2012). Their work was extended to topic independent classification by Misra and Walker (2013). Since it is the largest previously used corpus, we use the IAC corpus in our experiments. Lastly, Mukherjee and Liu (2012) , developed an SVM+Joint Topic Model classifier to detect (dis)agreement using 2,000 posts. They studied accommodation across (dis)agreement by classifying over 300,000 posts and explore the difference in accommodation across LIWC categories. While they did not implement accommodation, they found that it is more common in agreement for most categories, except for a few style dimensions (e.g. negation) where it is reversed. This paper highly motivates our inclusion of accommodation for (dis)agreement detection.

In other work, Opitz and Zirn (2013) detect (dis)agreement on sentences using the Authority and Alignments in Wikipedia Discussions corpus (Bender et al., 2011) which is different than the AWTP corpus used in this paper. In the future we would like to explore whether we could incorporate this corpus into ours. Wang and Cardie (2014) also detect (dis)agreement on the sentence and segment[1] level using this corpus and the IAC.

Our approach differs from prior work in that it explores (dis)agreement detection on a large, naturally occurring dataset where the annotations are derived from participant information. We explore new features representing aspects of conversational structure (e.g. sentence similarity) and the more difficult 3-way classification task of detecting agreement/disagreement/none.

3 Data

In this work we focus on direct (dis)agreement between quote-response (Q-R) posts in the three datasets described in the following subsections. Across all datasets we only include discussions of depth > 2 to ensure a response chain of at least three people and thus, a thread. We also excluded extremely large discussions to improve processing speed. We only consider entire posts in Q-R pairs.

3.1 Agreement by Create Debaters (ABCD)

Create Debate is a website where people can start a debate on a topic by asking a question. On this site, a debate can be:

- **open-ended**: there is no side
- **for-or-against**: two sided
- **multiple-sides**: three or more sides

In this paper, we only focus on debates of the for-or-against nature where there are two sides. For example, we use a debate discussing whether people are for or against abortion[2] in our examples throughout the paper. In this corpus, the participants in the debate choose what side they are on each time they participate in the discussion. Prior work (Abu-Jbara et al., 2012) has used the side label of this corpus to detect the subgroups in the discussion. We annotate the corpus as follows: the side label determines whether a post (the *Response*) is in agreement with the post prior to it (the *Quote*). If the two labels are the same, then they agree. If the two labels are different, they disagree. When the author is the same for both posts, there is no (dis)agreement as the second post is just a continuation of the first. Finally, the first post and its direct responses do not agree with anyone; the first post does not have a side as it is generally a question asking whether people are for, or against the topic of the debate. Examples of (dis)agreement and none are shown in Table 1. We call this corpus Agreement by Create Debaters or ABCD.

Dataset	Thread Count	Post Count	Agree	Disagree	None
ABCD	9981	185479	38195	60991	86293
IAC	1220	5940	428	1236	4276
AWTP	50	822	38	148	636

Table 2: Statistics for full datasets

Our dataset includes over 10,000 discussions which include 200,000 posts on a variety of topics. Additional statistics for ABCD are shown in Table 2. There are far more disagreements than agreements as people tend to be argumentative when they are debating a topic.

3.2 Internet Argument Corpus (IAC)

The second dataset we use is the IAC (Walker et al., 2012). The IAC consists of posts gathered from `4forums.com` discussions that were annotated on Mechanical Turk. The Turkers were provided with a Q-R pair and had to indicate the level of (dis)agreement using a scale of $[-5, 5]$ where -5 indicated high disagreement, 0 no (dis)agreement, and 5 high agreement. As in prior work with this corpus (Abbott et al., 2011; Misra and Walker, 2013), we converted the scalar values to (dis)agreement with $[-5, -2]$ as disagreement, $[-1, 1]$ as none, and $[2, 5]$ as agreement. In this dataset is it possible for multiple annotations to occur in a single post. We combine the annotation to the post level as follows. We ignored the none annotations unless there was no (dis)agreement. In all other cases, we use the average (dis)agreement score as the final score for the post. 10% of the posts had more than one annotation label. The number of annotations per class is shown in Table 2. Not all Q-R posts in a thread were annotated for agreement as is evident by the ratio of threads to post annotations.

3.3 Agreement in Wikipedia Talk Pages (AWTP)

Our last corpus is 50 Wikipedia talk pages (used to discuss edits) containing 822 posts (see full statistics in Table 2) that were manually annotated as the ATWP (Andreas et al., 2012). Although

[1] a segment is a portion of a post

[2] www.createdebate.com/debate/show/Abortion_9

smaller than the IAC, the advantage to this dataset is that each thread was annotated in its entirety. As in the create debate discussions, disagreement is more common than agreement due to the nature of the discussion. These annotations were on the sentence level where multiple sentences can be part of a single annotation. In 99% of the Q-R posts, there was just one pair of sentences that were annotated with a (dis)agreement label and we used that annotation for the post. When there was one more than one pair, we used the majority annotation. The post was labeled with none only when all sentences within the post had the none label. AWTP was annotated by three different people. Inter-Annotator Agreement (IAA) using the sentence pairs was very high because most annotations were none. Therefore, we computed IAA by randomly sampling an equivalent amount of sentences pairs per label from two of the annotators (A1 & A2) and had the third annotator (A3) annotate all of those sentence pairs. Cohen's κ for A1,A3 was .90 and for A2,A3 was .70 indicating high IAA.

4 Method

We model our data by posts. Each data point (the *Response*) is a single post and its label indicates whether it agrees, disagrees, or none, to the post it is responding to (the *Quote*). The following sections discuss the features used to train our model. Each feature is computed within the entire post. In addition, in all applicable features, we also indicate if the feature occurs in the first sentence of the post. Our analysis showed that (dis)agreement tends to occur in the first sentence of the response.

Meta-Thread Structure features include: 1) **The post is the root of the discussion**: This is useful because the root of the discussion tends to be a question (e.g., "Are you for or against abortion") and thus, does not express (dis)agreement. 2) **The reply was by the same author**: The second post is just a continuation of the first. 3) **The distance, or depth, of the post from the beginning of the discussion**: anyone that replied to the root (Depth of 1) has no (dis)agreement because the root is a question and therefore has no side. The average depth per thread is 4.9 in ABCD, 12.7 in IAC and 6.2 in ATWP, and 4) **The number of sentences in the response**: people who disagree tend to write more than those who agree.

Lexical Features are generated for each post. We use (1-3)gram features and also generate up to 4 possible Part of Speech (POS) tag features (Toutanova et al., 2003) for each word in the post. We include all unigram POS tags and perform Chi-Squared feature selection on everything else. In addition, we also generated small lists of negation terms (e.g. not, nothing; 11 terms in total), agreement terms (e.g. agree, concur; 16 terms in total), and disagreement terms (e.g. disagree, differ; 14 terms in total) and generate a binary feature for each list indicating that the post has one of the terms from the respective list of words. Finally, we also include a feature indicating whether there is a sentence that ends in a question as when someone asks a question, it may be followed by (dis)agreement, but it probably won't be in (dis)agreement with the post preceding it.

Lexical Stylistic Features that fall into two groups are included, **general**: ones that are common across online and traditional genres, and **social media**: ones that are far more common in online genres. Examples of general style features are exclamation points and ellipses. Examples of social media style features are emoticons and word lengthening (e.g. sweeeet).

Linguistic Inquiry Word Count The Linguistic Inquiry Word Count (LIWC) (Tausczik and Pennebaker, 2010) aims to capture the way people talk by categorizing words into a variety of categories such as negative emotion, past tense, and health and has been used previously in agreement (Abbott et al., 2011). The 2007 LIWC dictionary contains 4487 words with each word belonging in one or more categories. We use all the categories as features to indicate whether the response has a word in the category.

Sentiment By definition, (dis)agreement indicates whether someone has the same, or different, opinion than the original speaker. A sentence tagged with subjectivity can help differentiate between (dis)agreement and the lack thereof, while polarity can help differentiate between agreement and disagreement. We use a phrase-based sentiment detection system (Agarwal et al., 2009; Rosenthal et al., 2014) that has been optimized for lexical style to tag the sentences with opinion and polarity. For example, it produces the following tagged sentence "[That is soo true]/*Obj* [living with the guilt forever]/*neg* [know you murder you child]/*neg*..." We use the tagged sentence to generate several opinion-related features. We generate bag of words for all opinionated words in the opinion and polarity phrases, labeling each word as to which class it belongs to (opinion, positive, or negative). We also have binary features indicating the prominence of opinion and polarity (positive or negative).

Sentence Similarity A useful indicator for determining whether people are (dis)agreeing or not is if they are talking about the same topic. We use sentence similarity (Guo and Diab, 2012) to determine the similarity between the Q-R posts. For example the disagreement posts in Table 1 are similar because of the statements *"LIVE WITH THE GUILT FOREVER!!!!!!!"* and *"living with the guilt forever"*. We use the output of the system to indicate whether there are two similar sentences above some threshold and whether all the sentences are similar to one another.

Furthermore, we also look at similar Q-R phrases in conjunction with sentiment. We generate phrases using the Stanford parser (Socher et al., 2013) by adding reasonably sized branches of the parse tree as phrases. We then find the similarity (Guo and Diab, 2012) and opinion (Agarwal et al., 2009; Rosenthal et al., 2014) of the phrases and extract the unique words in the similar phrases as features. We hypothesize that this could help indicate disagreement, for example, if the word "not" was mentioned in one of the phrases, e.g. *"I do not see anything wrong with abortion =/"* vs *"I do see something wrong with abortion ..."*. We also include unique negation terms using the list described in the Lexical Feature section and features to indicate whether there is a similar phrase and if its opinion in the Q-R posts are of the same polarity (agree) or different polarity (disagree).

Accommodation When people speak to each other, they tend to take on the speaking habits and mannerisms of the person they are talking to (Giles et al., 1991). This phenomenon is known as *accommodation*. Mukherjee and Liu (2012) found that accommodation differs among people who (dis)agree. This strongly motivates using accommodation in (dis)agreement detection[3]. We partly capture this via sentence similarity which explores whether they share the same words. We also explore whether Q-R posts use the same syntax (POS, n-grams), copy lexical style, and use the same category of words (LIWC). We use the features as described in prior sections but only include ones that exist in the quote and response.

5 Experiments

All of our experiments were run using Mallet (McCallum, 2002). We experimented with Naive Bayes, Maximum Entropy (i.e. Logistic Regression), and J48 Decision Trees and found that Maximum Entropy consistently outperformed or there was no statistically significant difference to the other classifiers; we only show the results for Maximum Entropy here. We show our results in terms of None, Agreement, and Disagreement F-Score as well as macro-average F-score for all three classes. The ABCD and IAC datasets were split into 80% train, 10% development, and 10% test. We use the entire AWTP dataset as a test set because of its small size. All results shown are using a balanced training set by downsampling and the full test set. It is important to use a balanced dataset for training because the ratio of agreement/disagreement/none differs in each dataset. We tuned the features using the development set and ran an exhaustive experiment to determine which features provided the best results and use that best group of features as an additional experiment in the test sets.

In order to show the impact of our large dataset, we experimented with increasing the size of the training set by starting with 25 posts from each class and increased the size until the full dataset is reached (e.g. 25, 50, 100, ...). We also show a more detailed analysis of the various features using the full datasets. In all datasets, the best experiment includes the features found to be most useful during development and differs per dataset.

We compare our experiments to two baselines. The first is the majority class, which is none. Although none is more common, it is important to note that we would prefer to achieve higher f-score in the other classes as our goal is to detect (dis)agreement. The second baseline is n-grams, the commonly used baseline in prior work. We compute statistical significance using the Approximate Randomization test (Noreen, 1989; Yeh, 2000), a suitable significance metric for F-score.

5.1 Agreement by Create Debaters (ABCD)

Our first experiments were performed on the large ABCD dataset of almost 10,000 discussions described in the Data Section. We experimented with balancing and unbalancing the training dataset and the balanced datasets consistently outperformed the unbalanced datasets. Therefore, we only used

[3] Accommodation wasn't used to classify (dis)agreement.

Features	None	Agree	Disagree	Avg
majority	63.2	0.0	0.0	21.1
n-gram	45.7	35.6	41.3	40.9
n-grams+POS+lex.-style+ LIWC in R	58.7^1	42.2	51.6	50.8
Thread Structure	100	45.8	62.0	69.2
Accommodation	74.0	45.1	59.1	59.4
Thread+Accommodation	99.6	57.8	68.2	75.2
All	99.6	58.0	73.1	76.9
Best	100	58.5	73.0	77.6

Table 3: The effect, in F-score, of conversational structure in the ABCD corpus. Statistical significance is shown over majority$^\alpha$ and n-gram$^\beta$ baselines.

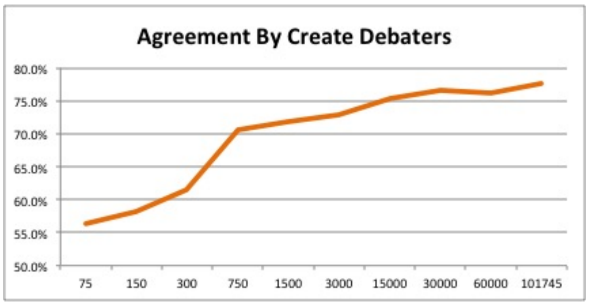


Figure 1: Average F-score as the ABCD training size increases when testing on the ABCD.

balanced datasets in the training set for the rest of the experiments. Table 3 shows how accommodation and meta-thread structure are very useful for detecting (dis)agreement. In fact, using n-grams, POS, LIWC, and lexical style features in just the response yields an average F-score of 50.8% whereas using POS, LIWC and lexical style in both the quote and response as well as sentence similarity yields a significant improvement of 8.6 points or 16.9% to an average F-score of 59.4%, indicating that conversational structure is very indicative of (dis)agreement. Using all features and the best features (computed using the development set) provide a statistically significant improvement at $\leq .05$ over both baselines. Our best results include all features except polarity with an average F-Score of 77.6%. Figure 1 shows that as the training size increases the results improve.

5.2 Internet Argument Corpus (IAC)

In contrast to prior work we detect (dis)agreement as a 3-way classification task: agreement, disagreement, none. Detecting (dis)agreement without including none pairs is unrealistic in a threaded discussion where the majority of posts will be neither agreement or disagreement. Additionally, we do not balance the test set as do Abbott et al (2011) and Walker et al (2013), but rather use all annotated posts to maintain a realistic agreement/disagreement/none ratio.

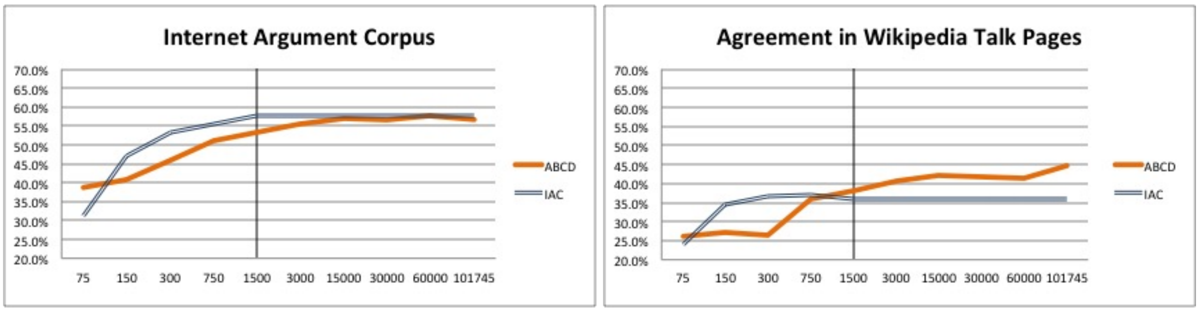
We experiment with using the small manually annotated in-domain IAC corpus and the large ABCD corpus. In contrast to the ABCD, we did not find accommodation to be significantly useful when training and testing using the IAC. We believe this is due to the large amount of none posts in the dataset (71.9%) where one does not expect accommodation to occur. However, in examining the average F-score for (dis)agreement, without none, we found that accommodation provides a 2.7 point or 11% improvement over only using features from the response. This improvement is masked by a 1.2 reduction in the none class where accommodation is not useful. The best IAC features differ depending on the training set and were computed using the IAC development set. Using the IAC training set, meta-thread structure, the LIWC, sentence similarity, and lexical style were most important. Using the ABCD corpus, the best features on the IAC development set were meta-thread structure, polarity, sentence similarity, the LIWC, and the negation/agreement/disagreement terms and question lexical features. We found it especially interesting that polarity and lexical features were useful on the ABCD while lexical style was useful for the IAC indicating clear variations in content across genres. Using the best features per corpus found from tuning towards the development sets (e.g. training and tuning on ABCD) provide a statistically significant improvement at $\leq .05$ over the n-gram baseline. The best and all (dis)agreement results provide a statistically significant improvement over the majority baseline. More detailed results are shown in Table 4. Finally, Figure 2a shows how increasing the size of the automatic ABCD training set improves the results compared to the manually annotated training set using the best feature set. Interestingly, there is little variation between the use of both datasets using the best features. We believe this is because thread structure is the most useful feature due to the large occurrence of none posts.

5.3 Agreement in Wikipedia Talk Pages (AWTP)

Our last set of experiments were performed on the AWTP which was annotated in-house. The advantage to the AWTP corpus is that the annotators were given the entire thread during annotation time, and annotated all (dis)agreement,

Features	IAC				ABCD			
	None	Agree	Disagree	Average	None	Agree	Disagree	Average
majority	85.1	0.0	0.0	28.4	85.1	0.0	0.0	28.4
n-gram	58.6	11.7	27.8	32.7	46.7	7.8	36.6	30.3
n-grams+POS+lexical-style+LIWC in R	54.1	12.0^{α}	29.7^{α}	31.9	43.9	13.6^{α}	30.1^{α}	29.2
Thread Structure	87.4^{β}	$25.3^{\alpha\beta}$	$50.0^{\alpha\beta}$	54.2^{β}	87.3^{β}	$26.4^{\alpha\beta}$	$53.8^{\alpha\beta}$	55.8^{β}
Accommodation	52.9	13.9^{α}	32.4^{α}	33.1	51.7	14.7^{α}	34.3^{α}	33.6
Thread+Accommodation	87.5^{β}	$26.5^{\alpha\beta}$	48.9^{β}	$54.3^{\alpha\beta}$	87.2^{β}	$28.0^{\alpha\beta}$	$55.5^{\alpha\beta}$	56.9^{β}
All	83.5^{β}	$28.8^{\alpha\beta}$	$50.4^{\alpha\beta}$	54.2^{β}	87.3^{β}	$27.0^{\alpha\beta}$	41.2^{α}	51.8
Best	87.4^{β}	$31.5^{\alpha\beta}$	$54.4^{\alpha\beta}$	57.8^{β}	87.3^{β}	$25.5^{\alpha\beta}$	$57.3^{\alpha\beta}$	56.7^{β}

Table 4: The effect, in F-score, of conversational structure in the IAC test set using the IAC and ABCD as training data. Results highlighted to indicate statistical significance over majority$^{\alpha}$ and n-gram$^{\beta}$ baselines.



(a)

(b)

Figure 2: Avg. F-score as the training size increases. The vertical line is the size of the IAC training set. The F-score succeeding the vertical line is the score at the peak size, included for contrast.

whether between Q-R pairs or not. In contrast, the IAC annotators were not provided with the entire thread. It was annotated only between Q-R pairs and even all Q-R pairs in a thread were not annotated. This means that each ATWP thread can be used for (dis)agreement detection in its entirety. Having fully annotated threads preserves the ratio of agreement/disagreement/none pairs better (the IAC has posts that are missing annotations).

We experiment with predicting (dis)agreement using the large naturally occurring ABCD dataset and the gold IAC dataset. Despite its advantage of gold labels, we found that using the ABCD as training consistently outperforms using the IAC as training on out-of-domain data, excluding when using just n-grams. In contrast to the other datasets, meta-thread structure and accommodation individually perform worse than using similar features found in the response alone. We believe this is because meta-thread structure is not strictly enforced in Wikipedia Talk Pages, providing an inaccurate representation of who is responding to who. Using all and the best features found during development (e.g. via training and tuning on ABCD) provide a statistically significant improvement at $\leq .05$ over the n-gram baseline for ABCD. The all and best (dis)agreement results provide a statistically significant improvement over the majority baseline for training on ABCD and IAC. More detailed results are shown in Table 3. We ran identical experiments to those performed on the IAC by increasing the training size of the ABCD corpus and IAC corpus to show their effects on the test set as shown in Figure 2b. The IAC dataset performs worse than using the ABCD dataset once the size of the ABCD training set exceeds the size of the IAC training set. This is further indication that automatic labeling is useful.

6 Discussion

We performed an error analysis to determine the kind of errors our system was making on 50 ABCD posts and 50 IAC posts from the development sets. In the ABCD posts we focused on agreement posts that were labeled incorrectly as our performance was worst in this class. Our analysis indicated that in most cases, 72.7% of the time, the error was due to the incorrect label; it should have been disagreement or none and not agreement as suggested by the side of the post. This is unsurprising as the label is determined using the side chosen by the post author. However, what is more surprising is that this was the common cause of error in the IAC

Features	IAC				ABCD			
	None	Agree	Disagree	Average	None	Agree	Disagree	Average
majority	87.2	0.0	0.0	29.1	87.2	0.0	0.0	29.1
n-gram	68.1	12.7	21.3	34.1	36.5	11.6	32	26.7
n-grams+POS+lexical-style+LIWC in R	64.1	12.1^{α}	22.7^{α}	33.0	54.0^{β}	$27.7^{\alpha\beta}$	$36.2^{\alpha\beta}$	39.3^{β}
Thread Structure	58.0	12.4^{α}	23.7^{α}	31.4	63.6^{β}	15.0^{α}	33.4^{α}	37.3
Accommodation	52.4	12.4^{α}	$30.7^{\alpha\beta}$	31.8	50.7^{β}	$17.5^{\alpha\beta}$	$40.1^{\alpha\beta}$	36.1^{β}
Thread+Accommodation	55.0	14.9^{α}	$37.2^{\alpha\beta}$	35.7	62.9^{β}	$21.3^{\alpha\beta}$	$52.2^{\alpha\beta}$	43.9^{β}
All	64.2	15.5^{α}	$36.4^{\alpha\beta}$	38.7	61.9^{β}	$25.8^{\alpha\beta}$	$43.5^{\alpha\beta}$	43.7^{β}
Best	59.3	14.4^{α}	$34.5^{\alpha\beta}$	36.1	63.6^{β}	$23.3^{\alpha\beta}$	$46.8^{\alpha\beta}$	44.4^{β}

Table 5: The effect, in F-score, of conversational structure in the AWTP test set using the IAC and ABCD as training data. Statistical significance is shown over majority$^{\alpha}$ and n-gram$^{\beta}$ baselines.

Dataset	Quote	Response	Description
ABCD	The same thing people use all words for; to convey information.	to convey information. Give me an example of when you are fully capable of saying this without offending someone.	The first sentence sounds like agreement but the second sentence is argumentative
IAC	Nowhere does it say, that she kept a gun in the bathroom emoticon_xkill	And nowhere does it say she went to her bedroom and retrieved a gun.	Agreement. It is an elaboration. Further context would help.

Table 6: Hard examples of (dis)agreement in ABCD and IAC

dataset as well, occurring 58.3% of the time. This is because the IAA using Cohen's κ among Amazon Turk workers for the IAC is low, averaging to .47 (Walker et al., 2012) across all topics. In addition, detecting agreement is hard as is evident in the incorrectly labeled examples in Table 6. Other errors were in posts where the agreement was a response, an elaboration, there was no (dis)agreement, and a conjunction indicating the post contained agreement and disagreement. To gain true insight into our model and gauge the impact of mislabeling, the labels of a small set of 60 threads (908 posts) were manually annotated to correct (dis)agreement errors resulting in 99 label changes. We allowed a post to be both agreement and disagreement and avoided changing labels to none as it is not a self-labeling option. This did not provide a significant change in F-score.

As is evident from our experiments, exploiting meta-thread structure and accommodation provide significant improvements. We also explored whether additional context would help by exploring the entire thread structure using general CRF. However, our experiments found that using CRF did not provide a significant improvement compared to using Maximum Entropy in the ABCD and AWTP corpora. This may be explained by our error analysis, which showed that in only 2/50 ABCD posts and 9/50 IAC posts further context beyond the Q-R posts would possibly help make it clearer whether it was agreement or disagreement.

7 Conclusion

We have shown that by exploiting conversational structure our system achieves significant improvements compared to using lexical features alone. In particular, our approach demonstrates the importance of meta-thread features, and accommodation between participants of an online discussion reflected in the semantic, syntactic and stylistic similarity between their posts. Furthermore, we use naturally occurring labels derived from Create Debate, to achieve improvements in detecting (dis)agreement compared to using smaller manually labeled datasets of the IAC and AWTP. The ABCD and AWTP datasets are available at `www.cs.columbia.edu/~sara/data.php`. This is promising for domains where no annotated data exists; the dataset can be used to avoid performing a time consuming and costly annotation effort. In the future we would like to take further advantage of existing manually annotated datasets by using domain adaptation to combine the datasets. In addition, our error analysis indicated that a significant amount of errors were due to mislabeling. We would like to explore improving results by using the system to automatically correct such errors in held-out training data and then using the corrected data to retrain the model.

Acknowledgments

This paper is based upon work supported by the DARPA DEFT Program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. We thank several anonymous reviewers for their constructive feedback.

References

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. 2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on LSM*, LSM '11, pages 2–11, Portland, Oregon. ACL.

Amjad Abu-Jbara, Mona Diab, Pradeep Dasigi, and Dragomir Radev. 2012. Subgroup detection in ideological discussions. In *Proceedings of the 50th Annual Meeting of the ACL*, ACL '12, pages 399–409, Jeju Island, Korea. ACL.

Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 24–32, 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 Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

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 LSM*, LSM '11, pages 48–57, Portland, Oregon. ACL.

Or Biran, Sara Rosenthal, Jacob Andreas, Kathleen McKeown, and Owen Rambow. 2012. Detecting influencers in written online conversations. In *Proceedings of the 2nd Workshop on LSM*, LSM '12, pages 37–45, Montreal, Canada. ACL.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on WWW*, WWW '12, pages 699–708, NYC, USA. ACM.

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 43rd Annual Meeting of the ACL*, page 669, Barcelona, Spain. ACL.

Sebastian Germesin and Theresa Wilson. 2009. Agreement detection in multiparty conversation. In *ICMI*, pages 7–14. ACM.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. In *Contexts of Accommodation*, pages 1–68. Cambridge University Press. Cambridge Books Online.

Weiwei Guo and Mona Diab. 2012. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the ACL*, pages 864–872, Jeju Island, Korea, July. ACL.

Sangyun Hahn, Richard Ladner, and Mari Ostendorf. 2006. Agreement/disagreement classification: Exploiting unlabeled data using contrast classifiers. In *Proceedings of HLT-NAACL*, pages 53–56, NYC, USA, June. ACL.

Ahmed Hassan, Amjad Abu-Jbara, and Dragomir Radev. 2012. Detecting subgroups in online discussions by modeling positive and negative relations among participants. In *Proceedings of the Joint Conference on EMNLP and CoNLL*, EMNLP-CoNLL '12, pages 59–70, Jeju Island, Korea. ACL.

Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of HLT-NAACL*, Edmonton, Canada. ACL.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The icsi meeting corpus. pages 364–367.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on?: Identifying perspectives at the document and sentence levels. In *Proceedings of the 10th Conference on CoNLL*, CoNLL-X '06, pages 109–116, NYC, USA. ACL.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet.

I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. 2005. The ami meeting corpus. In *Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research.*

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, Metz, France, August. ACL.

Arjun Mukherjee and Bing Liu. 2012. Analysis of linguistic style accommodation in online debates. In *Proceedings of COLING 2012*, pages 1831–1846, Mumbai, India, December. The COLING 2012 Organizing Committee.

Arjun Mukherjee and Bing Liu. 2013. Discovering user interactions in ideological discussions. In *Proceedings of the 51st Annual Meeting of the ACL*, pages 671–681, Sofia, Bulgaria, August. ACL.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses : An Introduction*. Wiley-Interscience, April.

Bernd Opitz and Cecilia Zirn. 2013. Bootstrapping an unsupervised approach for classifying agreement and disagreement. volume 85. Linköping Univ. Electronic Press.

Sara Rosenthal, Apoorv Agarwal, and Kathy McKeown. 2014. Columbia nlp: Sentiment detection of sentences and subjective phrases in social media. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, Dublin, Ireland.

Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the ACL*, pages 455–465, Sofia, Bulgaria, August. ACL.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods.

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 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 327–335.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, NAACL '03, pages 173–180, Edmonton, Canada. ACL.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eight International Conference on LREC (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Lu Wang and Claire Cardie. 2014. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 97–106, Baltimore, Maryland, June. Association for Computational Linguistics.

Wen Wang, Kristin Precoda, Colleen Richey, and Geoffrey Raymond. 2011a. Identifying agreement/disagreement in conversational speech: A cross-lingual study. In *INTERSPEECH*, pages 3093–3096. ISCA.

Wen Wang, Sibel Yaman, Kristin Precoda, Colleen Richey, and Geoffrey Raymond. 2011b. Detection of agreement and disagreement in broadcast conversations. In *Proceedings of the 49th Annual Meeting of the ACL*, pages 374–378. ACL.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 947–953, Stroudsburg, PA, USA. Association for Computational Linguistics.

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 WASSA*, WASSA '12, pages 61–69, Jeju Island, Korea. ACL.