Detecting Disagreement in Conversations using Pseudo-Monologic Rhetorical Structure

Kelsey Allen **Giuseppe Carenini** **Raymond T. Ng**
Department of Computer Science, University of British Columbia
Vancouver, BC, V6T 1Z4, Canada
{kelseyra, carenini, rng}@cs.ubc.ca

Abstract

Casual online forums such as Reddit, Slashdot and Digg, are continuing to increase in popularity as a means of communication. Detecting disagreement in this domain is a considerable challenge. Many topics are unique to the conversation on the forum, and the appearance of disagreement may be much more subtle than on political blogs or social media sites such as twitter. In this analysis we present a crowd-sourced annotated corpus for topic level disagreement detection in Slashdot, showing that disagreement detection in this domain is difficult even for humans. We then proceed to show that a new set of features determined from the rhetorical structure of the conversation significantly improves the performance on disagreement detection over a baseline consisting of unigram/bigram features, discourse markers, structural features and meta-post features.

1 Introduction

How does disagreement arise in conversation? Being able to detect agreement and disagreement has a range of applications. For an online educator, dissent over a newly introduced topic may alert the teacher to fundamental misconceptions about the material. For a business, understanding disputes over features of a product may be helpful in future design iterations. By better understanding how debate arises and propagates in a conversation, we may also gain insight into how authors' opinions on a topic can be influenced over time.

The long term goal of our research is to lay the foundations for understanding argumentative structure in conversations, which could be applied to NLP tasks such as summarization, information retrieval, and text visualization. Argumentative structure theory has been thoroughly studied in both psychology and rhetoric, with negation and discourse markers, as well as hedging and dispreferred responses, being known to be indicative of argument (Horn, 1989; Brown and Levinson, 1987). As a starting point, in this paper we focus on the detection of disagreement in casual conversations between users. This requires a generalized approach that can accurately identify disagreement in topics ranging from something as mundane as whether GPS stands for galactic positioning system or global positioning system, to more ideological debates about distrust in science.

Motivated by the widespread consensus in both computational and theoretical linguistics on the utility of discourse markers for signalling pragmatic functions such as disagreement and personal opinions (Webber and Prasad, 2008; Abbott et al., 2011; J. E. Fox-Tree, 2010), we introduce a new set of features based on the Discourse Tree (DT) of a conversational text. Discourse Trees were formalized by Mann and Thompson (1988) as part of their Rhetorical Structure Theory (RST) to represent the structure of discourse. Although this theory is for monologic discourse, we propose to treat conversational dialogue as a collection of linked monologues, and subsequently build a *relation graph* describing both rhetorical connections within user posts, as well as between different users. Features obtained from this graph offer significant improvements on disagreement detection over a baseline consisting of meta-post features, lexical features, discourse markers and conversational features. Not only do these features improve disagreement detection, but the discovered importance of relations known to be theoretically relevant to disagreement detection, such as COMPARISON (Horn, 1989), suggest that this approach may be capturing the essential aspects of the conversational argumentative structure.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1169–1180,
October 25-29, 2014, Doha, Qatar.

As a second contribution of this work, we provide a new dataset consisting of 95 topics annotated for disagreement. Unlike the publicly available *ARGUE* corpus based on the online debate forum 4forums.com (Abbott et al., 2011), our corpus is based on Slashdot, which is a general purpose forum not targeted to debates. Therefore, we expect that detecting disagreement may be a more difficult task in our new corpus, as certain topics (like discussing GPS systems) may be targeted towards objective information sharing without any participants expressing opinions or stances. Because of this, our corpus represents an excellent testbed to examine methods for more subtle disagreement detection, as well as the major differences between news-style and argument-style dialogue.

2 Related Work

In the past decade, there have been a number of computational approaches developed for the task of disagreement and controversy detection, particularly in synchronous conversations such as meetings (Somasundaran et al., 2007; Raaijmakers et al., 2008) and in monologic corpora such as news collections (Awadallah et al., 2012) and reviews (Popescu et al., 2005; Mukherjee and Liu, 2012).

In the domain of synchronous conversations, prosodic features such as duration, speech rate and pause have been used for spoken dialogue (Wang et al., 2011; Galley et al., 2004). Galley et al. (2004) found that local features, such as lexical and structural features, as well as global contextual features, were particularly useful for identifying agreement/disagreement in the ICSI meeting corpus. Germesin and Wilson (2009) also showed accuracies of 98% in detecting agreement in the AMI corpus using lexical, subjectivity and dialogue act features. However, they note that their system could not classify disagreement accurately due to the small number of training examples in this category. Somasundaran et al. additionally show that dialogue act features complement lexical features in the AMI meeting corpus (Somasundaran et al., 2007). These observations are taken into account with our feature choices, and we use contextual, discourse and lexical features in our analysis.

In the monologic domain, Conrad et al. (2012) recently found rhetorical relations to be useful for argument labelling and detection in articles on the topic of healthcare. Additionally, discourse markers and sentiment features have been found to assist with disagreement detection in collections of news documents on a particular topic, as well as reviews (Choi et al., 2010; Awadallah et al., 2012; Popescu et al., 2005).

In the asynchronous domain, there has been recent work in disagreement detection, especially as it pertains to stance identification. Content based features, including sentiment, duration, and discourse markers have been used for this task (Yin et al., 2012; Somasundaran and Wiebe, 2009; Somasundaran and Wiebe, 2010). The structure of a conversation has also been used, although these approaches have focused on simple rules for disagreement identification (Murakami and Raymond, 2010), or have assumed that adjacent posts always disagree (Agrawal et al., 2003). More recent work has focused on identifying users' attitudes towards each other (Hassan et al., 2010), influential users and posts (Nguyen et al., 2014), as well as identifying subgroups of users who share viewpoints (Abu-Jbara et al., 2010). In Slashdot, the h-index of a discussion has been used to rank articles according to controversiality, although no quantitative evaluation of this approach has been given, and, unlike in our analysis, they did not consider any other features (Gomez et al., 2008). Content based features such as polarity and cosine similarity have also been used to study influence, controversy and opinion changes on microblogging sites such as Twitter (Lin et al., 2013; Popescu and Pennacchiotti, 2010).

The simplified task of detecting disagreement between just two users (either a question/response pair (Abbott et al., 2011) or two adjacent paragraphs (Misra and Walker, 2013)) has also been recently approached on the *ARGUE* corpus. Abbott et al. (2011) use discourse markers, generalized dependency features, punctuation and structural features, while Misra and Walker (2013) focus on n-grams indicative of denial, hedging and agreement, as well as cue words and punctuation. Most similar to our work is that by Mishne and Glance (2006). They performed a general analysis of weblog comments, using punctuation, quotes, lexicon counts, subjectivity, polarity and referrals to detect disputative and non-disputative comments. Referrals and questions, as well as polarity measures in the first section of the post, were found to be most useful. However, their analysis did not

Type	Num	P/A	S/P	W/P	Num Authors	W/S	TBP	TT	TP	Length
Disagreement - C	19.00	1.02	3.21	65.86	15.84	20.47	4.60	50.90	16.21	49.11
Disagreement - NC	27.00	1.00	3.08	59.80	14.07	19.33	3.89	42.29	14.11	42.26
No disagreement - NC	28.00	1.03	2.85	57.25	10.29	19.94	6.83	50.12	10.50	28.00
No disagreement - C	21.00	1.00	3.69	69.66	6.29	20.22	6.14	18.22	6.29	19.81

Table 1: Characteristics of the four categories determined from the crowd-sourced annotation. All values except for the number of topics in the category are given as the average score per topic across all topics in that category. **Key: C and NC:** Confident ($score \geq 0.75$) and Not confident ($score < 0.75$), **Num:** Number of topics in category, **P/A:** Posts per author, **S/P:** Sentences per post, **W/P:** Words per post, **Num Authors:** Number of authors, **W/S:** Words per sentence, **TBP:** Time between posts (minutes), **TT:** Total time in minutes, **TP:** Total posts, and **Length:** Length of topic in sentences

take into account many features that have been subsequently shown to be relevant, such as discourse markers and conversational structure, and was hampered by a severe imbalance in the test set (with very few disputative comments).

Our method takes advantage of insights from many of these previous studies, focusing on discussion thread structure as well as text based features to form our basic feature set. It is unlike Mishne and Glance's work in that we incorporate several new features, have a balanced testing and training set, and only use comments from one type of online blog. Furthermore, it is a very different task from those so far performed on the *ARGUE* corpus, as we consider topics discussed by more than two users. We aim to compare our features to those found to be previously useful in these related tasks, and expect similar feature sets to be useful for the task of disagreement detection in this new corpus.

3 Corpus

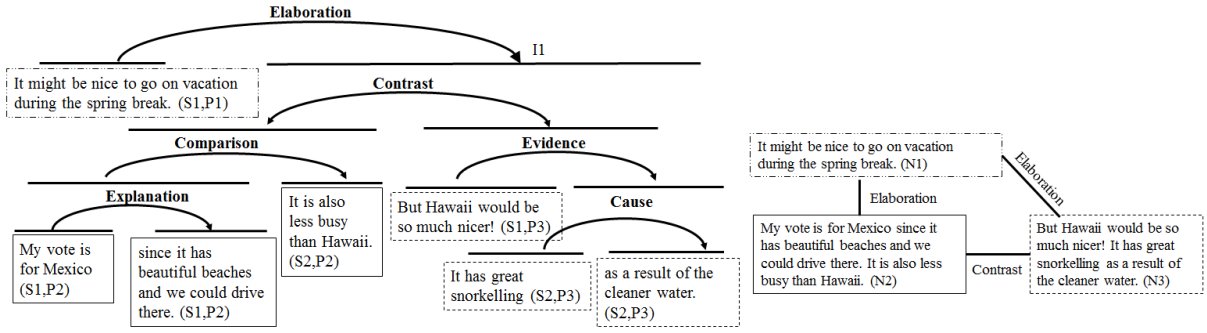
The corpus stems from the online forum Slashdot.[1] Slashdot is a casual internet forum, including sections for users to ask questions, post articles, and review books and games. For the task of disagreement detection, we focus our analysis on the section of the site where users can post articles, and then comment either on the article or respond to other users' posts. This results in a tree-like dialogue structure for which the posted article is the root, and branches correspond to threads of comments. Each comment has a timestamp at the minute resolution as well as author information (although it is possible to post on the forum anonymously). Additionally, other users can give different posts scores (in the range -1 to 5) as well as categorizing posts under "funny", "interesting", "informative", "insightful", "flamebait", "off topic", or "troll". This user moderation, as well as the formalized reply-to structure between comments, makes Slashdot attractive over other internet forums as it allows for high-quality and structured conversations.

In a previous study, Joty et al. (2013) selected 20 articles and their associated comments to be annotated for topic segmentation boundaries and labels by an expert Slashdot contributor. They define a topic as a subset of the utterances in a conversation, while a topic label describes what the given topic is about (e.g., Physics in videogames). Of the 98 annotated topics from their dataset, we filtered out those with only one contributing user, for a total of 95 topics. Next, we developed a Human Intelligence Task (HIT) using the crowdsourcing platform Crowdflower.[2] The objective of this task was to both develop a corpus for testing our disagreement detection system, as well as to investigate how easily human annotators can detect disagreement in casual online forums. For training, users were shown 3 sample topics, labelling them as containing disagreement or not. In each round, annotators were shown 5 topics, with a set of radio buttons for participants to choose "Yes", "No", or "Not sure" in response to asking whether or not the users in the conversation disagree on the topic. In order to limit the number of spam responses, users were shown test questions, which consisted of topics where there was obvious disagreement, as well as topics where there was obviously no disagreement (either agreement, or more news-style information sharing). We required that users correctly identify 4 of these test topics before they were allowed to continue with the task. Users were also shown test questions throughout the task, which, if answered incorrectly, would reduce the amount of money they received for the task, and ultimately disqualify them.

For each topic, five different judgements were obtained. We consider the trust of each partici-

[1] www.slashdot.org

[2] www.Crowdflower.com



(a) Discourse tree **(b)** Relation graph

Figure 1: Discourse tree (left) with extracted relation graph (right) for a sample conversation involving three users with three different posts P1, P2 and P3. N1, N2 and N3 are the corresponding nodes in the relation graph.

pant as the fraction of test questions which they answered correctly. Then, each topic is assigned a score according to a weighted average of the responses, with the weight being the trust of each participant:

$$score = A \sum_{users} (\frac{test_{correct}}{test_{total}})_{user_i} \times (0, 0.5, 1) \quad (1)$$

where 0, 0.5 and 1 represent the answers "No", "Not sure" and "Yes" to the question of disagreement existence, and A is a normalization factor. If the score is less than 0.5, its confidence would be $1 - score$ towards "No disagreement", whereas greater than 0.5 would be a confidence of $score$ towards "Disagreement". The average confidence score across all topics was 0.73. Our corpus consists of 49 topics without disagreement and 46 topics with disagreement. Interestingly, 22 topics had confidence scores below 55%, which suggests that subtle disagreement detection is a subjective and difficult task. Further statistics for the developed corpus are given in Table 1.

4 Features for Disagreement Detection

The features we use in our experiments combine information from conversational structure, rhetorical relations, sentiment features, n-gram models, Slashdot meta-features, structural features, and lexicon features.

4.1 Rhetorical Relation Graphs

Discourse markers have been found to aid in argument and disagreement detection, and for tasks such as stance identification (Abbott et al., 2011; Misra and Walker, 2013; Somasundaran and Wiebe, 2009). We aim to improve over discourse markers by capturing the underlying discourse structure of the conversation in terms of discourse relations.

In Rhetorical Structure Theory, Discourse trees are a hierarchical representation of document structure for monologues (Mann and Thompson, 1988). At the lowest level, Elementary Discourse Units (EDUs) are connected by discourse relations (such as ELABORATION and COMPARISON), which in turn form larger discourse units that are also linked by these relations. Computational systems (discourse parsers) have been recently developed to automatically generate a discourse tree for a given monologue (Joty et al., 2013). Although theoretically the rhetorical relations expected in dialogues are different from those in monologues (Stent and Allen, 2000), no sophisticated computational tools exist yet for detecting these relations reliably. The core idea of this work is that some useful (although noisy) information about the discourse structure of a conversation can be obtained by applying state-of-the-art document level discourse parsing to parts of the conversation.

More specifically, posts on a particular topic are concatenated according to their temporal order. This pseudo-monologic document is then fed to a publicly available document level discourse parser (Joty et al., 2013). A discourse tree such as that seen in Figure 1a is output by the parser. Then, we extract the novel *relation graph* (Figure 1b) from the discourse tree. In this graph, each node (N1, N2, N3) corresponds to a post (P1, P2, P3) and links aim to capture the argumentative structure. There are three cases when a link is added between two nodes in the relation graph. Firstly, links existing between two posts directly, such as the COMPARISON relation between P2 and P3, are added between the corresponding nodes in the re-

lation graph (N2 and N3). Secondly, links existing between fractions of posts in the discourse tree are added to the relation graph (e.g. if (S2,P2) was connected to (S1,P3) directly, N2 and N3 would have an additional link with that label). Finally, when posts are connected through internal nodes (such as P1 and I1 in Figure 1a), a labelled link is added for each post in the internal node to the relation graph (N1->N2 and N1->N3 in Figure 1b).

This relation graph allows for the extraction of many features that may reflect argumentative structure, such as the *number of connections*, the frequency of each rhetorical relation in the graph per post (*diff per post*), and the frequency as a percentage of all rhetorical relations (*diff percentage*). For example, COMPARISON relations are known to indicate disagreement (Horn, 1989), so we expect higher frequencies of this relation if the conversation contains argument. Features from the discourse tree such as the *average depth* of each rhetorical relation are also added to reflect the cohesiveness of conversation. Moreover, features combining the graph and tree representations, such as the ratio of the frequency of a rhetorical relation occurring between different posts to the average depth ($\frac{\text{CONTRAST between different posts}}{\text{Average depth of CONTRAST}}$), called *avg ratio* are implemented. These reflect the hypothesis that relations connecting larger chunks of text (or whole posts) may be more important than those connecting sentences or only partial posts.

Finally, the sub-trees corresponding to individual posts are used to extract the average frequency of rhetorical relations within a post (*same per post*) and the average frequency of a rhetorical relation with respect to other rhetorical relations in the post (*same percentage*). A measure of how often a rhetorical relation connects different users compared to how often it connects discourse units in the same post (*same to diff*), is also added. These capture the intuition that certain rhetorical relations such as CAUSE, EVIDENCE and EXPLANATION are expected to appear more within a post if users are trying to support their perspective in an argument. In total, there are 18 (rhetorical relations)×7 (*avg ratio, avg depth, same per post, same percentage, diff percentage, diff per post, same to diff*) + 1 (number of connections) = 127 features.

4.2 Discourse Markers

Motivated by previous work, we include a frequency count of 17 discourse markers which were found to be the most common across the *ARGUE* corpus (Abbott et al., 2011). Furthermore, we hypothesize that individual discourse markers might have low frequency counts in the text. Therefore, we also include an aggregated count of all 17 discourse markers in each fifth of the posts in a topic (e.g. the count of all 17 discourse markers in the first fifth of every post in the topic). Altogether, there are 5 aggregated discourse marker features in addition to the 17 frequency count features.

4.3 Sentiment Features

Sentiment polarity features have been shown to be useful in argument detection (Mishne and Glance, 2006). For this work, we use four sentiment scoring categories: the *variance*, *average score*, *number of negative sentences*, and *controversiality score* (Carenini and Cheung, 2008) of sentences in a post. These are determined using SoCAL (Taboada et al., 2011), which gives each sentence a polarity score and has been shown to work well on user-generated content.

Overall, we have two main classes of sentiment features. The first type splits all the posts in a topic into 4 sections corresponding to the sentences in each quarter of the post. The sentiment scores described above are then applied to each section of the posts (e.g. one feature is the number of negative sentences in the first quarter of each post). As a separate feature, we also include the scores on just the first sentence, as Mishne and Glance (2006) previously found this to be beneficial. This gives a total of 4×5 = 20 features. We refer to this set as "sentiment".

Motivated by the rhetorical features, our second main class of sentiment features aims to identify "more important" posts for argument detection by applying the four categories of sentiment scores to only those posts connected by each of our 18 rhetorical relations. This is done for both posts with an inner rhetorical connection (identified by the sub-tree for that post), as well as for posts connected by a rhetorical relation in the relation graph. This results in a total of (4 sentiment categories)×(2 (different + same post connections))×(18 rhetorical relations) = 144 features. This set is referred to as "RhetSent".

4.4 Fragment Quotation Graphs

As previously noted in (Murakami and Raymond, 2010; Gomez et al., 2008), the structure of discussion threads can aid in disagreement detection. In online, threaded conversations, the standard approach to extracting conversational structure is through reply-to relations usually present in online forums. However, if users strongly disagree on a topic (or sub-topic), they may choose to quote a specific paragraph (defined as a fragment) of a previous post in their reply. Being able to determine which specific fragments are linked by relations may then be useful for more targeted content-based features, helping to reduce noise. In order to address this, we use the Fragment Quotation Graph (FQG), an approach previously developed by Carenini et al. (2007) for dialogue act modelling and topic segmentation (Joty et al., 2011; Joty et al., 2013).

For our analysis, the FQG is found over the entire Slashdot article. We then select the subgraph corresponding to those fragments in the target topic. From the fragment quotation sub-graph, we are then able to extract features for disagreement detection such as the *number of connections*, *total number of fragments*, and the *average path length* between nodes which we hypothesize to be useful. We additionally extract the *h-index* (Gomez et al., 2008) and *average branching ratio per fragment* of the topic from the simpler reply-to conversational structure. In total, there are 8 FQG features.

4.5 N-gram models

As noted previously (Somasundaran and Wiebe, 2010; Thomas et al., 2006), it is often difficult to outperform a unigram/bigram model in the task of disagreement and argument detection. In this analysis, because of the very small number of samples, we do not consider dependency or part-of-speech features, but do make a comparison with a filtered unigram/bigram model. In the filtering, we remove stop words and any words that occur in fewer than three topics. This helps to prevent topic specific words from being selected, and limits the number of possible matches slightly. Additionally, we use a lexicon of bias-words (Recasens et al., 2013) to extract a bias-word frequency score over all posts in the topic as a separate feature.

4.6 Structural Features

Length features have been well documented in the literature to provide useful information about whether or not arguments exist, especially in online conversations that may be more informative than subjective (Biyani et al., 2014; Yin et al., 2012). In this work, length features include the *length of the post in sentences*, the *average number of words per sentence*, the *average number of sentences per post*, the *number of contributing authors*, the *rate of posting*, and the *total amount of time* of the conversation. This results in a total of 9 features.

4.7 Punctuation

Like many other features already described, frequency counts of '?','!',"","", and '.' are found for each fifth of the post (the first fifth, second fifth, etc.). These counts are then aggregated across all posts for a total of $5 \times 5 = 25$ features.

4.8 Referrals

Referrals have been found to help with the detection of disagreement (Mishne and Glance, 2006), especially with respect to other authors. Since there are no direct referrals to previous authors in this corpus, references to variations of "you", "they", "us", "I", and "he/she" in each fifth of the post are included instead, for a total of $5 \times 5 = 25$ features.

4.9 Meta-Post Features

Slashdot allows users to rank others' posts with the equivalent of a "like" button, changing the "score" of the post (to a maximum of 5). They are also encouraged to tag posts as either "Interesting", "Informative", "Insightful", "Flamebait", "Troll", "Off-topic" or "Funny". Frequency counts of these tags as a percentage of the total number of comments are included as features, as well as the overall fraction of posts which were tagged with any category. The *average score* across the topic, as well as the *number of comments with a score of 4 or 5*, are also added. These are expected to be informative features, since controversial topics may encourage more up and down-voting on specific posts, and generally more user involvement. This results in 9 meta-post features.

Feature Set	Random Forest					SVM				
	P	R	F1	A	ROC-AUC	P	R	F1	A	ROC-AUC
N-grams	0.71	0.57	0.63	0.67	0.69	0.52	0.60	0.56	0.53	0.53
Basic	0.69	0.67	0.68	0.69	0.73	0.62	0.62	0.62	0.63	0.67
Basic+N-grams	0.73	0.66	0.69	0.70	0.73	0.57	0.65	0.60	0.59	0.61
Basic+FQG	0.69	0.66	0.67	0.69	0.71	0.64	0.63	0.63	0.65	0.70
Basic+Sentiment	0.68	0.65	0.66	0.68	0.73	0.61	0.59	0.60	0.62	0.67
Basic+RhetStruct	0.71	0.70	0.70	0.71	0.73	0.73	0.70	**0.71**	**0.73**	**0.78**
Basic+RhetStruct+FQG	0.69	0.69	0.69	0.70	0.73	0.74	0.74	**0.74**	**0.75**	**0.80**
Basic+RhetAll	0.72	0.73	**0.73**	**0.73**	**0.75**	0.76	0.76	**0.76**	**0.77**	**0.79**
RhetStructOnly	0.69	0.72	0.71	0.71	0.72	0.76	0.72	**0.74**	**0.75**	**0.79**
RhetAllOnly	0.69	0.74	0.71	0.71	0.73	0.75	0.72	**0.73**	**0.75**	**0.78**
All	0.71	0.72	**0.71**	**0.72**	0.74	0.74	0.77	**0.75**	**0.76**	**0.77**

Table 2: Basic: Meta-post, all structural, bias words, discourse markers, referrals, punctuation **RhetAll:** Structural and sentiment based rhetorical features **All:** Basic, all rhetorical, sentiment and FQG features. The N-gram models include unigrams and bi-grams. All feature sets in the bottom part of the table include rhetorical reatures.

5 Experiments

Experiments were all performed using the Weka machine learning toolkit. Two different types of experiments were conducted - one using all annotated topics in a binary classification of containing disagreement or not, and one using only those topics with confidence scores greater than 0.75 (corresponding to the more certain cases). All results were obtained by performing 10 fold cross-validation on a balanced test set. Additionally, in-fold cross-validation was performed to determine the optimal number of features to use for each feature set. Since this is done in-fold, a paired t-test is still a valid comparison of different feature sets to determine significant differences in F-score and accuracy.

5.1 Classifiers

Two classifiers were used for this task: Random Forest and SVM. Random Forest was selected because of its ability to avoid over-fitting data despite large numbers of features for relatively few samples. For all runs, 100 trees were generated in the Random Forest, with the number of features to use being determined by in-fold optimization on the F-score. For the SVM classifier, we use the normalized poly-vector kernel with a maximum exponent of 2 (the lowest possible), and a C parameter of 1.0 (Weka's default value). This was chosen to avoid over-fitting our data. We additionally use a supervised in-fold feature selection algorithm, Chi-Squared, to limit over-fitting in the SVM. The number of features to be used is also optimized using in-fold cross-validation on the F-score. Both the SVM classifier and the Random Forest classifier were tested on the same training/testing fold pairs, with a total of 10 iterations.

6 Results

The results of the experiments are shown in Tables 2 and 3. In order to compare to previous analyses, unigram and bigram features are shown, as well as a combination of the basic features with the n-grams. When performing the experiments, we noticed that the n-gram features were hurting the performance of the classifiers when included with most of our other feature sets (or not changing results significantly), and therefore those results are not shown here. As seen in the table,

Feature Set	Random Forest					SVM				
	P	R	F1	A	ROC-AUC	P	R	F1	A	ROC-AUC
N-grams	0.70	0.70	0.70	0.71	0.77	0.63	0.70	0.66	0.63	0.66
Basic	0.74	0.69	0.72	0.72	0.77	0.73	0.70	0.72	0.71	0.78
Basic+FQG	0.72	0.67	0.69	0.69	0.76	0.73	0.63	0.68	0.69	0.76
Basic+Sentiment	0.71	0.65	0.68	0.68	0.76	0.73	0.67	0.70	0.70	0.76
Basic+RhetStruct	0.79	0.75	**0.77**	**0.77**	**0.78**	0.79	0.67	**0.72**	**0.74**	**0.79**
Basic+RhetStruct+FQG	0.76	0.71	0.73	0.73	0.77	0.74	0.64	0.69	0.70	0.78
Basic+RhetAll	0.77	0.75	**0.76**	**0.76**	**0.78**	0.72	0.69	0.71	0.70	0.76
RhetStructOnly	0.75	0.71	0.73	0.73	0.75	0.76	0.63	0.69	0.70	0.76
RhetAllOnly	0.73	0.76	0.74	0.73	0.76	0.67	0.62	0.65	0.65	0.67
All	0.73	0.69	0.71	0.70	0.76	0.71	0.70	0.70	0.69	0.74

Table 3: Precision, recall, F1, accuracy and ROC-AUC scores for the simpler task of identifying the cases deemed "high confidence" in the crowd-sourcing task.

the best performing feature sets are those that include rhetorical features under the SVM+χ^2 classifier. In fact, these feature sets perform significantly better than a unigram/bigram baseline according to a paired t-test between the best classifiers for each set ($p < 0.0001$). The inclusion of rhetorical structure also significantly outperforms the "basic" and "basic+N-grams" feature baselines (which includes discourse markers, referrals, punctuation, bias word counts and structural features) with respect to both the F-score and accuracy ($p < 0.02$ for all feature sets with rhetorical features). Overall, the feature sets "Basic+RhetAll" and "All" under the SVM+χ^2 classifier perform best. This performance is also better than previously reported results for the *ARGUE* Corpus (Abbott et al., 2011), even though the basic and unigram/bigram features perform similarly to that reported in previous analyses.

As an additional check, we also conduct experiments on the "high confidence" data (those topics with a confidence score greater than 0.75). These results are shown in Table 3. Clearly the basic features perform better on this subset of the samples, although the addition of rhetorical structure still provides significant improvement ($p < 0.001$). Overall, this suggests that the rhetorical, sentiment and FQG features help more when the disagreement is more subtle.

Structural	*Length of topic in sentences, total number of authors, quotes in first sentence, quotes in second sentence, questions in first sentence, questions in second sentence, referrals to you and they in first half of post*
Meta	*Number of comments with labels, number of comments labelled 'Flamebait', number of comments with scores of 4 or 5*
FQG	*Number of connections, Number of fragments, Maximum number of links from one node*
RhetStruct	COMPARISON *(same to diff, diff per post, diff percentage, avg ratio, same per post, same percentage),* EXPLANATION *(avg ratio, diff per post),* JOINT *(diff percentage),* ELABORATION *(average depth)*
Discourse Markers	*Aggregated first sentence, Aggregated middle, 'and', 'oh', 'but' frequency counts*
N-grams	*'don't ?', 'plus', 'private', 'anti', 'hey', 'present', 'making', 'developers'*
RhetSent	ELABORATION *(variance in same post),* ATTRIBUTION *(variance in same post),* CONTRAST *(range in same post)*
Sentiment	*Range of sentiment scores in first sentence of all posts, range of sentiment scores over all posts*

Table 4: Features found to be commonly selected over different iterations of the classifiers

7 Analysis and Discussion

In order to examine the quality of our features, we report on the rhetorical features selected, and show that these are reasonable and in many cases, theoretically motivated. Furthermore, we check whether the commonly selected features in each of our feature categories are similar to those found to be useful over the *ARGUE* corpus, as well as within other argument detection tasks in online forums.

The rhetorical features that are consistently selected are very well motivated in the context of argument detection. From the rhetorical structural features, we find COMPARISON relation features to be most commonly selected across all rhetorical feature sets. Other highly ranked features include the proportion of JOINT relations linking different authors, EXPLANATION relations between different authors, and the average depth of ELABORATION relations.

The COMPARISON relations are expected to indicate disagreement as motivated by theoretical considerations (Horn, 1989). The importance of other rhetorical relation features can also be explained by examining conversations in which they appear. In particular, EXPLANATION relations often link authors who share viewpoints in a debate, especially when one author is trying to support the claim of another. The JOINT relations are also very well motivated. In the extreme case, conversations with a very high number of JOINT relations between different users are usually news based. The high proportion of these relations indicates that many users have added information to the conversation about a specific item, such as adding new suggested videogame features to an ongoing list. Fewer JOINT relations seem to indicate disagreement, especially when found in conjunction with COMPARISON relations between different users. This appears to generally indicate that users are taking sides in a debate, and commenting specifically on evidence which supports their viewpoint.

The average depth of ELABORATION relations reveals how deep the perceived connections are between users in a conversation over time. Deeper ELABORATION connections seem to indicate that the conversation is more cohesive. Alone, this does not signify disagreement/agreement but does seem to signify argument-style over news-style dialogues. This is particularly helpful for differentiating between articles with many COMPARISON relations, as COMPARISON may be present in both news-style dialogues (e.g. comparing citation styles) as well as argument style dialogues (e.g. arguing over which of two operating systems is superior).

For the combined sentiment and rhetorical relations, range and variance in ELABORATION, CONTRAST and ATTRIBUTION within the same post are found to be the most informative features. Additionally, neither ATTRIBUTION nor CONTRAST are useful features when only their structural information is considered. In the case of ATTRIBUTION, we hypothesize that the added sentiment score within the post differentiates between a neutral attribution (which would not signify disagreement) and a negative or positive attribution (which may signify disagreement). For CONTRAST, the added sentiment helps to distinguish between responses such as "*We will be trying the Microsoft software. We won't, however, be able to test the Apple equivalent.*" and "*We will be trying the Microsoft software. We won't, however, be trying the inferior Apple equivalent.*" where the second example more likely signals, or even provokes, disagreement.

Outside of the rhetorical features, the discourse markers which are found to be the most useful in our experiments agree with those found in the *ARGUE* corpus (Abbott et al., 2011). Namely, 'oh', 'but', 'because' and 'and' are discovered to be the most informative features. We also find the aggregated discourse marker frequency count in the first part of each post to be useful.

Previous analysis on Slashdot as a social network (Gomez et al., 2008) suggests that the h-index of the conversation is relevant for detecting controversy in a posted article. We include the h-index as part of the Fragment Quotation Graph feature set, but surprisingly do not find this to be a useful feature. This may be due to our corpus involving relatively shallow conversational trees - the maximum h-index across all topics is three.

Comparing to Mishne and Glance's work, we also find quotations, questions and sentiment range near the beginning of a post to be very informative features. These are often selected across all feature sets which include the "basic" set.

The topics most often misclassified across all feature sets are those with relatively few sentences. In these cases, the rhetorical structure is not very well defined, and there is much less content available for detecting quotes, punctuation and referrals. Additionally, the feature sets which only use rhetorical and sentiment features consistently misclassify the same set of conversations (those that have lower quality discourse trees with few connections). When combined with the "basic" feature set, these errors are mitigated, and the topics which the "basic" features miss are picked up by the rhetorical features. This leads to the best overall accuracy and F-score.

7.1 Discourse Parser

A major source of error in detecting disagreement arises because of inaccuracies in our discourse parser. In particular, document-level discourse parsing is a challenging task, with relatively few parsers available at the time of this analysis (Joty et al., 2013; Hernault et al., 2010). We chose to use the discourse parser developed by Joty et al. which both identifies elementary discourse units in a text, and then builds a document-level discourse tree using Conditional Random Fields. Because their approach uses an optimal parsing algorithm as opposed to a greedy parsing algorithm, they are able to achieve much higher accuracies in relation and structure identification than other available parsers. Here, results from their parser on the standard RST-DT dataset are presented since there is no currently available dialogic corpora to compare to.

	RST-DT			Instructional
Metrics	Joty	HILDA	Human	Joty
Span	83.84	74.68	88.70	81.88
Nuclearity	68.90	58.99	77.72	63.13
Relation	55.87	44.32	65.75	43.60

Table 5: Joty et al. document-level parser accuracy of the parser used in this paper. The parser was originally tested on two corpora: RST-DT and Instructional. HILDA was the state-of-the-art parser at that time. Span and Nuclearity metrics assess the quality of the structure of the resulting tree, while the Relation metric assesses the quality of the relation labels.

Examining the relation labels confusion matrix

	T-C	T-O	T-CM	M-M	CMP	EV	SU	CND	EN	CA	TE	EX	BA	CO	JO	S-U	AT	EL
T-C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
T-O	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T-CM	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	7
M-M	0	0	0	10	0	0	0	0	0	0	0	1	1	0	0	0	1	3
CMP	0	0	0	1	4	0	0	1	0	1	0	3	3	0	1	1	0	2
EV	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	2	11
SU	0	0	0	0	0	0	8	0	0	0	0	0	0	0	1	0	0	12
CND	0	0	0	0	0	0	0	22	0	0	0	0	1	3	0	0	3	2
EN	0	0	0	0	0	0	0	1	24	1	0	0	0	0	0	0	1	7
CA	0	0	0	0	0	0	0	0	2	3	0	4	2	2	7	0	3	11
TE	0	0	0	1	0	0	0	1	2	0	7	1	9	1	9	0	3	4
EX	0	0	0	1	0	0	0	0	1	5	0	12	0	1	3	0	3	12
BA	0	0	0	1	0	0	0	1	0	1	4	1	19	2	6	1	5	12
CO	0	0	0	1	2	0	0	2	0	1	3	2	2	33	7	0	0	9
JO	0	0	0	0	0	0	1	2	0	1	1	1	1	2	57	1	0	13
S-U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	85	1	0
AT	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	272	9
EL	0	1	0	0	0	0	0	0	14	6	1	8	1	0	8	2	2	359

Figure 2: Confusion matrix for relation labels on RST-DT. The X-axis represents predicted relations, while the Y-axis corresponds to true values. The relations are Topic-Change (T-C), Topic-Comment (T-CM), Textual Organization (T-O), Manner-Means (M-M), Comparison (CMP), Evaluation (EV), Summary (SU), Condition (CND), Enablement (EN), Cause (CA), Temporal (TE), Explanation (EX), Background (BA), Contrast (CO), Joint (JO), Same-Unit (S-U), Attribution (AT) and Elaboration (EL).

for the discourse parser in Figure 2, some of the chosen rhetorical features make even more sense. In particular, the confusion of ELABORATION and EXPLANATION may account for the perceived importance of ELABORATION relations in the analysis. Likewise, CAUSE (which may be present when users attribute positive or negative qualities to an entity, signalling disagreement) is often confused with JOINT and ELABORATION which were often picked as important features by our classifiers.

8 Conclusions and Future Work

In this paper, we have described a new set of features for detecting disagreement in online blog forums. By treating a written conversation as a series of linked monologues, we can apply a document level discourse parser to extract a discourse tree for the conversation. We then aggregate this information in a *relation graph*, which allows us to capture post-level rhetorical relations between users. Combining this approach with sentiment features shows significantly improved performance in both accuracy and F-score over a baseline consisting of structural and lexical features as well as referral counts, punctuation, and discourse markers. In building our new crowd-sourced corpus from Slashdot, we have also shown the challenges of detecting subtle disagreement in a dataset that contains a significant number of news-style discussions.

In future work, we will improve sentiment features by considering methods to detect opinion-topic pairs in conversation, similar to Somasundaran and Wiebe (2009). Additionally, we will incorporate generalized dependency and POS features (Abbott et al., 2011), which were not used in this analysis due to the very small number of training samples in our dataset. The fragment quotation graph features did not perform as well as we expected, and in future work we would like to investigate this further. Furthermore, we will also explore how to create a discourse tree from the thread structure of a conversation (instead of from its temporal structure), and verify whether this improves the accuracy of the relation graphs, especially when the temporal structure is not representative of the reply-to relationships.

Finally, we plan to apply our novel feature set to other corpora (e.g., *ARGUE*) in order to study the utility of these features across genres and with respect to the accuracy of the discourse parser. This may provide insights into where discourse parsers can be most effectively used, as well as how to modify parsers to better capture rhetorical relations between participants in conversation.

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 LSM*, pages 2-11.

Amjad Abu-Jbara, Mona Diab, Pradeep Dasigi, and Dragomir Radev. 2012. Subgroup detection in ideological discussions. In *Proceedings of ACL*, pages 399-409.

Rakesh Agrawal, Sridhar Rajagopalan, Ramakrishnan Srikant, and Yirong Xu. 2003. Mining newsgroups using networks arising from social behavior. In *Proceedings of WWW*, pages 529-535.

Rawia Awadallah, Maya Ramanath, Gerhard Weikum. 2012. Harmony and Dissonance: Organizing the People's Voices on Political Controversies. In *Proceedings of WSDM*, pages 523-532.

Prakhar Biyani, Sumit Bhatia, Cornelia Caragea, Prasenjit Mitra. 2014. Using non-lexical features for identifying factual and opinionative threads in online forums. In *Knowledge-Based Systems*, in press.

Penelope Brown and Stephen Levinson. 1987. *Politeness: Some universals in language usage.* Cambridge University Press.

Giuseppe Carenini and Jackie Cheung. 2008. Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality. In *Proceedings of INLG*, pages 33-41.

Giuseppe Carenini, Raymond Ng, Xiaodong Zhou. 2007. Summarizing Email Conversations with Clue Words. In *Proceedings of WWW*, pages 91-100.

Yoonjung Choi, Yuchul Jung, and Sung-Hyon Myaeng. 2010. Identifying controversial issues and their subtopics in news articles. In *Proceedings of PAISI*, pages 140-153.

Alexander Conrad, Janyce Wiebe and Rebecca Hwa. 2012. Recognizing Arguing Subjectivity and Argument Tags. In *ACL Workshop on Extra-Propositional Aspects of Meaning*, pages 80-88.

Jean E. Fox Tree. 2010. Discourse markers across speakers and settings. *Language and Linguistics Compass*, 3(1):1-113.

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 ACL*, pages 669-es.

Sebastian Germesin and Theresa Wilson. 2009. Agreement Detection in Multiparty Conversation. In *Proceedings of International Conference on Multimodal Interfaces* pages 7-14.

Vicenc Gomez, Andreas Kaltenbrunner and Vicente Lopez. 2008. Statistical Analysis of the Social Network and Discussion Threads in Slashdot. In *Proceedings of WWW*, pages 645-654.

Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev. 2010. What's with the attitude?: identifying sentences with attitude in online discussions. In *Proceedings of EMNLP*, pages 1245-1255.

Hugo Hernault, Helmut Prendinger, David A. duVerle and Mitsuru Ishizuka. 2010. HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue and Discourse*, 1(3):1-33.

Laurence R. Horn. 1989. *A natural history of negation.* Chicago University Press.

Shafiq Joty, Giuseppe Carenini, and Chin-Yew Lin. 2011. Unsupervised modeling of dialog acts in asynchronous conversations. In *Proceedings of IJCAI*, pages 1807-1813.

Shafiq Joty, Giuseppe Carenini, Raymond Ng and Yashar Mehdad. 2013. Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis. In *Proceedings of ACL*.

Shafiq Joty, Giuseppe Carenini and Raymond Ng. 2013. Topic Segmentation and Labeling in Asynchronous Conversations. *Journal of AI Research*, 47:521-573.

Ching-Sheng Lin, Samira Shaikh, Jennifer Stromer-Galley, Jennifer Crowley, Tomek Strzalkowski, Veena Ravishankar. 2013. Topical Positioning: A New Method for Predicting Opinion Changes in Conversation. In *Proceedings of LASM*, pages 41-48.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243-281.

Gilad Mishne and Natalie Glance. 2006. Leave a reply: An analysis of weblog comments. In *Proceedings of WWW*.

Amita Misra and Marilyn Walker. 2013. Topic Independent Identification of Agreement and Disagreement in Social Media Dialogue. In *Proceedings of SIGDIAL*, pages 41-50.

Arjun Mukherjee and Bing Liu. 2012. Modeling review comments. In *Proceedings of ACL*, pages 320-329.

Akiko Murakami and Rudy Raymond. 2010. Support or Oppose? Classifying Positions in Online Debates from Reply Activities and Opinion Expressions. In *Proceedings of the International Conference on Computational Linguistics*, pages 869-875.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, Deborah Cai, Jennifer Midberry, Yuanxin Wang. 2014. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning* 95:381-421.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting Product Features and Opinions from Reviews. In *Proceedings of HLT/EMNLP*, pages 339-346.

Ana-Maria Popescu and Marco Pennacchiotti. 2010. Detecting controversial events from twitter. In *Proceedings of CIKM*, pages 1873-1876.

Stephan Raaijmakers, Khiet Truong, Theresa Wilson. 2008. Multimodal subjectivity analysis of multiparty conversation. In *Proceedings of EMNLP*, pages 466-474.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of ACL*, pages 16501659.

Swapna Somasundaran, Josef Ruppenhofer, Janyce Wiebe. 2007. Detecting Arguing and Sentiment in Meetings. In *Proceedings of SIGDIAL Workshop on Discourse and Dialogue*.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of ACL*, pages 226-234.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of NAACL, Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116124.

Amanda Stent and James Allen. 2000. Annotating Argumentation Acts in Spoken Dialog. *Technical Report*.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Journal of Computational Linguistics*, 37(2):267-307.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, pages 327-335.

Wen Wang, Sibel Yaman, Kristen Precoda, Colleen Richey, and Geoffrey Raymond. 2011. Detection of agreement and disagreement in broadcast conversations. In *Proceedings of ACL*, pages 374-378.

Bonnie Webber and Rashmi Prasad. 2008. Sentence-initial discourse connectives, discourse structure and semantics. In *Proceedings of the Workshop on Formal and Experimental Approaches to Discourse Particles and Modal Adverbs*.

Jie Yin, Paul Thomas, Nalin Narang, and Cecile Paris. 2012. Unifying local and global agreement and disagreement classification in online debates. In *Proceedings of Computational Approaches to Subjectivity and Sentiment Analysis*, pages 61-69.