How can you say such things?!?: Recognizing Disagreement in Informal Political Argument

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree,
Robeson Bowmani, and Joseph King
University of California Santa Cruz
abbott|maw@soe.ucsc.edu,panand|foxtree@ucsc.edu

Abstract

The recent proliferation of political and social forums has given rise to a wealth of freely accessible naturalistic arguments. People can "talk" to anyone they want, at any time, in any location, about any topic. Here we use a Mechanical Turk annotated corpus of forum discussions as a gold standard for the recognition of disagreement in online ideological forums. We analyze the utility of meta-post features, contextual features, dependency features and word-based features for signaling the disagreement relation. We show that using contextual and dialogic features we can achieve accuracies up to 68% as compared to a unigram baseline of 63%.

1 Introduction

The recent proliferation of political and social forums has given rise to a wealth of freely accessible naturalistic arguments. People can "talk" to anyone they want, at any time, in any location, about any topic. Their conversations range from current political topics such as national health care to religious questions such as the meaning of biblical passages. See Figure 1. We aim to automatically derive representations of the discourse structure of such arguments and to gain a deeper theoretical and empirical understanding of the linguistic reflexes of perlocutionary acts such as persuasion (Austin, 1965).

The study of the structure of argumentative communication has a long lineage in psychology (Cialdini, 2000) and rhetoric (Hunter, 1987), but the historical lack of a large corpus of naturalistic examples has limited empirical work to a handful of genres (e.g., editorials or simulated negotiations). Argumentation is above all tactical. Thus being able to effectively model it would afford us a glimpse of pragmatics beyond the conversational turn. More practically, an increasing portion of information and opinion exchange online occurs in natural dialogue, in forums, in webpage comments, and in the back

Topic	Q-R: Post
Evolution	**Q:** How can you say such things? The Bible says that God CREATED over and OVER and OVER again! And you reject that and say that everything came about by evolution? If you reject the literal account of the Creation in Genesis, you are saying that God is a liar! If you cannot trust God's Word from the first verse, how can you know that the rest of it can be trusted? **R:** It's not a literal account unless you interpret it that way.
Gay marriage	**Q:** Gavin Newsom- I expected more from him when I supported him in the 2003 election. He showed himself as a family-man/Catholic, but he ended up being the exact oppisate, supporting abortion, and giving homosexuals marriage licenses. I love San Francisco, but I hate the people. Sometimes, the people make me want to move to Sacramento or DC to fix things up. **R:** And what is wrong with giving homosexuals the right to settle down with the person they love? What is it to you if a few limp-wrists get married in San Francisco? Homosexuals are people, too, who take out their garbage, pay their taxes, go to work, take care of their dogs, and what they do in their bedroom is none of your business.
Abortion	**Q:** Equality is not defined by you or me. It is defined by the Creator who created men. **R:** Actually I think it is defined by the creator who created all women. But in reality your opinion is gibberish. Equality is, like every other word, defined by the people who use the language. Currently it means "the same". People aren't equal because they are not all the same. Any attempt to argue otherwise is a display of gross stupidity.

Figure 1: Sample Quote/Response Pairs

and forth of short messages (e.g., Facebook status updates, tweets, etc.) Effective models of argumentative discourse thus have clear applications in automatic summarization, information retrieval, or predicting real-world events such as how well a new product is being received or the outcome of a popular vote on a topic (Bollen et al., 2011).

In this paper, we focus on an important initial task for the recognition of argumentative structure: automatic identification of agreement and disagreement. We introduce the ARGUE corpus, an annotated collection of 109,553 forum posts (11,216 discussion threads) from the debate website 4forums.com. On 4forums, a person starts a discussion by posting a topic or a question in a particular category, such as society, politics, or religion. Some example topics can be seen in Table 1. Forum participants can then post their opinions, choosing whether to respond directly to a previous post or to the top level topic (start a new thread). These discussions are essentially dialogic; however the affordances of the forum such as asynchrony, and the ability to start a new thread rather than continue an existing one, leads to dialogic structures that are different than other multiparty informal conversations (Fox Tree, 2010). An additional source of dialogic structure in these discussions, above and beyond the thread structure, is the use of the quote mechanism, in which participants often break a previous post down into the components of its argument and respond to each component in turn. Many posts include quotations of previous posts. Because we hypothesize that these posts are more targeted at a particular proposition that the poster wants to comment on, than posts and replies in general, we focus here on understanding the relationship between a quoted text and a response, and the linguistic reflexes of those relationships. Examples of quote/response pairs for several of our topics are provided in Figure 1.

The most similar work to our own is that of Wang & Rose (2010) who analyzed Usenet forum quote/response structures. This work did not distinguish agreement vs. disagreement across quote/response pairs. Rather they show that they can use a variant of LSA to improve accuracy for identifying a parent post, given a response post, with 70% accuracy. Other similar work uses Congressional debate transcripts or blogs or other social media to develop methods for distinguishing agreement from disagreement or to distinguish rebuttals from out-of-context posts (Thomas et al., 2006; Bansal et al., 2008; Awadallah et al., 2010; Walker et al., ; Burfoot, 2008; Mishne and Glance, 2006; Popescu and Pennacchiotti, 2010). These methods are directly applicable, but the genre of the language is so different from our informal forums that the results are not directly comparable. Work by Somasundaran & Wiebe (2009, 2010) has examined debate websites and focused on automatically determining the stance of a debate participant with respect to a particular issue. This work has treated each post as a text to be classified in terms of stance, for a particular topic, and shown that discourse relations such as concessions and the identification of argumentation triggers improves performance . Their work, along with others, also indicates that for such tasks it is difficult to beat a unigram baseline (Pang and Lee, 2008). Other work has focused on the social network structure of online forums (Murakami and Raymond, 2010; Agrawal et al., 2003). However, Agarwal's work assumed that adjacent posts always disagree, and did not use any of the information in the text. Murakami & Raymond (2010) show that simple rules defined on the textual content of the post can improve over Agarwal's results.

Section 2 discusses our corpus in more detail, describes how we collected annotations using Mechanical Turk, and presents results of a corpus analysis of the use of particular discourse cues. Section 3 describes how we set up classification experiments for distinguishing agreement from disagreement, and Section 4 presents our results for agreement classification. We also characterize the linguistic reflexes of this relation. We analyze the utility of meta-post features, contextual features, dependency features and word-based features for signaling the disagreement relation. We show that using contextual and dialogic features we can achieve accuracies up to 68% as compared to a unigram baseline of 63%.

2 Data and Corpus Analysis

Table 1 provides an overview of some of the characteristics of our corpus by topic. Figure 2 shows the wording of the survey questions that we posted for each quote/response as Mechanical Turk hits.

Topic	Discs	Posts	NumA	P/A	A>1P	PL	Agree	Sarcasm	Emote	Attack	Nasty
evolution	872	10292	580	17.74	76%	576	10%	6%	16%	13%	9%
gun control	825	7968	411	19.39	66%	521	11%	8%	21%	16%	12%
abortion	564	7354	574	12.81	69 %	454	9%	6%	31%	16%	12%
gay marriage	305	3586	342	10.49	69%	522	13%	9%	23%	12%	8%
existence of God	105	1581	258	6.13	66%	569	11 %	7%	26%	14%	10%
healthcare	81	702	112	6.27	64%	522	14%	10%	34%	17%	17%
communism vs. capitalism	38	585	110	5.32	59%	393	23%	8%	15%	8%	0%
death penalty	25	500	138	3.62	62%	466	25%	5%	5%	5%	5%
climate change	40	361	116	3.11	55%	375	20%	9%	17%	26%	17%
marijuana legalization	13	160	72	2.22	38%	473	5%	2%	20%	5%	5%

Table 1: Characteristics of Different Topics. **KEY**: Number of discussions and posts on the topic (**Discs, Posts**). Number of authors (**NumA**). Posts per author (**P/A**). Authors with more than one post (**A > 1P**). Median post Length in Characters (**PL**). The remainder of the columns are the annotations shown in Figure 2. Percentage of posts that agree (**Agree%**), use sarcasm (**Sarcasm%**), are emotional (**Emote**), attack the previous poster (**Attack**), and are nasty (**Nasty**). The scalar values are threshholded at -1,1.

Our corpus is derived from a debate oriented internet forum called 4forums.com. It is a typical internet forum built on the vBulletin software. People initiate discussions (threads) and respond to others' posts. Each thread has a tree-like dialogue structure. Each post has author information and a timestamp with minute resolution. Many posts include quotations of previous posts. For this work we chose to focus on quotations because they establish a clear relationship between the quoted text and the response.

Our corpus consists of 11,216 discussions and 109,553 posts by 2764 authors. We hand annotated discussions for topic from a set of previously identified contentious political and social issues. The website is tailored to a US audience and our topics are somewhat US centric. Table 1 describes features of our topics in order of decreasing discussion count. When restricted to these topics, the corpus consists of 2868 discussions, 33,089 posts, and 1302 authors.

Many posts include quotations. Overall 60,382 posts contain one or more quotation. Within our topics of interest, nearly 20,000 posts contain quotations. We defined a quote-response pair (Q-R pair) where the response was the portion of the responding post directly following a quotation but preceding any additional quotations.

We selected 10,003 Q-R pairs from the topics of interest for a Mechanical Turk annotation task. These were biased by cue word to ensure adequate data for discourse marker analysis (See Section 2.1. For this task we showed annotators seven Q-R pairs and asked them to judge Agreement/Disagreement and a set of other measures as shown in Figure 2. Most of our measures were scalar; we chose to do this because previous work on estimating the relationship between MTurk annotations and expert annotations suggest that taking the means of scalar annotations could be a good way to reduce noise in MTurk annotations (Snow et al., 2008). For all of the measures annotated, the Turkers were not given additional definitions of their meaning. For example, we let Turkers to use their native intuitions about what it means for a post to be sarcastic, since previous work suggests that non-specialists tend to collapse all forms of verbal irony under the term sarcastic (Bryant and Fox Tree, 2002). We did not ask Turkers to distinguish between sarcasm and other forms of verbal irony such as hyperbole, understatement, rhetorical questions and jocularity (Gibbs, 2000).

Agreement was a scalar judgment on an 11 point scale [-5,5] implemented with a slider. The annotators were also able to signal uncertainty with an CAN'T TELL option. Each of the pairs was annotated by 5-7 annotators. We showed the first 155 characters of each quote and each response. We also provided a SHOW MORE button which expanded the post to its full length. After annotation, we removed a number of Q-R pairs in cases where a clear link between the quote and a previous post could not be established, e.g. the source quoted was not another post, but the NY Times. This left us with 8,242 Q-R pairs for our final analysis. Resampling to a natural distribution left us with 2,847 pairs which we used to build our machine learning test set. We used the remaining annotated and unannotated pairs for de-

velopment.

Type	α	Survey Question
S	0.62	**Agree/Disagree**: Does the respondent agree or disagree with the prior post?
S	0.32	**Fact/Emotion**: Is the respondent attempting to make a fact based argument or appealing to feelings and emotions?
S	0.42	**Attack/Insult**: Is the respondent being supportive/respectful or are they attacking/insulting in their writing?
B	0.22	**Sarcasm**: Is the respondent using sarcasm?
S	0.46	**Nice/Nasty**: Is the respondent attempting to be nice or is their attitude fairly nasty?

Figure 2: Mechanical Turk Annotations (Binary = B and Scalar = S) and level of agreement as Krippendorff's α.

Figure 3 provides examples from the end points and means of the annotations for three of the questions, Respect/Insult, Sarcasm, and Fact/Emotion. Nice/Nasty and Respect/Insult are strongly correlated by worker annotations $r(54003) = 0.84$, $p <$ 2.2e-16 and both weakly correlated with Agree/Disagree ratings ($r(54003) = 0.32$ and $r(54003)=0.36$, respectively; $p <$ 2.2e-16) and Fact/Emotion ratings ($r(54003) = 0.32$ and $r(54003)=0.31$, respectively; $p <$ 2.2e-16), while Agree/Disagree and Fact/Emotion ratings show the smallest correlation, $r(54003)=0.11$, $p <$ 2.2e-16. For the linguistic marker correlations discussed below we averaged scores across annotators, a process which sharpened correlations (e.g., Respect/Insult means correlate with Agree/Disagree means more strongly ($r(5393) = 0.51$) as well as Nice/Nasty means ($r(5393) = 0.91$; Agree/Disagree is far less correlated with Fact/Emotion ($r(5393) = 0.07$). Interannotator agreement was computed using Krippendorff's α (due to the variability in number of annotators that completed each hit), assuming an ordinal scale for all measures except sarcasm; see Figure 2. The low agreement for Sarcasm accords with native intuition – it is the class with the least dependence on lexicalization and the most subject to inter-speaker stylistic variation. The relatively low results for Fact/Emotion is perhaps due to the emotional charge many ideological arguments engender; informal examination of posts that showed the most disagreement in this category often showed a cutting comment or a snide remark at the end of a post, which was was ignored by some annotators and evidence for others (one Emotional post in Figure 3 is clearly an insult, but was uniformly labeled as -5 by all annotators).

2.1 Discourse Markers

Both psychological research on discourse processes (Fox Tree and Schrock, 1999; Fox Tree and Schrock, 2002; Groen et al., 2010) and computational work on agreement (Galley et al., 2004) indicate that discourse markers are strongly associated with particular pragmatic functions. Because of their salient position, we test the role of turn-initial markers in predicting upcoming content (Fox Tree and Schrock, 2002; Groen et al., 2010). Based on manual inspection of a subset of the corpus, we constructed a list of 20 discourse markers; 17 of these occurred at least 50 times in a quote response (upper bound of 700 samples): *actually*, *and*, *because*, *but*, *I believe*, *I know*, *I see*, *I think*, *just*, *no*, *oh*, *really*, *so*, *well*, *yes*, *you know*, *you mean*. All of their occurrences became part of the 10,003 Q-R pairs annotated.

The top discourse markers highlighting disagreement were *really* (67% read a response beginning with this marker as prefacing a disagreement with a prior post), *no* (66%), *actually* (60%), *but* (58%), *so* (58%), and *you mean* (57%). At this point, the next most disagreeable category was the unmarked category, with about 50% of respondents interpreting an unmarked post as disagreeing. On the other hand, the most agreeable marker was *yes* (73% read a response beginning with this marker as prefacing an agreement) followed by *I know* (64%), *I believe* (62%), *I think* (61%), and *just* (57%). The other markers were close to the unmarked category: *and* (50%), *because* (51%), *oh* (51%), *I see* (52%), *you know* (54%), and *well* (55%).

The overall agreement on sarcasm was low, as in other computational work on recognizing sarcasm (Davidov et al., 2010). At most, only 31% of respondents agreed that the material after a discourse marker was sarcastic, with the most sarcastic markers being *you mean* (31%), *oh* (29%), *really* (24%), *so* (22%), and *I see* (21%). Only 15% of respondents rated the unmarked category as sarcastic (e.g., fewer than 1 out of 6 respondents). The cues *I think* (10%), *I believe* (9%), and *actually* (10%) were the least sarcastic markers.

Taken together, these ratings suggest that the cues *really*, *you mean*, and *so* can be used to indicate both

Class	Very High Degree	Neutral	Very Low Degree
Insult or Attack	Well, you have proven yoruself to be a man with no brain, that is for sure. The definition that was given was the one that scientists use, not the layperson.	The empire you defend is tyrannical. They are responsible for the death of millions.	Very well put.
	Is that what you said right before they started banning assault weapons?...Obviously, you're gullible. Since you're such a brainiac and all, why don't you visit the UN website and see what your beloved UN is up to?	Bad comparisons. A fair comparison would be comparing the total number of defensive gun uses to the total number of gun crimes (not just limiting it to gun homicides).	In some cases yes, in others no. If the mutation gives a huge advantage, then there will be a decline in the size of the gene pool for a while (eg when the Australian rabbit population...
Sarcasm	My pursuit of happiness is denied by trees existing. Let's burn them down and destroy the environment. It's much better than me being unhappy.	An interesting analysis of that article you keep quoting from the World Net Daily [url]	I would suggest you look at the faero island mouse then. That is a new species, and it is not man doing it, but rather nature itself.
	Like the crazy idea the Earth goes around the Sun.	Indeed there is no diffrence it is still a dead baby but throwing a baby in a trash can and leaving it for dead is far more cruel than abortion.	Too late, drug usage has already created those epidemics. Legalizing drugs may increase some of them temporarily, but they already exist.
Emotion-based Argument	Really! You can prove that most pro-lifers don't care about women?...it is idiotic thinking like this that makes me respect you less and less.	Fine by me. First, I don't consider having a marriage recognized by government to be a "right". Second, I've said many times I don't think government should be in the marriage business at all.	Sure. Here is an explanation. The 14C Method. That is from the Radiocarbon WEB info site by the Waikato Radiocarbon Dating Lab of the University of Waikato (New Zeland).
	I love Jesus John the Beloved is my most favorite writer throughout time If you think I have a problem with a follower of Jesus your wrong. I have a problem with the Christians	I agree that the will to survive is an amazing phenomenon when put to the test. But I do not agree with your statement of life at *any* cost. There will always be a time when the humane/loving thing to do is to let an infant/child/adult go.	Heller is about determining the answer to a long standing question on the nature of the Second Amendment, and how much gun control is legally allowed. Roe v. Wade is about finding legal precedent for the murder of unborn children. I see absolutely no comparison between the two.

Figure 3: Sample Responses for the Insult, Sarcasm, and Fact/Feeling spectrums

disagreement and sarcasm. However, *but*, *no*, and *actually* can be used for disagreement, but not sarcasm. And *I know* (14% sarcastic, similar to None), *I believe*, and *I think* can be used for non-sarcastic agreement.

From informal analyses, we hypothesized that *really* and *oh* might indicate sarcasm. While we found evidence supporting this for *really*, it was not the case for *oh*. Instead, *oh* was used to indicate emotion; it was the discourse marker with the highest ratings of feeling over fact.

Despite the fact that it would seem that disagreement would be positively correlated with sarcasm, disagreement and sarcasm were not related. There were two tests possible. One tested the percentage of people who identified an item as disagreeing against the percentage of people who identified it as sarcasm, $r(16)$ = -.27, p = .27 (tested on 17 discourse markers plus the None category). The other tested the degree of disagreement (from -5 to +5) against the percentage of people who identified the post as sarcastic, $r(16)$ = -.33, p = .18.

However, we did observe relationships between sarcasm and other variables. Two results support the argument that sarcasm is emotional and personal. The more sarcastic, the nastier (rather than nicer), $r(16)$ = .87, $p < .001$. In addition, the more sarcastic, the more emotional (over factual) respondents were judged to be, $r(16)$ = .62, p = .006 Taken together, these analyses suggest that sarcasm is emotional and personal, but not necessarily a sign of disagreement.

3 Machine Learning Experimental Setup

For our experiments we used the Weka machine learning toolkit. All results are from 10 fold cross-validation on a balanced test set. Unless otherwise mentioned, we used thresholds of 1 and -1 on the mean agreement judgment to determine agreement

and disagreement respectively. We omitted those Q-R pairs which were judged neutral (mean annotator judgment in the (-1,1) range).

As described above, from the original 10,003 Q-R pairs we applied certain constraints (notably requirement that we be able to identify the originating post) which left us with 8,242. We then resampled to obtain a natural distribution leaving us with 2,847 pairs. Applying the (-1,1) threshold and balancing the result yielded a test set of 682 Q-R pairs.

3.1 Classifiers

Our experiments used two simple classifiers: NaiveBayes and JRip. NaiveBayes makes a strict independence assumption and can be swamped by the sheer number of features we used, but it is a solid baseline and does a decent job of suggesting which features are more powerful. JRip is a rule based classifier which produces a compact model suitable for human consumption and quick application. JRip is not without its own limitations but, for our task, it shows better results than NaiveBayes. The model it builds uses only a handful of features.

3.2 Feature Extraction

Our aim was to develop features for the automatic identification of agreement and disagreement that would do well on the task and provide useful baselines for comparisons with previous and future work. Features are grouped into sets as shown in Table 2 and discussed in more detail below.

Set	Description/Examples
MetaPost	Non-lexical features. E.g. posterid, time between posts, etc.
Unigrams, Bigrams	Word and Word Pair frequencies
Cue Words	Initial unigram, bigram, and trigram
Punctuation	Collapsed into one of the following: ??, !!, ?!
LIWC	LIWC measures and frequencies
Dependencies	Dependencies derived from the Stanford Parser.
Generalized Dependencies	Dependency features generalized with respect to POS of the head word and opinion polarity of both words.

Table 2: Feature Sets, Descriptions, and Examples

Unigrams, Bigrams, Trigrams. Results of previous work suggest that a unigram baseline can be difficult to beat for certain types of debates (Walker et al., ; Somasundaran and Wiebe, 2010). Thus we derived both unigrams and bigrams as features. We captured the final token as a feature by padding with -nil- tokens when building the bigrams. See below for comments on initial uni/bi/tri-grams.

MetaPost Info. Previous work suggested that non-lexical features like poster ids and the time between posts might contain indicators of disagreement. People on these forums get to know one another and often enjoy repeatedly arguing with the same person. In addition, we hypothesized that the "heat" of a particular conversation could be correlated with rapid-fire exchanges, as indicated by short time periods between posts.

Thus these features involve structure outside of the quote/response text. This includes author information, time between posts, the $\log_{10}$ of the time between posts, the number of other quotes in the response, whether the quote responds to a post by the response's author, the percent of the quoted post which is actually quoted, whether the quoted post is by the same author as the response (there were only an handful of these), whether the response mentions the quote author by name, and whether the response is longer than the quote.

The forum software effectively does this annotation for us so there is no reason not to consider it as a clue in our quest to understand and interpret online dialogue.

Discourse Markers. Previous work on dialogue analysis has repeatedly noted the discourse functions of particular discourse markers, and our corpus analysis above also suggests their use in this particular dataset (Hirschberg and Litman, 1993; Fox Tree, 2010; Schiffrin, 1987; Di Eugenio et al., 1997; Moser and Moore, 1995). However, because discourse markers can be stacked up *Oh, so really* we decided to represent this feature as post initial unigrams, bigrams and trigrams.

Repeated Punctuation. Informal analyses of our data suggested that repeated sequential use of particular types of punctuation such as **!!** and **??** did not mean the same thing as simple counts or frequencies of punctuation across a whole post. Thus we developed distinct features for a subset of these repetitions.

LIWC. We also derived features using the Linguistics Inquiry Word Count tool (LIWC-2001) (Pennebaker et al., 2001). LIWC classifies words

into 69 categories and counts how many words get classified into each category. Some LIWC features that we expect to be important are words per sentence (WPS), pronominal forms, and positive and negative emotion words.

Dependency and Generalized Dependency. We used the Stanford parser to extract dependency features for each quote and response (De Marneffe et al., 2006; Klein and Manning, 2003). The dependency parse for a given sentence is a set of triples, composed of a grammatical relation and the pair of words for which the grammatical relation holds (rel_i, w_j, w_k), where rel_i is the dependency relation among words w_j and w_k. The word w_j is the HEAD of the dependency relation.

Following (Joshi and Penstein-Rosé, 2009) we extracted generalized dependency features by leaving one dependency element lexicalized and generalizing the other to part of speech. Joshi & Rose's results suggested that this approach would work better than either fully lexicalized or fully generalized dependency features.

Opinion Dependencies. Somasundaran & Wiebe (2009) introduce the concept of features that identify the TARGET of opinion words. Inspired by this approach, we used the MPQA dictionary of opinion words to select the subset of dependency and generalized dependency features in which those opinion words appear. For these features we replace the opinion words with their positive or negative polarity equivalents.

Cosine Similarity. This feature is based on previous work on threading. We derive cosine-similarity measure using tf-idf vectors where the document frequency was derived from the entire topic restricted corpus.

Annotations. We also add features representing information that we do not currently derive automatically, but which might be automatically derived in future work based on annotations in the corpus. These include the topic and Mechanical Turk annotations for Fact/Emotion, Respect/Insult, Sarcasm, and Nasty/Nice, which could reasonably be expected to be recognized independently of Agreement/Disagreement.

Feature type	Selected Features
Meta	number-of-other-quotes, percent-quoted, author-quote-USERNAME
Initial n-gram	*yes, so, I agree, well said, really?, I don't know*
Bigram	*that you, ? -nil-, you have, evolution is*
Dependency	dep-nsubj(agree, i), dep-nsubj(think, you), dep-prep-with(agree, you)
Opinion Dependency	dep-opinion-nsubj(negative, you), dep-opinion-dep(proven, negative), dep-opinion-aux(positive, to)
Annotations	topic-gay marriage, mean-response-nicenasty, mean-unsure-sarcasm

Table 3: Some of the more useful features for each category, using χ^2 for feature selection.

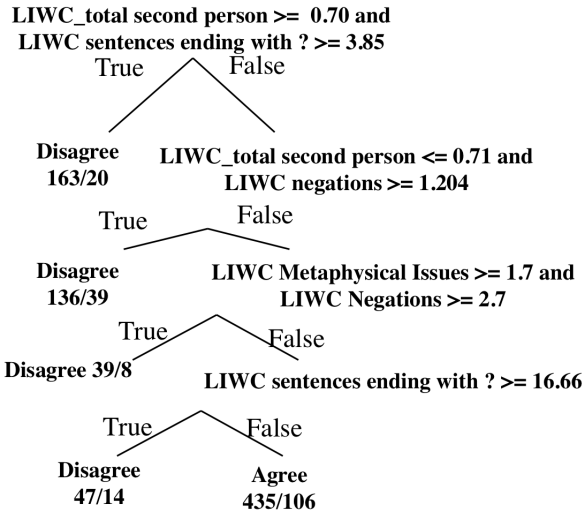

Figure 4: Sample model learned using JRip. The numbers represent (total instances covered by a rule / number incorrectly labeled). This particular model was built on development data.

4 Results

Table 3 shows features which were selected for each of our feature categories using a χ^2 test for feature selection. These results vindicate our interest in discourse markers as cues to argument structure, as well as the importance of the generalized dependency features and opinion target pairs (Wang and Rosé, 2010; Somasundaran and Wiebe, 2009). Figure 4 shows a sample model learned using JRip.

We limit our pair-wise comparisons between classifiers and feature sets to those corresponding to par-

Feats	NB	JRipχ^2
Uni,UniCue	0.578	0.626
BOW	0.598	0.654
Meta	0.579	0.588
Response Local	0.600	0.666
Quote Local	0.531	0.588
Both Local	0.601	**0.682**
Meta+Local	0.603	0.654
All	0.603	0.632
Just Annotations	0.765	0.814
All+Annotations	0.603	0.795

Table 4: Accuracies on a balanced test set (random baseline: 0.5). NB = NaiveBayes. JRipχ^2 = Jripper with χ^2 feature selection on the training set during cross validation. **BOW** = Unigrams, CueWords, Bigrams, Trigrams, LIWC, Repeated Punctuation. **Response/Quote/Both Local** uses only those features which exist in the text of the response or quote respectively. It consists of LIWC, dependencies, generalized dependencies, the various n-grams, and length measures.

ticular hypotheses. We conducted five tests with Bonferroni correction to .01 for a .05 level of significance.

While we hypothesized that more sophisticated linguistic features would improve over unigram features alone, a paired t-test using the results in Table 4 indicate that there is no statistical difference between the performance of JRip using only response local features (JRip,ResponseLocal), as compared to the Unigram,UniCue features ($t(9) = 2.18$, $p = .06$).

However, a paired t-test using the results in Table 4 indicate that there is a statistical difference between the performance of JRip using local features from both the quote and the response, (JRip,BothLocal) as compared to the Unigram,UniCue features ($t(9) = 3.94$, $p = .003$). This shows that the contextual features do matter, even though (JRip,BothLocal) does **not** provide significant improvements over (JRip,Response Local) ($t(9) = .92$, $p = .38$).

In general, examination of the table suggests that the JRip classifier performs better than Naive Bayes. A paired t-test indicates that there is a statistical difference between the performance of JRip using local features from both the quote and the response, (JRip,BothLocal) (JRip,BothLocal) and Naive Bayes using local features from both the quote and the response, (NB,BothLocal) ($t(9) = 3.43$, $p = .007$).

In addition, with an eye toward the future, we examined whether automatic recognition of sarcasm, attack/insult, fact/feeling nice/nasty could possibly improve results for recognizing disagreement. Using the human annotations as a proxy for automatic results, we get classification accuracies of over 81% (JRip,JustAnnotations). This suggests it might be possible to improve results over our best current results (JRip,BothLocal) ($t(9) = 6.09$, $p < .001$).

Another interesting fact, is that despite its use in previous work for threading, the cosine similarity between the quote and response did not improve accuracy for the classifiers we tested, over and above the use of text-based contextual features. Further investigation is required to draw conclusions about this or similar metrics (LSA, PMI, etc.).

5 Discussion and Conclusion

In this paper, we have introduced a new collection of internet forum posts, the ARGUE corpus, collected across a range of ideological topics, and containing scalar Agreement/Disagreement annotations over quote-response pairs within a post. We have demonstrated that we can achieve a significant improvement over a unigram baseline agreement detection system using features from both a response and the quote being responded to.

Beyond agreement, the ARGUE corpus contains finer-grained annotations for degrees of insult, nastiness, and emotional appeal, as well as the presence of sarcasm. We have demonstrated that these classes (especially insult and nastiness) correlate with agreement. While the utility of these classes as features for agreement detection is dependent on how easily they are learned, in closing we note that they also afford us a richer understanding of how argumentative conversation flows. In section 2.1.2, we outlined how they can yield understanding of the potential functions of a discourse particle within a particular post. They may as allow us to understand the extent to which participants react in kind, rewarding insult with insult or kindness in turn. In future work, we hope to turn to these conversational dynamics.

In future work, it would be useful to build a ternary classifier which labels

Agree/Disagree/Neutral, thus reflecting the true distribution of these dialogue acts in the data. Additionally, the proportion of agreeing utterances varies widely across media so it may be desirable to add an appropriate prior when adapting the model to a new dataset.

Acknowledgments

This work was funded by Grant NPS-BAA-03 to UCSC and IARPA Grant on Persuasion in Dialogue to UCSC by subcontract from the University of Maryland. We'd like to thank Craig Martell for helpful discussions over the course of this project, and the anonymous reviewers for useful feedback. We would also like to thank Michael Minor and Jason Aumiller for their contributions to scripting and the database.

References

R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. 2003. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th international conference on World Wide Web*, pages 529–535. ACM.

J.L. Austin. 1965. *How to do things with words*. Oxford University Press, New York.

R. Awadallah, M. Ramanath, and G. Weikum. 2010. Language-model-based pro/con classification of political text. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 747–748. ACM.

M. Bansal, C. Cardie, and L. Lee. 2008. The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. *Proceedings of COLING: Companion volume: Posters*, pages 13–16.

J. Bollen, H. Mao, and X. Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*.

G.A. Bryant and J.E. Fox Tree. 2002. Recognizing verbal irony in spontaneous speech. *Metaphor and symbol*, 17(2):99–119.

C. Burfoot. 2008. Using multiple sources of agreement information for sentiment classification of political transcripts. In *Australasian Language Technology Association Workshop 2008*, volume 6, pages 11–18.

Robert B. Cialdini. 2000. *Influence: Science and Practice (4th Edition)*. Allyn & Bacon.

D. Davidov, O. Tsur, and A. Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116. Association for Computational Linguistics.

M.C. De Marneffe, B. MacCartney, and C.D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454. Citeseer.

Barbara Di Eugenio, Johanna D. Moore, and Massimo Paolucci. 1997. Learning features that predict cue usage. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, ACL/EACL 97*, pages 80–87.

J.E. Fox Tree and J.C. Schrock. 1999. Discourse Markers in Spontaneous Speech: Oh What a Difference an Oh Makes. *Journal of Memory and Language*, 40(2):280–295.

J.E. Fox Tree and J.C. Schrock. 2002. Basic meanings of you know and I mean. *Journal of Pragmatics*, 34(6):727–747.

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

M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 669–es. Association for Computational Linguistics.

R.W. Gibbs. 2000. Irony in talk among friends. *Metaphor and Symbol*, 15(1):5–27.

M. Groen, J. Noyes, and F. Verstraten. 2010. The Effect of Substituting Discourse Markers on Their Role in Dialogue. *Discourse Processes: A Multidisciplinary Journal*, 47(5):33.

Julia Hirschberg and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

John E. Hunter. 1987. A model of compliance-gaining message selection. *Communication Monographs*, 54(1):54–63.

M. Joshi and C. Penstein-Rosé. 2009. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 313–316. Association for Computational Linguistics.

D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

G. Mishne and N. Glance. 2006. Leave a reply: An analysis of weblog comments. In *Third annual workshop on the Weblogging ecosystem*. Citeseer.

Margaret G. Moser and Johanna Moore. 1995. Investigating cue selection and placement in tutorial discourse. In *ACL 95*, pages 130–137.

A. Murakami and R. Raymond. 2010. Support or Oppose? Classifying Positions in Online Debates from Reply Activities and Opinion Expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 869–875. Association for Computational Linguistics.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

J. W. Pennebaker, L. E. Francis, and R. J. Booth, 2001. *LIWC: Linguistic Inquiry and Word Count*.

A.M. Popescu and M. Pennacchiotti. 2010. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1873–1876. ACM.

Deborah Schiffrin. 1987. *Discourse Markers*. Cambridge University Press, Cambridge, U.K.

R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

S. Somasundaran and J. Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 226–234. Association for Computational Linguistics.

S. Somasundaran and J. Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.

M. Thomas, B. Pang, and L. Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics.

Marilyn Walker, Rob Abbott, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. Cats Rule and Dogs Drool: Classifying Stance in Online Debate.

Y.C. Wang and C.P. Rosé. 2010. Making conversational structure explicit: identification of initiation-response pairs within online discussions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 673–676. Association for Computational Linguistics.