Collaborative Research: RI: Processing Opinion Sharing Dialogs in Social Media
Walker, Anand, Whittaker, Fox Tree, Martell

1 Introduction

When people are making decisions, such as deciding where to go for dinner, what kind of car to buy, or how to vote in a local election, they often discuss and evaluate options with others, seeking out opinions and sharing their own. For thousands of years, these conversations have taken place in both formal and informal settings, in classrooms, courtrooms, and in formal organized debates, as well as in social or workplace settings, as illustrated in Fig. 1. However, with the increased popularity of discussion forums and social media the manner in which we consume other people's opinions is radically changing to being more **dialogic**. Our private dialogs have grown truly public, both in terms of who we can now converse with and whose opinions are available for the world to see, and these dialogs occur at scale. There are a huge number of opinion oriented websites with different dialogic properties concerning news, products, entertainment, health and politics. At last count, one site `4forums.com` that we scraped for our corpus, had 11,800 different dialogs totaling 84 million words. See Figs. 2 and 3.

Figure 1: Formal and informal debate over time.

The release of the Penn TreeBank (PTB) has led to great strides in processing the language of traditional media settings, while shared resources for work on dialog have lagged far behind. But more and more of the information available on the web is becoming dialogic, from forum conversations to comment threads on newspaper articles. And these conversations, such as those in Figs. 2 and 3, are very different from newspaper articles or broadcast news. Subjective genres in traditional media tend to be both monologic and formal, while online debates are strongly dialogic, interpersonal, and colloquial, often containing emotional and colorful language. Unlike a newspaper article, these dialogs give a strong sense of the individual expressing their opinion. Conversants engage in repeated discussion and get to know one another even though they have never met in person (e.g., 93% of the posts in `fourforums` are generated by 12% of the community).

PostID	Turn	Stance
S1-1:	Studies have shown that using the death penalty saves 4 to 13 lives per execution. That alone makes killing murderers worthwhile.	PRO
S2-1:	What studies? I have never seen ANY evidence that capital punishment acts as a deterrant to crime. I have not seen any evidence that it is "just" either.	CON
S1-2:	When Texas and Florida were executing people one after the other in the late 90's, the murder rates in both states plunged, like Rosie O'donnel off a diet.	PRO
S2-2:	That's your evidence? What happened to those studies? In the late 90s a LOT of things were different than the periods preceding and following the one you mention. We have no way to determine what of those contributed to a lower murder rate, if indeed there was one. You have to prove a cause and effect relationship and you have failed.	CON

Figure 2: Death Penalty discussion from `Convinceme.net`. Posts are linked using rebuttal links, showing poster's stance on the topic under debate.

The snippet in Fig. 2 gives a sense of the dialogic back-and-forth characteristic of this genre – questions (some rhetorical), targeted attacks on other discussants, and attention to *facets* of the debate (here, the deterrence facet of the capital punishment debate) in terms of particular propositions central to the discussion. This snippet also illustrates the implicit labels afforded by several discussion sites; in the case of `Convinceme`, participants explicitly mark dialogic links as rebuttals, providing a ready proxy for stance.

Fig. 3 provides further examples of the range of dialogic behaviors found in these online conversations. To start, Figs. 2 and 3 illustrate the frequent use of cues to mark discourse relations, e.g. comparison and contingency relations are marked by *When* in S1-2 and *If indeed* in S2-2 [110]. Dialog strategies also often include non-information seeking **questions**, which are intended to elicit responses by challenging another's evidence or assumptions: *And what is wrong with giving homosexuals the right to settle down with the person they love?* (R3 in Fig. 3). Utterances may also be strongly **emotional or highly rational**, e.g. contrast *What is it to you if a few limp-wrists get married in San Francisco?* (R3 in Fig. 3) with *It's not a literal account unless you read it that way* (R1 in Fig. 3). The language used is **social and non-literal**; irony and sarcasm abound, e.g., *Really? Well, when I have a kid, I'll be sure to just leave it in the woods, since it can apparently care for itself* (R5 in Fig. 3, see also see Q2 and R2). Insults are common: *But in reality your opinion is gibberish* (R3 in Fig. 3), and *Here come the Christians, thinking they can know everything by guessing, and commiting the genetic fallacy left and right* (R7 in Fig. 3). **Subjective dialog acts** such as sarcasm and insults targeted at other conversants and their opinions are very frequent. A pilot annotation study on 10,003 Quote/Response pairs from `4forums` indicates that about 12% of posts are sarcastic, 23% are emotional, and 12% are insulting or nasty, behaviors that are uncommon in traditional media [1, 142, 141].

Topic	Quote **Q**, Response **R**	**Stance**
Evolution	**Q1**: 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?	CON
	R1: It's not a literal account unless you interpret it that way.	PRO
	———	
	Q2: I jsut voted. sorry if some people actually have, you know, LIVES and don't sit around all day on debate forums to cater to some atheists posts that he thiks they should drop everything for. emoticon-rolleyes emoticon-rolleyes emoticon-rolleyes As to the rest of your post, well, from your attitude I can tell you are not Christian in the least. Therefore I am content in knowing where people that spew garbage like this will end up in the End.	CON
	R2: No, let me guess . . . er . . . McDonalds. No, Disneyland. Am I getting closer?	PRO
Gay marriage	**Q3**: 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.	CON
	R3: 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.	PRO
Abortion	**Q4**: Equality is not defined by you or me. It is defined by the Creator who created men.	CON
	R4: 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.	PRO
	———	
	Q5: The key issue is that once children are born they are not physically dependent on a particular individual.	CON
	R5 Really? Well, when I have a kid, I'll be sure to just leave it in the woods, since it can apparently care for itself.	PRO
Gun Control	**Q6**: How about a sin tax of $100 each time you buy a gun and $10 each time you buy a bullet?	PRO
	R6: How about a sin tax of $100 dollars each time you log on and $10 dollars a word for each time you speak one? It's fair because it would help pay for all the lying propaganda damage you do to society. Rights come with responsibilities. On guns, no can do. SCOTUS has ruled many times that a right freely stated in the Constitution cannot be compelled to purchase a license nor a fee to exercise. NEXT!	CON
Existence of God	**Q7**: okay, well i think that you are just finding reasons to go against Him. I think that you had some bad experiances when you were younger or a while ago that made you turn on God. You are looking for reasons, not very good ones i might add, to convince people.....either way, God loves you. :)	PRO
	R7: Here come the Christians, thinking they can know everything by guessing, and commiting the genetic fallacy left and right.	CON

Figure 3: Sample Quote/Response Pairs from `4forums.com` with Mechanical Turk annotations for Stance.

The information in the growing body of online opinion dialogs provides unique opportunities and challenges for the field. On the one hand, the sheer amount of data now available promises the possibility of better empirical understanding of dialog and its structure *at scale*. And given that an increasing portion of our collective knowledge and opinion is being encoded online in dialogic text, a host of aggregation and extraction technology (e.g., question answering, opinion mining, behavior prediction, and summarization) will have to contend with understanding the content in this genre sooner or later [162, 22, 13, 88, 87, 19, 96, 115, 26, 159, 126, 48, 122, 123, 156, 154, 117]. However, as we have worked on this data over the past few years, one basic point has become clear: the informal, adversarial, interpersonal, dialogic nature of this data renders it particularly vexing for the tools that have been built on the traditional monologic sources present in PTB [76]. Admittedly, work on subjectivity and sentiment has also focused on user-generated content, but very little to date has contended with dialog [156, 154, 162, 115, 26, 97, 159, 118, 107, 16, 133, 122, 123, 55, 57].

The sentiment literature is not unique in this: there has been little computational research on identifying the discourse and dialog relations within and between turns in informal conversation. There are no NLP tools for recognizing socio-emotional utterance categories such as sarcasm or insults [156, 32, 125]. There is limited work that aims at resolving chains of event reference in dialog as we see in Fig. 2 [41, 40, 130, 23, 91]. Computational models of dialog structure, in the main, aim only to account for the structure of task-oriented or tutoring dialogs [8, 52, 71, 128, 129, 23, 28, 108, 135, 72, 36, 53, 29, 82, 85, 35, 37, 84, 83, 86]. Previous work on summarizing dialog aimed only to extract phenomena specific to meetings, such as action items or decisions [90, 89, 47, 61, 153, 62, 24].

In short, we believe that in order to meaningfully extract information from these dialogs, we need better computational models of informal dialog. Our pilot studies involving over 102 million words of dialog (described below) suggest that developing better computational models requires that we first understand which aspects of opinion sharing dialogs people find most salient. We are using human summarization as a probe for exploring saliency, and we describe a pilot study suggesting that dialogic summaries orient at the highest tier to propositional content, but that attribution of beliefs to the stance holder and other social, subjective and dialogic properties are also reliably repeated elements in human summaries. These pilots also suggest we need to represent (proxies for) the underlying propositions, rather than just the terms associated with them because in our data, people on both sides of an issue often use the same terms. We also believe that it is critical to model socio-emotional properties of dialog, and model some aspects of dialog structure within and across turns [11, 149, 110]. Our research therefore focuses on the following specific **aims**:

- DIALOG CORPUS RESOURCE: provide our final corpus as a resource to the dialog community. This includes 1000 sample dialogue summaries with Pyramid annotations, as well as annotations of stance, subjective dialog acts, discourse relations and across-turn relations;
- SUBJECTIVE DIALOG ACTS: automatically identify subjective dialog acts such as sarcasm and insults and distinguish them from objective dialog acts;
- IDENTIFYING CENTRAL PROPOSITIONS: identify the propositions (abstract objects) that are under discussion that are central to a speaker's argument;
- STANCE: identify the speaker's stance in terms of whether the speaker is entrenched on one side or another of an issue, or is undecided, or holds a more nuanced middle position. Are two speakers agreeing or disagreeing? Are they presenting different arguments, or focused on the same arguments?

The goals of the proposed research are to produce computational tools that can characterize and select dialogs across these three aspects. Of these, the subjective dialog acts are the least well-understood, and to fill that in beyond sarcasm and insult, we will use human summaries of dialogic snippets like Fig. 2 to discover what dialog acts are most salient to readers. We will build data at scale using a cycle of human annotation, implicit labeling (like rebuttal labeling, abstract anaphora, and quotation), and bootstrap learning. This cycle is designed to maximize our annotation potential: some highly salient dialogic aspects are relatively rare (e.g., sarcasm at 12%), and bootstrapping will allow us to resample to more efficiently to find and annotate them. While the long-term goal of this research may be summarization or extraction, our final goal is to have a suite of tools that will allow us to select dialogs with specific properties. This can only be accomplished once we have trained the automatic classifiers of interest to a sufficient level of precision. For example, we expect in the long run it will be useful to a range of different applications to have classifiers for sarcasm and for stance, and algorithms that can identify central propositions of an issue (e.g. morality or deterrance arguments against the Death Penalty).

The Team. The proposed research team brings together a unique combination of interdisciplinary analyses and methods applied to computational dialogic processing techniques in social media. We have research experience in the analysis of social media, large scale online conversation and meetings (Whittaker), experimental studies of informal communication and the differential effects on participants and audiences, as well as the cognitive and affective effects of discourse markers and quotation (Fox Tree), linguistic analyses of the semantic, pragmatics and discourse structure of online conversation and chat (Anand), and computational modeling of informal dialog (Martell, Walker). This includes experience in annotating corpora and releasing it to the public to support other researchers (Whittaker: AMI corpus, Anand: Internet Argument Corpus, Martell:NPS Chat Corpus, Walker:DARPA Communicator Corpus).

2 Background and Theoretical Orientation

Recent work, including our own, has shown that processing these online conversations is challenging [122, 123, 133, 142, 141, 97, 68, 64]. This is because these dialogs exhibit many contextually-dependent and subjective properties. In this section, we first describe our extensive efforts to systematically create and curate a **dialog corpus** with a range of different properties. We have already made a subset of this corpus public [140]. Then, because of the relative novelty of this area, we describe a **pilot study** on this data, which we use to motivate our research goals and to scope the **relevant theoretical** literature. We then summarize relevant prior work.

Data Source	Dialog Affordances Available from Site	# Dialogs	Word Count
Convinceme	**Rebuttals, Topic, Stance**	3072	3,650,000
4forums	**Reply links, Quoting, Debate Topic**	11,800	84,300,000
CreateDebate	**Support, Oppose and Clarification Links, Stance, Quoting, Topic**	11,876	15,000,000

Table 1: Dialog sources in our corpus: with types of discussion, available affordances from the site, media types and size of the resource.

Dialog Corpus. Table 1 lists the dialog sources in our corpus along with some of their affordances. To date, our corpus contains a total of 26478 dialogs from `Convinceme`, `4forums` and `CreateDebate` consisting of 102 million words. Collecting data from multiple sites increases the number of dialogs per topic, but there are also differences in dialog structure and in meta-information across different sites (see the affordances column listing dialog metadata in Table 1). These site specific affordances provide rich participant supplied dialog annotations that can be exploited in our research. For example, an important type of affordance from a site is the support provided by `4forums` for **quoting** another person's post: See the Q/R pairs in Fig 3. We included 10,000 Q/R pairs in our inital release of our dialogue resource because Q/R pairs represent critical dialogic information, manifesting an explicit effort by conversants to focus their response on a particular aspect of another's turn [144]. Because the site's layout encourages quoting, 72.3% of all posts contain at least one quote. Our initial analysis also shows that these dialogues are rich in explicitly signalled discourse relations. For example, 8,731 Q/Rs used in one of our pilots included the following counts of discourse markers in turn-initial position: *well* (605), *and* (582), *so* (583), *actually* (322), *but* (265), *oh* (228), *I think* (182), *because* (162), *just* (113), *really* (100), *I believe* (69), *I know* (59), *you know* (49), *you mean* (50), *I see* (47), *I dunno* (10), and *you think* (7), as well as 233 instances of *yes*, 220 instances of *no*.

TOPIC	D	POSTS	REBUT %	P/A	A > 1p	CHARS
Cats v. Dogs	3	162	40%	1.68	26%	242
Firefox vs. IE	2	233	40%	1.28	16%	167
Mac vs. PC	7	126	47%	1.85	24%	347
Superman/Batman	4	146	34%	1.41	21%	302
2nd Amendment	6	134	59%	2.09	45%	385
Abortion	10	607	70%	2.82	43%	339
Climate Change	6	207	69%	2.97	40%	353
Communism vs. Capitalism	6	207	70%	3.03	47%	348
Death Penalty	12	331	62%	2.44	45%	389
Evolution	16	818	76%	3.91	55%	430
Exist God	16	852	77%	4.24	52%	336
Gay Marriage	6	560	65%	2.12	29%	401
Healthcare	5	112	80%	3.24	56%	280
Marijuana Legalization	5	236	52%	1.55	26%	423

Figure 4: Some properties of `Convinceme`. KEY: D = number of dialogs. POSTS = total posts across all dialogs. REBUT% = percentage of posts linked together into a rebuttal chain. P/A is average posts per author. A > 1p = percentage of authors with more than one post. CHARS = average characters per post.

`Convinceme` provides a richer type of dialogic system affordance allowing participants to indicate that their stance on a prior post is a REBUTTAL as illustrated in Fig. 2. Long chains of rebuttals are common, and these are useful for exploring sequential information about disagreement. CreateDebate also supports inbuilt rebuttals (called oppose links) but also provides two other types of explicit argumentation links: support and clarification. To date, we have only scraped CreateDebate and created a database of these explicit links, i.e. we have not yet begun to explore how we can use these argumentation links. Another important determinant of dialog properties can be seen across topic. Table 4 provides more detail for `Convinceme` of how dialogic properties and number of dialogs vary by topic. Table 4 separates idealogical topics (below the line) from playful (Cats vs. Dogs) and technical (Firefox vs. IE) topics [123]. The Rebut% column in Table 4 indicates that idealogical topics tend to be more highly threaded and contested, perhaps because the participants are

more invested in the discussion. Compare the 34% rebuttal rate for Superman vs. Batman to the 80% rebuttal rate for Healthcare. Posts per author also vary, ranging from a low of 1.28 for Firefox vs. IE, up to 4.24 posts per author for Existence of God, again possibly revealing strong personal involvement. These differences by topic have implications for algorithms for stance side classification and agreement recognition.

Pilot Study. As we have been conducting research on these dialogs over the last several years, one of our fundamental questions has been what aspects of these dialogs people are oriented to. They are clearly different from traditional media as exemplified in Figs. 2 and 3. But in what ways? Our intuitions suggested that the attributes indicated in Figs. 2 and 3 were an important part of what makes these dialogs engaging. To follow this up more systematically we conducted a pilot study on dialog summarization. Summarization provides an efficient, open-ended mechanism for determining the central aspects of a dialog: at the individual level, each reader conveys what they subjectively deem the central or most salient aspects of the dialog. In addition, summarization is something that any native speaker can do; it does not require training. Thus, as an annotation of salience it is highly amenable to crowd sourcing. Finally, once a set of reference summaries are collected, the pyramid evaluation scheme provides a useful way to calculate global salience, as well as reliability across human perceptions [93, 92].

If asked to summarize these dialogues, what would people decide was important to mention? Is it their dialogic nature? Is it the fact that they are social? Is it that they provide multiple perspectives on an issue? Is it because they are rich in subjective elements?

The content of opinion sharing dialogs can be divided into three types. At a global level, it is a *social* interaction between two or more individuals. But it is also a dialog, a sequence of utterances, linked to particular speakers with particular *dialogic* goals. Finally, at the *propositional* level, these discussions are stand-ins for the comparisons between the larger group of people who hold a stance in the general issue. We hypothesized that summarisers of these debates would be sensitive to the social, dialogic, and propositional levels in constructing their summaries, and this was borne out in the pilot, where there was also strikingly high agreement among the 10 pilot subjects about core facets of the discussion and the elements that they focused upon.

PostID	Turn
S1-1:	Bush has been eagerly desperate in injecting US military troops to middle eastern countries which lead to speculations on the 9/11 being a conspiracy. He poured billions expecting to get bigger from the availability of oil in the countries by which he (undoubtedly) exploited.
S2-1:	Only a stupid fool believes 9/11 was a consiracy.
S1-2:	Well there are a handful of "stupid fools" out there, and you can't force your ideology on them
S2-2:	I'm not trying to force my ideology on anyone. People should be free to be stupid fools.

Figure 5: Dialog #5 from Summarization Pilot. Posts linked with rebuttal links.

We gave 8 short dialogs between two participants to 10 subjects and asked them to: *Imagine you are trying to summarize the following dialogs for a friend. Please read each of the following 8 short dialogs carefully, summarizing the main points in your own words at the end of each dialog*. An example dialog is in Fig. 5. The 8 dialogs covered a range of topics that are representative of different types of online discussions: evolution, political TV shows, abortion, musical taste, Middle East politics, mind-brain dichotomy, Russian foreign policy, a soccer player's transfer between countries. Our dialogs were chosen to have variable levels of agreement between the participants (from directly opposed to largely agreeing) and levels of commitment on the parts of the participants (from strongly held views to ambivalent). Dialogs were also chosen to include social language including humor *YAY World War 3!!!!!!!!!!* (in the context of Russian foreign policy), emoticons ('lol'), strong language including insults *Only a stupid fool believes 9/11 was a conspiracy*, quoting, and affiliative comments *just joshin' ya Americans :]*. Our analysis of the summaries supported our expectation that listeners would be oriented to 'content' properties that are not strictly propositional. The LHS of the pyramid annotation in Fig. 6 enumerates the SCUs (summary content units) that we found in our pyramid style analysis of the summaries produced by the 10 pilot subjects for the dialog in Fig. 5. While we can neither include all the original dialogs, nor their summaries, due to lack of space, we see consistent behavior across all the pilot summaries. The summaries almost always talk about which conversant hold which stance and whether the speakers agree or disagree about particular points. In other words, as expected, the summaries point out particular beliefs (central propositions) but importantly, these are attributed to someone.

At the social level, summarizers were sensitive to the quality of interpersonal interaction of participants, commenting on levels of coordination (*building on each other* and *shar[ing] in person one's frustration* vs. *The rest of the discussion is an attempt between the speakers to decide exactly what Paul's position is* vs. *arguing* and emotional states of participants during the discussion (*getting defensive* and being *taken aback*, being *in shock* and *impressed*). See Fig. 6. Summarizers also paid attention to the dialogic structure of the discussion in three ways. First, summaries included explicit reference to who said what; fully 90% of summaries included 'play by play' devices characterizing the utterances in temporal succession; see Fig. 7 for one for the discussion in Fig. 5. Moreover, summarizers frequently went beyond the text itself to characterize participant emotions and subjective characterizations of dialog acts such as *Person 2 is skeptical*, The other speaker is more concerned with, *The first user again confronts*, *the second person taunts the first one*, *The second poster tries to make light of the situation*) as well uses of adverbs like *sympathetically* and *sarcastically* which can be found in Fig. 6, and reactions to other's dialog acts, e.g. *defensive*. We noted many instances of evaluative language by the summarizers towards the participants' conduct (e.g., calling something a *pointless argument* or saying that a participant *couldn't say anything else to complete his argument, so he just makes sarcastic comments*. We also noted that concession seemed to be highly salient when they occurred.

SCU	Used by summarizer?										Tot	Tier
	A	B	C	D	E	F	G	H	I	J		
a. why some feel 9/11 was a conspiracy	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10	6
S1 suggests (a)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10	6
b. only stupid people believe that	✓	✓	✓	✓		✓	✓	✓		✓	8	5
S2 says (b)	✓	✓	✓	✓		✓	✓	✓		✓	8	5
c. people have the freedom to be stupid			✓	✓		✓	✓			✓	5	4
S2 states (c)			✓	✓		✓	✓			✓	5	4
d. Bush expected profit in Middle East		✓	✓		✓		✓				4	3
S1 says (d)		✓	✓		✓		✓				4	3
e. you can't force your beliefs on others			✓	✓		✓	✓				4	3
S1 argues (e)			✓	✓		✓	✓				4	3
f. 9/11 truthers			✓							✓	2	2
g. Bush injects US army to Middle East				✓			✓				2	2
S1 says (g)				✓			✓				2	2
h. many truthers out there				✓						✓	2	2
S1 says (h)				✓						✓	2	2
S1 criticizes Bush			✓				✓				2	2
S1 sympathetic to truthers	✓					✓					2	2
S2 rejects S1's arguments								✓		✓	2	2
S2 replies sarcastically				✓			✓				2	2
i. those people are still stupid	✓		✓								2	2
S2 says (i)	✓		✓								2	2
S1 gets defensive						✓				✓	2	2
j. argument re: troops in Middle East				✓	✓						2	2
k. Bush's war wasted money				✓							1	1
S1 says (k)				✓							1	1
l. Bush pushing for war		✓									1	1
S1 says (l)		✓									1	1
S2 misspelling: consiracy	✓										1	1
S1 doesn't agree with Bush's war tactics						✓					1	1
m. stupid fools do exist							✓				1	1
S2 further taunts S1							✓				1	1
n. 9/11 is definitely not a conspiracy									✓		1	1
Σ SCUs	8	8	14	16	5	11	16	5	3	11		
Recall	.25	.25	.44	.50	.16	.34	.50	.16	.09	.34		
PYRAMID SCORE	.80	.83	1.0	.93	.77	.93	.96	.92	.59	.91		

Figure 6: Results of Pyramid on Bush Dialog.

At the propositional level, many summaries contained explicit statements of the CENTRAL PROPOSITIONS on the table (e.g., *S1 and S2 discuss geopolitics regarding Russia and the threat they pose.*). They also frequently convey the STANCE that participants bear towards the propositions that are introduced by their fellow participants (e.g., *The second speaker then takes up the first speaker's point that*), often characterizing people in terms of their position (*The pro this-is-evolution speaker does not accept these points as valid arguments*); only one summarizer consistently omitted any reference to the participants (e.g., *Some people believe that 9/11 was a conspiracy because they think Bush is just trying to control the oil market by sending in tons of troops.*).

> Person 1 criticizes Bush for exploiting other nations for oil, and explains how his actions have led to 9/11 conspiracy theories. Person 2 states that only idiots are 9/11 truthers. Person 1 argues that you can't force your beliefs on others. Person 2 states that people have the freedom to be stupid.

Figure 7: Sample 'play by play' summary for the dialog in Fig. 5

To more systematically index the salient elements of an opinion sharing dialog, we conducted a pyramid evaluation on two of our discussion summaries. Based on our three-level distinction, we identified mentions of propositions, attributions and stances toward propositions, mentions of dialogic and social relations, and summarizer editorializing as separate SCUs. The bottom of the pyramid table for the debate in Fig. 5 in Fig. 6 shows clearly that, quantitatively, the results are promising. While one of the reasons we did the pilot summary was our uncertainty what people would do, we found that 100%, 80%, and 50% of participants agreed on the top three tiers of SCUs. All of these

were attributed statements of propositions; the social and dialogic features of this conversation were less agreed upon (all but one had weight 2 of a possible 10), but only 2 summarizers failed to mention any social or dialogic features of the text. This suggests that while such features are clearly perceivable, the salience of particular features will be listener-centric as one would expect [78, 102]. Still, individual pyramid scores were high (0.83 and above for six participants), indicating good correspondence between the propositions chosen by different individuals and the overall weighted mean of propositions chosen by all summarizers.

Relevant Theoretical Background. Before our pilot we wondered whether social psychology theories of persuasion were relevant to our research. Theories of persuasion identify three main factors affecting whether and how persuasive messages change audience beliefs: (1) the ARGUMENT, i.e. the propositions discussed and the discourse and semantic relations between propositions; (2) the AUDIENCE and their prior beliefs and social identity; and (3) the SOURCE (speaker) of the argument such as whether the source is expert, attractive, powerful, trustworthy or credible [17, 39, 63, 18, 25, 80, 77, 27, 102, 101, 78, 59, 131, 95, 43, 10, 33, 79, 101, 34, 139, 5]. While we are not interested in belief change *per se*, it seems plausible *a priori* that these same factors are perceivable to readers of opinion sharing dialog in general, independent of whether reading a dialog triggers belief change. The pilot study supports our intuition, where the factors considered important in theories of persuasion appeared in the highly reliable summaries produced in our pilot. Thus summaries made reference to the argument itself, the properties of the audience as well as to the attributes of the person making the argument.

It is also clear that previous work on subjectivity is highly relevant. Work on subjectivity is inspired by work in linguistics and literary theory that focuses on how opinions and emotions are expressed linguistically in context [156, 157, 160]. The goal is to identify and characterize expressions of private states. Private state is a general covering term for opinions, evaluations, emotions, and speculations. For example, the utterance *You have to provide a cause and effection relationship and you have failed* (S2-2 in Fig. 2) expresses a negative evaluation of S1's reasoning. The phrase *Here come the Christians, thinking they can know everything by guessing and committing the genetic fallacy left and right* (**R7** in Fig. 3) also reflects the private state of the speaker, and a negative view of *Christians*. Phrases such as *fallacy* are clearly subjective, but also emphasizers like *left and right*. Thus, subjectivity cues such as experiencer verbs, adjectives, and subjective nouns will clearly be useful for identifying subjectivity in our dialogs [67, 138, 155, 157, 15, 14, 158, 163, 58, 156, 155]. However our dialogs contain what we call SUBJECTIVE DIALOG ACTS, novel types of subjectivity that occur primarily in dialog such as sarcasm and insults directed at the other participants. Moreover, our dialogs contain numerous expressions of subjectivity that would be unlikely to occur in newspaper articles such as rhetorical questions whose goal seems to be express incredulity at something the other conversant has said *That's your evidence?*, and various cue words indicating emotional state such as *Oh* and *Really?* that we see in Figs. 2 and 3, as documented in previous research by PI Fox Tree [45, 21, 44]. Consequently, we hope that the large scale nature of our corpus will support bootstrapped learning of some of these cues to increase the breadth and comprehensiveness of resources for identifying subjectivity [118, 158].

In addition, we hope to systematically examine many types of dialog behaviors that have been previously studied in a limited way. Most work on dialog structure has focused on task oriented dialogs, where the participants have a shared goal. That does not seem to be the case in our data. In our dialogs, event reference chains are frequent as previously noted for advice dialogs [143], SwitchBoard [41], and TRIPS planning dialogs [23]. We see one example of this in in Fig. 2. We thus expect to be able to draw from theories of event reference [148, 147, 11], that suggest that deictic references to abstract objects select antecedents on the "right frontier " of the discourse, or the most recently completed dialog act or " synchronizing unit" or "grounding unit" [41, 23]. We believe discourse deictics will be a critical cue for identifying central propositions as we discuss in Sec. 3.2. Previous work on event reference also suggests how to use predications on deictics, i.e. *That's right* or *That's true* to restrict the range of possible referring functions that can be used to coerce the "proxy" (the actual linguistic sentence or the verb phase), to the event, state, proposition, or fact that is the abstract object evoked by the deictic.

We also expect our work on dialog structure to draw on research on discourse relations both within and across turns [74, 52, 161, 106, 12, 146, 30, 104, 116]. Relations of COMPARISON and CONTINGENCY are common in our data, and there are many uses of explicit discourse cues for these relations [110]. Again, previous studies of the uses of discourse relations in dialog have been very limited [127, 7, 134] and have often simply focused on identifying agreement or disagreement [47, 133, 16, 1].

3 Research Plan

The overall research method is shown in Fig. 8. We have already begun the development and curation of a large scale corpus of opinion dialogs (26,748 dialogs totaling 102M words), and collected three types of annotations as depicted at the **top** of Fig. 8: (1) annotations that we can download from the site as meta-information listed in Table 1; (2) crowdsourced annotations we have collected in pilot annotation studies including the dialog summaries we discussed above in our pilot summary study; and (3) annotations for what we are calling implicit markup dialog behaviors, such as the use of discourse markers, event reference and quoting. See below. Our pilot studies have been carried out on the subset of topics listed in the first column of Table 4.

The middle of Fig. 8 illustrates the next step, processing the corpus using off-the-shelf tools and our own tools to extract syntactic and semantic structures and feature representations from the dialogs and their annotations. We then carry out machine learning experiments with a range of features that we posit are predictive of the dialog phenomena that we wish to automatically identify. Based on the pilot summary task and our previous work in agreement and stance classification, we know that it will be important to be able to identify: (1) subjective dialog acts such insult and sarcasm; (2) propositions that are central to the discussion; and (3) evaluative stance. We believe that these goals require identifying some discourse relations within and across conversational turns, as we discuss below. In addition, we will add to this list based upon the most commonly salient SCUs in our full-scale summarization task.

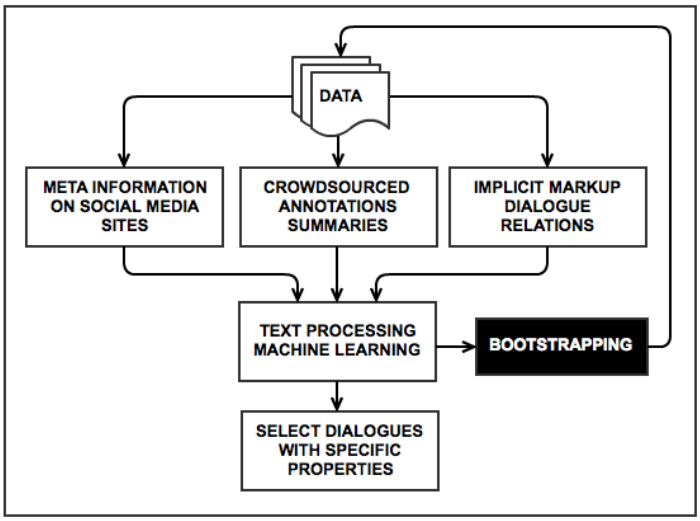

Figure 8: Overview of Research Plan.

Having identified these elements, starting with some crowdsourced annotations, it might seemt natural to pursue a traditional supervised learning approach – annotating for and training a classifier for each element. However, our goal in the proposed research is to be able to recognize these salient elements at scale, across a variety of topics and argumentative styles. Our previous work on stance detection has demonstrated that the optimal feature set varies widely across topic [142, 141], and so we do not believe that conventional machine learning mechanisms will adequately allow us to train classifiers on a narrow range of topics and generalize, at scale, to a variety of novel domains. Instead, we propose to achieve results by bootstrapping from relatively high-precision signals present in the dialogs themselves to fuller coverage over the dataset as a whole, across hundreds of topics, as depicted in the bootstrapping box in the RHS of Fig. 8. Below in Sec. 3.1, we provide a more detailed description of the bootstrapping process we have piloted and describe our initial results bootstrapping classifiers to recognize sarcasm.

Our final goal, shown at the bottom of Fig. 8 is to have a suite of tools that will allow us to select dialogs with specific properties. This can only be accomplished once we have trained the automatic classifiers of interest to a sufficient level of precision. For example, given classifiers for sarcasm and for stance, and algorithms that can identify central propositions of an issue (e.g. morality or deterrance arguments against the Death Penalty), then we expect in the long run that it will be useful in applications of this work to be able to automatically find a highly sarcastic argument representing the PRO side of the *deterrence* facet for the Death Penalty topic. We believe that this is a pre-requisite program of research to any possibility of automatic summarization of opinion sharing dialogs. Below, in Sec. 3.1 we first describe the types of subjective dialog acts that we find in our corpus and our research plan to develop ways to detect them. Then we describe our planned research on identifying central propositions in Sec. 3.2, and finally in Sec. 3.3 we describe our research plan for identifying participants' stances toward central propositions.

3.1 Subjective Dialog Acts

To illustrate the planned flow of our research, we first provide a worked example of the bootstrapping component shown collapsed on the RHS of Fig. 8. We have been carrying out a pilot study of boostrapping for sarcastic dialog acts using the model in Fig. 10, adapted from the work of Riloff & Wiebe to reflect the unique properties of our data and our focus here on sarcasm [118, 132]. The top of Fig. 10 assumes that a corpus of opinion dialogs such as those in Figs. 2 and 3 is available. The left circle reflects the assumption that there are linguistic cues that can identify the category of interest with high precision (this is called "Known Subjective Vocabulary" in [118]). In the case of sarcasm, there is no "Known Sarcastic Vocabulary", nor is there annotation for sarcasm already available. In addition, sarcasm is context dependent in a significant percentage of cases, and it is not a unitary phenomenon, e.g. it includes jocularity, understatement and hyberbole [50, 42, 21].

Category	Annotation Question	Percent
Fact/Emotion:	Is the respondent attempting to make a fact based argument or appealing to feelings and emotions?	34/23
Respect/Insult:	Is the respondent being supportive/respectful or are they attacking/insulting in their writing?	
Nice/Nasty:	Is the respondent attempting to be nice or is their attitude fairly nasty?	49/12
Sarcasm:	Is the respondent using sarcasm?	12

Figure 9: Mechanical Turk Pilot Annotations. The final two columns provide counts for the Q/R and the P1,P2,P3 subsets of the 4forums data.

So we start with an annotation study of two types of 4forums data: (1) quote/response pairs such as those in Fig. 3, and (2) threads consisting of three posts in sequence (referred to as P1,P2,P3 dataset). For 10,000 posts in each dataset, we asked 7 subjects to indicate their perceptions of the respondents' dialog intentions using the annotation questions in Fig. 9. All questions elicited scalar responses on a scale from 1 to 7 in order to reduce noise [121], except for the sarcasm question, which asked for a binary response. Fig. 11 provides posts judged as having very high or very low socio-emotional language. Annotation reliability α values ranged from .22 for sarcasm to .46 for nice/nasty. While these α values are low, the data is still useful because of the many annotations. For our experiments, we use examples that pass a high agreement threshhold, e.g. 3 out of 7 annotators said the turn was sarcastic. The pilot annotation suggests that about 12% of the utterances are sarcastic. If we could bootstrap an improved sarcasm classifier that could select 30% sarcastic utterances on unannotated data, we could collect a large scale corpus of sarcastic utterances, in all variations, more efficiently. Previous empirical work on automatic identification of sarcasm has depended on the use of the `#sarcasm` tag in Twitter, or the assumption that all articles are sarcastic on news sites such as *The Onion* [32].

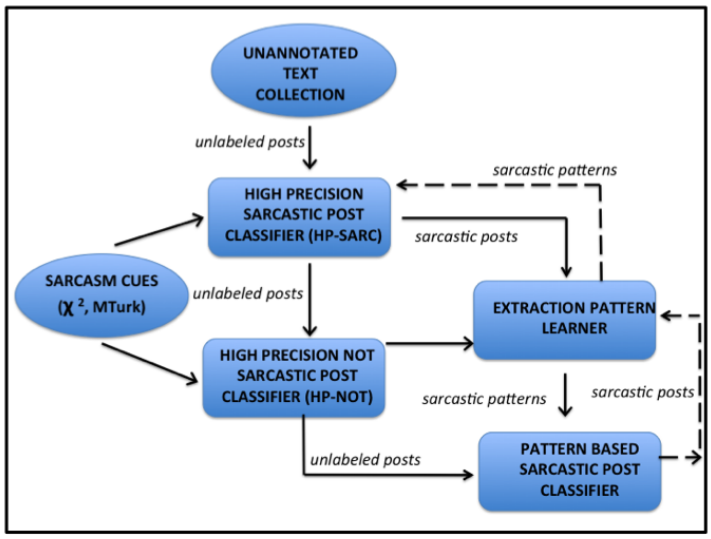

Figure 10: Bootstrapping Flow

Annotation	Very High Degree	Very Low Degree
Insulting, Respectful	Well, you have proven yourself 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.	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...
Sarcastic, Not-Sarcastic	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.	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.
Emotional, Rational	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.	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 Zealand).

Figure 11: Sample Responses selected as extremes on the Insult, Sarcasm, and Fact/Feeling spectrums

An initial classification experiment on the Q/R data using a range of features (ngrams, generalized opinion dependency features, cue words, context etc) and learning algorithms, achieves disappointing classification accuracies ranging from 53% to 60% over a baseline of 50%. Riloff & Wiebe's method (Fig. 10) first develops a very high precision classifier using multiple highly reliable cues for the classes of interest (here sarcastic vs. not sarcastic). Then the utterances identified by the high precision classifier are used to generate more general features using a general set of syntactic patterns. These more general features are tested on the **training** set and only those instantiations with high precision are retained. These retained features are then used on an unannotated corpus to increase the size of the dataset and improve the classifier. See Fig. 10. Because we don't have a prior "Known Sarcastic Vocabulary" we have piloted two different methods for discovering the initial 'vocabulary', and begun to experiment with combinations of cues that might be helpful to create a higher precision classifier. The first method simply takes the utterances annotated as sarcastic, extracts unigrams, bigrams and trigrams and then applies χ^2 feature selection to determine which ngrams are the most reliable predictors of the sarcasm class. The second method involves a second annotation round on previously annotated sarcastic utterances where we ask annotators to explicitly mark the cues to sarcasm by selecting words and phrases.

This second method is closer to the use of the MPQA corpus by Riloff & Wiebe to determine which cues are strongly subjective [160]. However, we were unsure as to whether Mechanical Turk annotators would be reliable at selecting cues and whether this task would make sense to them [6]. An initial collection of 20 annotations suggested that there might be a lot of variation across annotators. Sarcasm is also known to be highly variable in form, and to depend on context for its interpretation in some cases [124, 50, 21]. Therefore for this first pilot study we collected 100 annotations for 100 sarcastic posts. Figs. 12 shows a plot of average annotator agreement (ITA) as a function of the number of annotators, computed using Pearson correlation counts on the unigrams, bigrams and trigrams that were selected for markup. We show the results up to 40 annotators, but for all cases (unigrams, bigrams, trigrams and ngrams), ITA plateaus at around 20 annotators and is about 90% with 10 annotators, suggesting that in future annotation studies 10 annotators would be sufficient to get highly reliable results.

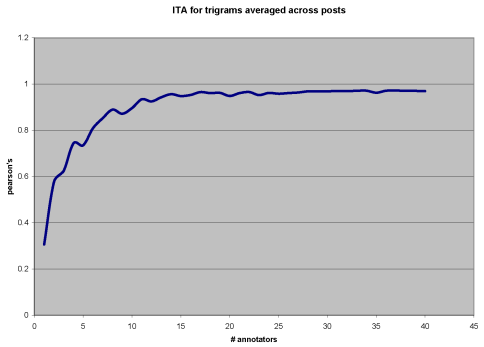

Figure 12: Interannotator Agreement for sarcasm trigrams with 40 annotators

The cues we learn from the annotation method are shown in Fig. 13. Like Riloff & Wiebe, what we are looking for are cues that will give us very high precision at the expense of possibly low recall in an initial High Precision Classifier. See Fig. 10. The aim of the bootstrapping steps: Extraction Pattern Features and Pattern Based Classifier, is then to increase recall while keeping precision high. Our initial results suggest that, although sarcasm is less well studied than subjectivity, and may have less reliable cues, that combining multiple cues into an initial High Precision Classifier should allow us to increase our accuracy for sarcasm classification. For example, Fig. 13 suggests that developing Extraction Pattern Features from dialog turns that include multiple cues such as *Oh yeah* and *just* or *and, oh* and *you mean* could lead to a higher precision sarcasm classifier. We are also applying this method to the *Nasty/Nice* annotation question in Fig. 9. These experiments are still in progress, but we consider these initial results to be highly promising.

indicators	FREQ	%SARC
{oh yeah}	11	55%
{oh}	337	29%
{you mean}	135	21%
{just, oh yeah}	3	67%
{and, oh yeah}	9	56%
{oh, like a}	7	43%
{and, that, oh yeah}	6	67%
{and, oh, you mean}	6	50%
{oh, like, you are}	16	38%

Figure 13: Selected sarcasm cues from annotators, their frequency and accuracy

In addition to this bottom up approach, we plan to explore whether, as in the case of subjectivity, we can use the theoretical literature to come up with an initial set of high precision cues for other types of subjectivity in dialog [20, 94, 114, 46, 66, 60]. For example, PI Fox Tree's previous work, and an analysis of the discourse markers in our annotations suggested positive correlations between *Oh* and emotional words from LIWC [100], and between sarcasm and discourse markers *you mean, oh, really, so*, and *I see*. PI Fox Tree also found a negative correlation between sarcasm and the markers *I think*, *I believe*, and *actually*. The

only previous work on insulting behavior that we are aware of is Spertus' research on identifying flames in email, which used a combination of theoretically motivated observations of flames and learning on a corpus [125]. We believe it will be possible to bootstrap the recognition of insults using lists of politeness forms such as hedges and tag questions [20, 94, 114, 46, 66, 60], or by writing patterns by hand to recognize some types of name-calling: these are often definite referring expressions like *The Liberals* or *The Christians*. Other potential cues for subjectivity include written affective markers such as punctuation e.g. exclamation points and asterisks; [136, 152] and discourse markers such as *oh*, *actually*, and *really* that are proposed to lead to both positive and negative emotional inferences [2, 3, 4, 38, 120], including argumentativeness [119], or ironic intent [54, 152].

3.2 Identifying Central Propositions

Clearly, to understand an opinion sharing dialog, it is necessary to recognize the propositions **central** to the discussion, as well as the relations between central propositions. We assume that central propositions are both those which "drive" the discussion, serving as repeated targets for anaphoric reference, subjective predication, and restatement and those which are perceived as most salient by readers. Previous computational work suggests a number of methods for identifying the targets of subjectivity or opinions [160, 138, 99, 56, 107]. This includes finding frequent collocations, or identifying the target using evaluative predicates and patterns defined on syntactic trees or POS tags [55, 57, 123]. However, in most previous work, these targets are realized by named entities or noun phrases such as people (Clinton, Obama, Romney), products (Firefox, IE), or physical amenities such as comfy beds or free breakfasts [98, 138, 115, 26, 159]. In our dialogs the targets are frequently abstract objects such as propositions or event descriptions, that can be referred to with deictic pronouns [11, 149, 65]. See Fig. 2. We are not aware of any other work that attempts to automatically identify the abstract objects evoked in a dialog as the targets of opinions. Moreover, we are not aiming to identify all of the abstract objects, rather we hope to identify the central ones.

How can we determine which propositions are most central to the dialog? Our approach will be to identify proxies for central propositions from two different types of data as illustrated in Fig. 8. First, as we assume central propositions are the most salient to readers, we plan to collect and utilize a small corpus of dialog summaries (see **Summary Corpus Collection** below) Second, because we assume that central propositions drive the discussion, we propose a novel idea of exploring the use of **Explicit Dialog Behaviors**, i.e. the behavior of the dialog participants themselves as a type of implicit markup of participants' subjective notion of the centrality of propositions. There are two kinds of behavior we can consider as implicit markup: (1) discourse cues marking relations of COMPARISON and CONTINGENCY and their arguments; (3) contextual cues in dialog such as uses of event reference and uses of the verbatim quoting mechanism supported by the site.

Summary Corpus Collection. Our interest in summarization is because we believe that summarization provides an efficient, open-ended mechanism for determining the central aspects of a dialog, and summarization is something that any native speaker can do [93, 92]. We believe that collecting dialogic summaries will focus our computational work on the important aspects of our dialogs and provide a resource for future work. We also believe that it is a necessary pre-requisite to meaningful work on summarization of dialog. In addition, the pilot results discussed at length Sec. 2 demonstrate that Pyramid evaluation of dialog summaries is a **highly reliable** and therefore promising approach to indexing the salience of propositions, because Pyramid SCU annotation is based on repeated key elements [93, 92]. We will expand our summary corpus by extending our pilot.

We will conduct a series of summary collection experiments on Mechanical Turk over the first two years, so that at the end of these two years we have collected a total of 1000 pyramid annotated summaries. See Timeline and Budget Justification. Our main goal is to expand on the pilot with a broader sample of dialogs and more participants. We predict that this will solidly document the key phenomena emerging from our pilot that the top tier of the SCUs are the propositional content, but that attributions to the belief-holder and social and dialogic aspects of the dialogs are also well represented in the lower tiers. Stimuli will be selected to include dialogs that contain at least four turns per contributor (minimum 8 turns), that present at least two differing perspectives on an issue, and that contain at least 6 dialogic features and at least 4 expressions of evaluative stance. Dialogic features may include uses of particular discourse cues, using punctuation such

as capitalization, asterisks, exclamation points, or ellipses, name calling, sarcasm, and strong agreement vs. disagreement. Since we are going to iteratively collect summaries, analyze them, and then collect more, we plan to iteratively focus our summary corpus collection on properties that look to be the most interesting as we proceed. Initially we expect to sample for highly dialogic and context dependent dialogs and use as a control a few instances of monologic posts that we see in some instances on `Convinceme`, and we hope to explore the role of quoting, event reference and discourse markers. We believe that this data collection will inform the three key computational tasks that we plan to undertake in this project. It will also help future researchers to determine: (1) what a shallow representation of the dialog must contain to support indexing of dialogs by meaningful properties; and (2) how a future system might compress the original dialog into something more summary-like.

Explicit Dialog Behaviors. We are interested in investigating whether the behavior of the dialog participants themselves are useful for identifying the central propositions of the discourse. We will exploit three different types of explicit dialog behaviors: use of explicit discourse cues, discourse deixis and verbatim quoting.

We expect lexicalized explicit discourse markers of the CONTINGENCY and COMPARISON discourse relations, such as *Because, So, If-Then* [110], to be useful for identifying central propositions [74, 52, 161, 106, 12, 146, 30, 104, 116]. These are highly frequent in our corpus: e.g. there are 1187 examples of turn initial *Because* in a sample of 10,000 Q/R pairs from `4forums`. Arguments of discourse relations are abstract objects [11, 148], such as events, states and propositions, for which the discourse relations hold. Recognize discourse relations involves the following subtasks: (1) identify discourse uses of connectives, (2) identify the arguments of discourse connectives, (3) identify the senses (i.e., semantics) of the relation. For example, we posit that it could be useful to recognize automatically that the relation CONTINGENCY connects the two propositions in Fig. 2: *Studies have shown that* **using the death penalty saves 4 to 13 lives per execution**. **That alone makes killing murderers worthwhile**. As in other work on recognizing discourse relations, we plan to use the highest level PDTB relations, and apply shallow discourse parsing to identify the arguments of the relations. We will also focus exclusively on explicitly signalled discourse relations since accuracies are much higher when an explicit connective is used [81, 105, 69, 151, 164, 73, 70, 150, 145, 103, 110, 109, 111, 112, 49, 69, 70]. PDTB discourse relations may also be useful for recognizing which propositions are the focus of discussion **across speakers** [7, 127, 134, 47, 133, 16, 1]. One suggestive piece of evidence is that our `4forums` corpus has high frequencies of turn-initial discourse connectives in **reply** contexts, suggesting cross-speaker discourse relations similar to those that occur within a speaker turn, e.g. *Because*, *So* and *But*. See Table 2. Contrasts may be used to portray the opponent's position in an unflattering light, while concessions may be used to acknowledge a fact but dispute its importance, in light of other facts about to be presented, e.g. *Sure most communism has failed,* **however** *this is not due to a fundamental flaw in the theory,* **rather** *it is due to a flaw in implementation.* This concedes that most communism has failed, but denies the conclusion of the opposing person with contrastive markers of *however, rather*.

Comparison:Contrast/Concession
Sure most communism has failed, **however** this is not due to a fundamental flaw in the theory, **rather** it is due to a flaw in implementation.
The processes underlying natural selection are not set in place by minds (divine or human) **but rather** by the operation of nature via (ultimately) physics and chemistry.
You prefer to cling to medieval mythology **rather than** accept science.
The problem being that it is forced on students as the truth **rather than** a possibility.

Table 2: Examples of contrasts and concessions from 4forums.com pivoting on terms involving "rather". The bolded terms establish the relevant function.

Because both psychological research on discourse processes [45, 46, 51] and computational work on agreement [47] indicates that discourse markers are strongly associated with particular pragmatic functions, we conducted a pilot on our `4forums` corpus to examine how turn-initial markers may predict upcoming content [46, 51]. Based on manual inspection of our 20,000 post subset of `4forums`, we constructed a list of 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*. 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%). Based on this pilot, we expect our corpus to provide a rich resource for further work on the role of discourse markers and discourse relations. Moreover we expect to see high frequencies of relations that are specific to dialogic discourse.

However such explicit signalling is not the only indicator of centrality. We also hypothesize that **contextual** devices such as ABSTRACT ANAPHORA and QUOTING function as implicit markers by the dialog participants of the central propositions under discussion. Discourse deictic references such as *That alone* in S1-1 and *That's your evidence?* in S2-2 in Fig. 2 are an indicator of which propositions are under discussion [148, 23, 91, 11, 65], as is verbatim quoting as illustrated in Fig. 3. Previous computational work on abstract anaphora has primarily elucidated the underlying cognitive and linguistic mechanisms in understanding monologic text, [148, 11], and there is limited empirical work on resolving propositional anaphora in dialog [65, 91, 23, 41, 130]. However, constraints on distance for possible antecedents for deictics restrict the possibilities of what can be under discussion, and we plan to build directly on the algorithms described by Byron and by Eckert & Strube and adapt them to our dialogs and their unique properties.

3.3 Identifying Participants' Stances toward Central Propositions

Stance is defined as an overall subjective position held by a person towards an object, idea or position [122]. Much computational work on stance classification in argumentative genres has both a) focused on determining poster stance at a macroscopic level (e.g., death penalty) and b) treated stance as a binary distinction (pro vs. con). In monologic soapboxing, as occurs on debate sites such as forandagainst.com, both of these assumptions are reasonable, as each "debate" is actually set up as a forced choice between two choices and posters are simply asked to argue in favor of their position. Opinion sharing dialog, however, is far more complex than these idealizations. First, because it is dialogic, even ostensibly soapboxing sites like ConvinceMe.net may end up with deep discussion trees where the issue under dispute is relatively microscopic. Fig. 2 provides an example of discussions of the deterrance facet from the debate framed as being about morality shown at the top of Fig. 14. At a practical level, this dialogic setting can make determining macroscopic stance more difficult.

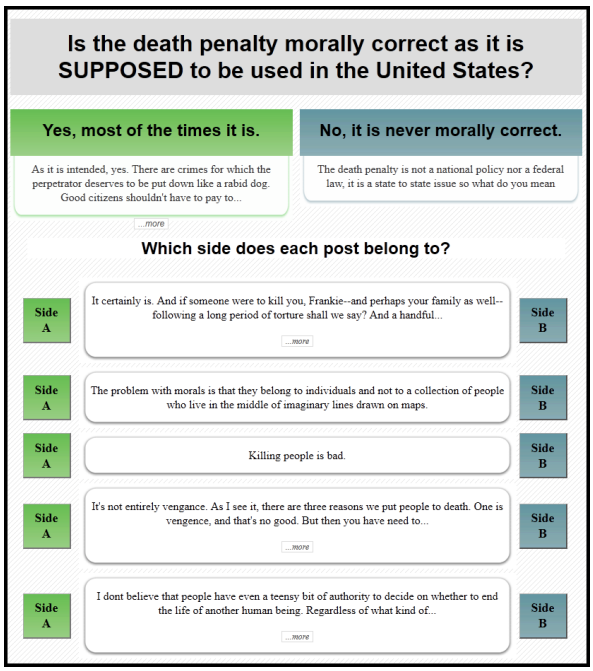

Figure 14: A Mechanical Turk HIT for the Death Penalty Topic framed in terms of the Morality facet.

Fig. 14 shows how we set up the context with the framing post and its sides, but provided no further context for the annotators. Fig. 15 shows the result of Mechanical Turk annotation of ConvinceMe post stance classification *in isolation*; the average human accuracy is 79%, but rebuttals are correctly labelled only 71% of the time, while non-rebuttals are correct 84% of the time. We believe that this shows clearly that richer contextual representations are needed for automatic identification of stance. In addition, research shows that dialogic settings are more encouraging to people who do not self-label into one of two canonical polar categories. Polling data from the Pew Research Center's General Social Survey shows that over the past twenty years between 5-15% of Americans refuse to classify themselves as firmly PRO or CON on the issues of abortion, gay marriage, and gun control [113]. For 200 discussions on `4forums` on 5 socio-political topics (abortion, gay marriage, gun control, evolution, and existence of god), we asked annotators to classify each participant as PRO, CON, or OTHER. Across these topics, 25% of discussants were classified as OTHER (Fleiss's κ = 0.65); annotators divided these participants into those who did not reveal enough information about their stance (10%) and those who were frame

challenging (15%).

A:	Hey I am pro-choice and believe that life begins at conception. However - that doesn't mean that B, C, or D get to make blanket decisions for total strangers without their input and without regard to their circumstances. There are simply better ways to address abortion than stepping on the rights and privacy of pregnant women, strapping down their options, demonizing them for their sexuality and trivializing everything they are facing.
B:	Don't lump me into a group unless you know where I stand. I'm pro-life except in cases of rape, incest or endangerment of the mother's life. I don't like abortions, but sometimes they are justified. There are too many potential parents wanting children they can't have themselves to allow random abortions just because the pregnant female or her family doesn't want the baby....and life does begin at conception.
C:	Like all the universe, biological evolution is either rational, lawful, limited, constrained and directed by universal forces, or it is an illusion. I happen to believe my eyes and ears, and accept its reality. Therefore I cannot accept either the irrational chance-based, "anything-is-possible" fantasy of RMNS darwin, nor the irrational fantasy of an all-powerful, "He can do anything!" magician god....
D:	You'd like to pretend to the world that the debate over darwinian evolution is a case of, "science versus superstition". If that were the truth darwinian evolution would have won in a walk many decades past. But that is not the truth. Not even close. The truth is that RMNS is not science at all, and the only reason anyone promotes it is because it is a basic underpinning for materialist/mechanist atheism. The notion of "spontaneous generation" is its fundamental superstition.
E:	Government should not be involved in marriage in the first place. People should not be required to obtain a license (permission) to get married. I suspect marriage licenses probably came into being as way for government to control who married whom; or, to be more precise: to make sure certain people (of a race, religion, nationality) did not marry certain other people (of a race, religion, nationality).

Figure 16: Middle Positions

Fig. 16 provides examples of middle positions identified. In opinion sharing dialogs, such participants, who agree with some propositions on each side of an issue, prove confounding for others, and can complicate the back-and-forth that is so characteristic of opinionated discourse.

For these two reasons – the presence of centrist positions and the sometimes targeted nature of stance conveyed in a discourse – we believe that opinion sharing dialogs are best characterized in terms of the stances participants bear toward the central propositions of the discussion. Consider the extract in Table 3 from a discussion on abortion, where two participants joust about whether pregnancy is life-threatening and/or physically harmful. Because S1 and S2 discuss the same issues, lexical features alone are not effective predictors of macroscopic stance; indeed, it is only when one identifies the propositions under question and the participants' stances toward them that we can effectively characterize both what is going on in the discourse and the participants' macroscopic stances.

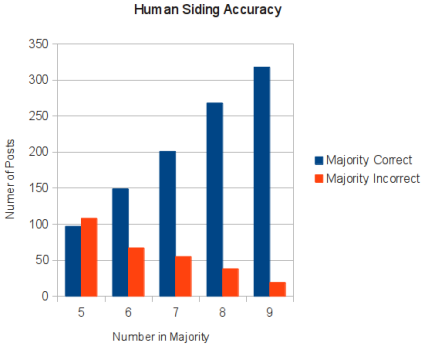

Figure 15: Accuracy of human stance classification for Convinceme. **Accuracy** = % posts where the majority correctly identify the post's side.

S1-1:	Even if you want a baby, it's unbelievably hard to go through and something your body never fully recovers from....You can't force somebody to go through this **life-threatening condition** that is always *physically harmful*.
S2-1:	"**Life threatening condition** that is always *physically harmful*"? What a giant load of steamy BS. Rarely is pregnancy *physically harmful* and even rarer is it **life threatening**.
S1-2:	Pregnancy is ALWAYS **physically harmful**. You try carrying a load of extra weight about and see what that does to your heart.
S2-2:	So, tell me, if it is ALWAYS **life-threatening**, why are there so few **life-threatening conditions** mentioned above?

Table 3: An extract from an abortion debate from ConvinceMe.net, with central propositions in bold.

In determining stances towards central propositions, we will follow two approaches. First, we will build on previous work on macroscopic stance classification in terms of stance towards key terms. In this work on both Congressional debates and on online forums, agreement relations between utterances are identified in two ways: (1) by using the polarity of mentions of other speaker's names at a global level, or the polarities of mentions of second person pronouns [133, 16, 55, 57], or (2) by assuming that all replies are disagreements [87]. Our initial work on automatic identification of stance side has built on this research. We also construct a social network structure as a graph-based representation of the dialog, but in our case it is based on the structure of replies in the dialog (agreement, disagreement). Our pilots show absolute improvements in stance classification accuracy ranging from 0 to 27% depending on topic. In our dialogs the percentage of disagreements varies from about 80 to 90%. An important determiner in whether the graph algorithms increase accuracy for stance classification is the percentage of posts linked into a dialogic structure. Importantly, this approach converges reasonably when partitioning the graph in two, and one key question during the proposed research will be how to extend this to multiple competing perspectives and middle positions.

One direction we have been piloting, which the extract in Fig. 3 is indicative of, is finding syntactic and semantic markers that reliably signal a speaker's (lack) of commitment to a proposition. For example, in Table 3 **S2** clearly marked his position by the context in which the propositional proxies occur, i.e. placing them within questions or conditionals, alongside adverbs such as 'rarely', or within the scope of negation [31]. We characterize these markers as explicit indications of **lack of commitment**. Table 4 provides a list of commitment and non-commitment markers from the theoretical literature [9, 137], along with frequencies in our pilot corpus. We have tested this idea with methods that treat negation and quotation as noncommitment for two topics in the pilot corpus, but which do not explicitly identify the central propositions. This modest alteration produced a 2% improvement overall for classifying at the post level in isolating, a significant result ($p < .05$). We plan to extend this approach to other commitment markers, along with using argument zones to identify propositions central to an issue.

Category	Examples	Commitment class	Convinceme N	4Forums N
conditionals	if ...	non-commitment	8995	152556
questions	...?	non-commitment	11307	89473
challenging	You think, said; I doubt	non-commitment	5021	110245
skeptical adverbs	rarely, maybe, never	non-commitment	6840	116672
assertive	I think, said; you acknowledge, know	commitment	7467	126653
actuality adverbs	actually, amazingly, certainly, unfortunately	commitment	4224	88000

Table 4: Commitment Markers and their frequencies.

Ultimately, one important question of the proposed research is how rich a discourse context is necessary to effectively characterize the stances of participants. We have shown, for example, that classifying the macroscopic stance of a post is significantly improved by including information from its parent, even if that information is simply the parent unigram counts. But we have not systematically investigated how helpful further history would be. Other work that uses some type of contextual information for discourse processing, such as recognizing implicit discourse relations, or recognizing agreement vs. disagreement, also uses rudimentary representations of context or no contextual representation or uses only global information about the speakers [104, 75, 47, 22, 87, 123, 133, 16].

DEBATE TITLE	N	PRO FRAMING POST	CON FRAMING POST
Should the U.S continue death penalty executions?	39	Kill them all !	Let them rot in prison !
Should child molesters face the death penalty	28	Yes, fry the bastards	No, just imprison them.
Death Penalty; justice?	22	They should be put to death. An eye for an eye.	They should have life in prison. Two wrongs don't make a right.
Is the death penalty morally correct as it is SUPPOSED to be used in the United States?	12	Yes, most of the times it is.	No, it is never morally correct.
whether to abolish death penalty	31	abolish death penalty = Life without parole	execution

Figure 17: Debate Framings mapped to the Death Penalty topic in `Convinceme`. **N** = number of dialog turns.

Moreover, although we have manually mapped how our general topics are framed at the start of particular discussions, as in Fig. 17, we have not utilized this information, even though it likely often serves as an excellent indicator of some central propositions. For example, Fig. 14 illustrates how posts may be oriented when the dialog is framed according to the value of *Morality*. Thus, while our initial results show a promising improvement for some representations of context, we intend to focus in the first year on developing richer representations of context, including exploring the use of sequential information and conditional probabilies [47], as well as the use of framing information.