Stance Classification using Dialogic Properties of Persuasion

Marilyn A. Walker, Pranav Anand, Robert Abbott and Ricky Grant
Baskin School of Engineering & Linguistics Department
University of California Santa Cruz
Santa Cruz, Ca. 95064, USA
maw,panand,abbott,rgrant@soe.ucsc.edu

Abstract

Public debate functions as a forum for both expressing and forming opinions, an important aspect of public life. We present results for automatically classifying posts in online debate as to the position, or STANCE that the speaker takes on an issue, such as **Pro** or **Con**. We show that representing the dialogic structure of the debates in terms of agreement relations between speakers, greatly improves performance for stance classification, over models that operate on post content and parent-post context alone.

1 Introduction

Public debate functions as a forum for both *expressing* and *forming* opinions. Three factors affect opinion formation, e.g. the perlocutionary uptake of debate arguments (Cialdini, 2000; Petty and Cacioppo, 1988; Petty et al., 1981). First, there is the ARGUMENT itself, i.e. the propositions discussed along with the logical relations between them. Second is the SOURCE of the argument (Chaiken, 1980), e.g. the speaker's expertise, or agreement relations between speakers. The third factor consists of properties of the AUDIENCE such as prior beliefs, social identity, personality, and cognitive style (Davies, 1998). Perlocutionary uptake in debates primarily occurs in the audience, who may be undecided, while debaters typically express a particular position or STANCE on an issue, e.g. **Pro** or **Con**, as in the online debate dialogues in Figs. 1, 2, and 3.

Previous computational work on debate covers three different debate settings: (1) congressional debates (Thomas et al., 2006; Bansal et al., 2008; Yessenalina et al., 2010; Balahur et al., 2009; Burfoot et al., 2011); (2) company-internal discussion sites (Murakami and Raymond, 2010; Agrawal et al., 2003); and (3) online social and political public forums (Somasundaran and Wiebe, 2009; Somasundaran and Wiebe, 2010; Wang and Rosé, 2010; Biran and Rambow, 2011). Debates in online public forums (e.g. Fig. 1) differ from debates in congress and on company discussion sites in two ways.

Post	Stance	Utterance
P1	PRO	I feel badly for your ignorance because although there maybe a sliver of doubt that mankind may have evolved from previous animals, there is no doubt that the Earth and the cosmos have gone through evolution and are continuing to do so
P2	CON	As long as there are people who doubt evolution, both lay and acedamia, then evolution is in doubt. And please don't feel bad for me. I am perfectly secure in my "ignorance".
P3	PRO	By that measure, as long as organic chemistry, physics and gravity are in doubt by both lay and acedamia, then organic chemistry, physics and gravity are in doubt. Gravity is a theory. Why aren't you giving it the same treatment you do to evolution? Or is it because you are ignorant? Angelic Falling anyone?
P4	CON	I'm obviously ignorant. Look how many times i've been given the title. "Gravity is a theory. Why aren't you giving it the same treatment you do to evolution?" Because it doesn't carry the same weight. ;P

Figure 1: All posts linked via rebuttal links. The topic was "Evolution", with sides "Yes, I Believe" vs. "No, I Dont Believe".

First, the language is different. Online debaters are highly involved, often using emotional and colorful language to make their points. These debates are also personal, giving a strong sense of the indi-

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 592–596,
Montréal, Canada, June 3-8, 2012.

vidual making the argument, and whether s/he favors emotive or factual modes of expression, e.g. *Let me answer.... NO!* (P2 in Fig. 3). Other common features are sarcasm, e.g. *I'm obviously ignorant. Look how many times i've been given the title* (P4 in Fig. 1), questioning another's evidence or assumptions: *Yes there is always room for human error, but is one accident that hasn't happened yet enough cause to get rid of a capital punishment?* (P2 in Fig. 3), and insults: *Or is it because you are ignorant?* (P3 in Fig. 1). These properties may function to engage the audience and persuade them to form a particular opinion, but they make computational analysis of such debates challenging, with the best performance to date averaging 64% over several topics (Somasundaran and Wiebe, 2010).

Post	Stance	Utterance
P1	Superman	Batman is no match for superman. Not only does he have SUPERnatural powers as opposed to batman's wit and gadgetry, but his powers have increased in number over the years. For example, when Superman's prowess was first documented in the comics he did not have x-ray vision. It wasn't until his story was told on radio that he could see through stuff. So no matter what new weapon batman could obtain, Superman would add another SUPERnatural weapon to foil the Caped crusader.
P2	Batman	Superman GAVE Batman a krytonite ring so that Batman could take him down should he need to. Superman did this because he knows Batman is the only guy that could do it.
P3	Superman	But, not being privy to private conversations with S-man, you wouldn't know that, being the humble chap that he is, S-man allowed batman the victory because he likes the bat and wanted him to mantain some credibility. Honest.
P4	Batman	Hmmm, this is confusing. Since we all know that Supes doesn't lie and yet at the time of him being beaten by Batman he was under the control of Poison Ivy and therefore could NOT have LET Batman win on purpose. I have to say that I am beginning to doubt you really are friends with Supes at all.

Figure 2: All posts linked via rebuttal links. The topic was "Superman vs. Batman"

Second, the affordances of different online debate sites provide differential support for dialogic relations between forum participants. For example, the research of Somasundaran and Wiebe (2010), does not explicitly model dialogue or author relations. However debates in our corpus vary greatly by topic on two dialogic factors: (1) the percent of posts that are rebuttals to prior posts, and (2) the number of posts per author. The first 5 columns of Table 2 shows the variation in these dimensions by topic.

Post	Stance	Utterance
P1	CON	69 people have been released from death row since 1973 these people could have been killed if there cases and evidence did not come up rong also these people can have lost 20 years or more to a false coviction. it is only a matter of time till some one is killed yes u could say there doing a good job now but it has been shown so many times with humans that they will make the human error and cost an innocent person there life.
P2	PRO	Yes there is always room for human error, but is one accident that hasn't happened yet enough cause to get rid of a capital punishment? Let me answer...NO! If you ban the death penalty crime will skyrocket. It is an effective deterannce for crime. The states that have strict death penalty laws have less crime than states that don't (Texas vs. Michigan) Texas's crime rate is lower than Michigan and Texas has a higher population!!!!

Figure 3: Posts linked via rebuttal links. The topic was "Capital Punishment", and the argument was framed as "Yes we should keep it" vs. "No we should not".

In this paper we show that information about dialogic relations between authors (SOURCE factors) improves performance for STANCE classification, when compared to models that only have access to properties of the ARGUMENT. We model SOURCE relations with a graph, and add this information to classifiers operating on the text of a post. Sec. 2 describes the corpus and our approach. Our corpus is publicly available, see (Walker et al., 2012). We show in Sec. 3 that modeling source properties improves performance when the debates are highly dialogic. We leave a more detailed comparison to previous work to Sec. 3 so that we can contrast previous work with our approach.

2 Experimental Method and Approach

Our corpus consists of two-sided debates from Convinceme.net for 14 topics that range from playful debates such as Superman vs. Batman (Fig. 2 to more heated political topics such as the Death Penalty (Fig. 3. In total the corpus consists of 2902 two-sided debates (36,307 posts), totaling 3,080,874 words; the topic labelled debates which we use in our experiments contain 575,818 words. On Convinceme, a person starts a debate by posting a topic or a question and providing sides such as *for* vs. *against*. Debate participants can then post arguments for one side or the other, essentially self-

labelling their post for stance. Convinceme provides three possible sources of dialogic structure, SIDE, REBUTTAL LINKS and TEMPORAL CONTEXT. Timestamps for posts are only available by day and there are no agreement links. Here, we use the self-labelled SIDE as the stance to be predicted.

Set/Factor	Description
Basic	Number of Characters in post, Average Word Length, Unigrams, Bigrams
Sentiment	LIWC counts and frequencies, Opinion Dependencies, LIWC Dependencies, negation
Argument	Cue Words, Repeated Punctuation, Context, POS-Generalized Dependencies, Quotes

Table 1: Feature Sets

We construct features from the posts, along with a representation of the parent post as context, and use those features in several base classifiers. As shown in Table 1, we distinguish between basic features, such as length of the post and the words and bigrams in the post, and features capturing **sentiment** and subjectivity, including using the LIWC tool for emotion labelling (Pennebaker et al., 2001) and deriving generalized dependency features using LIWC categories, as well as some limited aspects of the **argument** structure, such as cue words signalling rhetorical relations between posts, POS generalized dependencies, and a representation of the parent post (context). Only rebuttal posts have a parent post, and thus values for the context features.

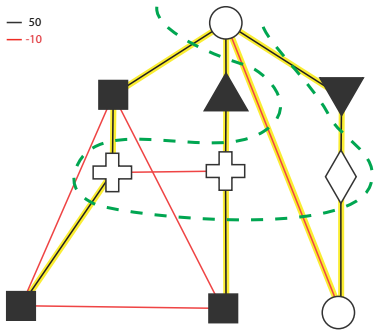

Figure 4: Sample maxcut to ConvinceMe siding. Symbols (circle, cross, square, triangles) indicate authors and fill colors (white,black) indicate true side. Rebuttal links are marked by black edges, same-author links by red; weights are 50 and -10, respectively. Edges in the maxcut are highlighted in yellow, and the nodes in each cut set are bounded by the green dotted line.

We then construct a graph (V,E) representing the dialogue structure, using the rebuttal links and author identifiers from the forums site. Each node V of the graph is a post, and edges E indicate dialogic relations of agreement and disagreement between posts. We assume only that authors always agree with themselves, and that rebuttal links indicate disagreement. Agreement links based on the inference that if A, B disagree with C they agree with each other were not added to the graph.

Maxcut attempts to partition a graph into two sides. Fig. 4 illustrates a sample result of applying MaxCut. Edges connecting the partitions are said to be cut, while those within partitions are not. The goal is to maximize the sum of cut edge weights. By making edge weights high we reward the algorithm for cutting the edge, by making edge weights negative we penalize the algorithm for cutting the edge. Rebuttal links were assigned a weight +100/(number of rebuttals). Same author links were assigned a weight -60/(number of posts by author). If author A rebutted author B at some point, then a weight of 50 was assigned to all edges connecting posts by author A and posts by author B. If author B rebutted author A as well, that 50 was increased to 100. We applied the MaxCut partitioning algorithm to this graph, and then we orient each of the components automatically using a traditional supervised classifier. We consider each component separately where components are defined using the original (pre-MaxCut) graph. For each pair of partition side $p \in \{P_0, P_1\}$ and classifier label $l \in \{L_0, L_1\}$, we compute a score $S_{p,l}$ by summing the margins of all nodes assigned to that partition and label. We then compute and compare the score differences for each partition. $D_p = S_{p,L_1} - S_{p,L_0}$ If $D_{P_0} < D_{P_1}$, then nodes in partition P_0 should be assigned label L_0 and nodes in P_1 should be assigned label L_1. Likewise, if $D_{P_0} > D_{P_1}$, then nodes in partition P_0 should be assigned label L_1 and nodes in P_1 should be assigned label L_0. If $D_{P_0} = D_{P_1}$, then we orient the component with a coin flip.

3 Results and Discussion

Table 2 summarizes our results for the base classifier (JRIP) compared to using MaxCut over the social network defined by author and rebuttal links. We report results for experiments using all the fea-

	Topic Characteristics					MaxCut Algorithm				JRIP Algorithm			
Topic	Posts	Rebs	P/A	A> 1p	MLE	Acc	F1	P	R	Acc	F1	P	R
Abortion	607	64%	2.73	42%	53%	82%	0.82	0.78	0.88	55%	0.55	0.52	0.59
Cats v. Dogs	162	40%	1.60	24%	53%	80%	0.78	0.80	0.76	61%	0.55	0.59	0.51
Climate Change	207	65%	2.92	41%	50%	64%	0.66	0.63	0.69	61%	0.62	0.60	0.63
Comm. v. Capitalism	214	62%	2.97	46%	55%	70%	0.67	0.66	0.68	53%	0.49	0.48	0.49
Death Penalty	331	60%	2.40	45%	56%	35%	0.31	0.29	0.34	55%	0.46	0.48	0.44
Evolution	818	66%	3.74	53%	58%	82%	0.78	0.78	0.79	56%	0.49	0.48	0.50
Existence Of God	852	76%	4.16	51%	56%	75%	0.73	0.70	0.76	52%	0.49	0.47	0.51
Firefox v. IE	233	38%	1.27	15%	79%	76%	0.47	0.44	0.49	72%	0.33	0.34	0.33
Gay Marriage	560	56%	2.01	28%	65%	84%	0.77	0.74	0.81	60%	0.43	0.43	0.44
Gun Control	135	59%	2.08	45%	63%	37%	0.24	0.21	0.27	53%	0.24	0.30	0.20
Healthcare	112	79%	3.11	53%	55%	73%	0.71	0.69	0.72	60%	0.49	0.56	0.44
Immigration	78	58%	1.95	33%	54%	33%	0.21	0.23	0.19	53%	0.39	0.48	0.33
Iphone v. Blackberry	25	44%	1.14	14%	67%	88%	0.80	0.86	0.75	71%	0.46	0.60	0.38
Israel v. Palestine	64	33%	3.37	53%	58%	85%	0.82	0.79	0.85	49%	0.48	0.42	0.56
Mac v. PC	126	37%	1.85	24%	52%	19%	0.18	0.17	0.18	46%	0.46	0.45	0.48
Marijuana legalization	229	45%	1.52	25%	71%	73%	0.56	0.52	0.60	63%	0.34	0.35	0.34
Star Wars vs. LOTR	102	44%	1.38	26%	53%	63%	0.62	0.60	0.65	63%	0.62	0.60	0.65
Superman v. Batman	146	30%	1.39	20%	54%	50%	0.40	0.44	0.37	56%	0.47	0.52	0.43

Table 2: Results. **KEY**: Number of posts on the topic (**Posts**). Percent of Posts linked by Rebuttal links (**Rebs**). Posts per author (**P/A**). Authors with more than one post (**A > 1P**). Majority Class Baseline (**MLE**).

tures with χ^2 feature selection; we use JRIP as the base classifier because margins are used by the automatic MaxCut graph orientation algorithm. Experiments with different learners (NB, SVM) did not yield significant differences from JRIP. The results show that, in general, representing dialogic information in terms of a network of relations between posts yields very large improvements. In the few topics where performance is worse (Death Penalty, Gun Control, Mac vs. PC, Superman vs. Batman), the MaxCut graph gets oriented to the stance sides the wrong way, so that the cut actually groups the posts correctly into sides, but then assigns them to the wrong side. For Maxcut, as expected, there are significant correlations between the % of Rebuttals in a debate and Precision (R = .16) and Recall (R= .22), as well as between Posts/Author and Precision (R = .25) and Recall (R = .43). This clearly indicates that the degree of dialogic behavior (the graph topology) has a strong influence on results per topic. These results would be even stronger if all MaxCut graphs were oriented correctly.

(Somasundaran and Wiebe, 2010) present an unsupervised approach using ICA to stance classification, showing that identifying argumentation structure improves performance, with a best performance averaging 64% accuracy over all topics, but as high as 70% for some topics. Other research classifies the speaker's side in a corpus of congressional floor debates (Thomas et al., 2006; Bansal et al., 2008; Balahur et al., 2009; Burfoot et al., 2011). Thomas et al (2006) achieved accuracies of 71.3% by using speaker agreement information in the graph-based MinCut/Maxflow algorithm, as compared to accuracies around 70% via an an SVM classifier operating on content alone. The best performance to date on this corpus achieves accuracies around 82% for different graph-based approaches as compared to 76% accuracy for content only classification (Burfoot et al., 2011). Other work applies MaxCut to the reply structure of company discussion forums, showing that rules for identifying agreement (Murakami and Raymond, 2010), defined on the textual content of the post yield performance improvements over using reply structures alone (Malouf and Mullen, 2008; Agrawal et al., 2003)

Our results are not strictly comparable since we use a different corpus with different properties, but to our knowledge this is the first application of MaxCut to stance classification that shows large performance improvements from modeling dialogic relations. In future work, we plan to explore whether deeper linguistic features can yield large improvements in both the base classifier and in MaxCut results, and to explore better ways of automatically orienting the MaxCut graph to stance side. We also hope to develop much better context features and to make even more use of dialogue structure.

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.

A. Balahur, Z. Kozareva, and A. Montoyo. 2009. Determining the polarity and source of opinions expressed in political debates. *Computational Linguistics and Intelligent Text Processing*, pages 468–480.

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.

O. Biran and O. Rambow. 2011. Identifying justifications in written dialogs. In *2011 Fifth IEEE International Conference on Semantic Computing (ICSC)*, pages 162–168. IEEE.

C. Burfoot, S. Bird, and T. Baldwin. 2011. Collective classification of congressional floor-debate transcripts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1506–1515. Association for Computational Linguistics.

S. Chaiken. 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology*, 39(5):752.

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

M.F. Davies. 1998. Dogmatism and belief formation: Output interference in the processing of supporting and contradictory cognitions. *Journal of personality and social psychology*, 75(2):456.

R. Malouf and T. Mullen. 2008. Taking sides: User classification for informal online political discourse. *Internet Research*, 18(2):177–190.

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.

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

Richard E. Petty and John T. Cacioppo. 1988. The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology*, 46(1):69–81.

R.E. Petty, J.T. Cacioppo, and R. Goldman. 1981. Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, 41(5):847.

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.

M. Walker, P. Anand, J. Fox Tree, R. Abbott, and J. King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*.

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.

A. Yessenalina, Y. Yue, and C. Cardie. 2010. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1046–1056. Association for Computational Linguistics.