Proceedings of the 51st Annual Meeting of the Association for
Computational Linguistics (ACL), Short Papers, Sofia, Bulgaria, August 2013, pp. 816--821.

Extra-Linguistic Constraints on Stance Recognition in Ideological Debates

Kazi Saidul Hasan and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{saidul,vince}@hlt.utdallas.edu

Abstract

Determining the stance expressed by an author from a post written for a two-sided debate in an online debate forum is a relatively new problem. We seek to improve Anand et al.'s (2011) approach to debate stance classification by modeling two types of soft extra-linguistic constraints on the stance labels of debate posts, user-interaction constraints and ideology constraints. Experimental results on four datasets demonstrate the effectiveness of these inter-post constraints in improving debate stance classification.

1 Introduction

While a lot of work on document-level opinion mining has involved determining the polarity expressed in a customer review (e.g., whether a review is "thumbs up" or "thumbs down") (see Pang and Lee (2008) and Liu (2012) for an overview of the field), researchers have begun exploring new opinion mining tasks in recent years. One such task is *debate stance classification*: given a post written for a *two-sided* topic discussed in an online debate forum (e.g., *"Should abortion be banned?"*), determine which of the two sides (i.e., *for* and *against*) its author is taking.

Debate stance classification is potentially more interesting and challenging than polarity classification for at least two reasons. First, while in polarity classification sentiment-bearing words and phrases have proven to be useful (e.g., "excellent" correlates strongly with the positive polarity), in debate stance classification it is not uncommon to find debate posts where stances are not expressed in terms of sentiment words, as exemplified in Figure 1, where the author is *for* abortion.

> The fetus is simply a part of the mother's body and she can have an abortion because it is her human rights. Also I take this view because every woman can face with situation when two lives are at stake and the moral obligation is to save the one closest at hand — namely, that of the mother, whose life is always more immediate than that of the unborn child within her body. Permission for an abortion could then be based on psychiatric considerations such as prepartum depression, especially if there is responsible psychiatric opinion that a continued pregnancy raises the strong probability of suicide in a clinically depressed patient.

Figure 1: A sample post on abortion.

Second, while customer reviews are typically written independently of other reviews in an online forum, the same is not true for debate posts. In a debate forum, debate posts form *threads*, where later posts often support or oppose the viewpoints raised in earlier posts in the same thread.

Previous approaches to debate stance classification have focused on three debate settings, namely congressional floor debates (Thomas et al., 2006; Bansal et al., 2008; Balahur et al., 2009; Yessenalina et al., 2010; Burfoot et al., 2011), company-internal discussions (Murakami and Raymond, 2010), and online social, political, and ideological debates in public forums (Agrawal et al., 2003; Somasundaran and Wiebe, 2010; Wang and Rosé, 2010; Biran and Rambow, 2011; Hasan and Ng, 2012). As Walker et al. (2012) point out, debates in public forums differ from congressional debates and company-internal discussions in terms of language use. Specifically, online debaters use colorful and emotional language to express their points, which may involve sarcasm, insults, and questioning another debater's assumptions and evidence. These properties can potentially make stance classification of online debates more challenging than that of the other two types of debates.

Our goal in this paper is to improve the state-of-the-art supervised learning approach to debate stance classification of online debates proposed by Anand et al. (2011), focusing in particular on *ideological debates*. Specifically, we hypothesize that there are two types of soft extra-linguistic constraints on the stance labels of debate posts that,

Domain	Number of posts	"for" posts (%)	% of posts in a thread	Average thread length
ABO	1741	54.9	75.1	4.1
GAY	1376	63.4	74.5	4.0
OBA	985	53.9	57.1	2.6
MAR	626	69.5	58.0	2.5

Table 1: Statistics of the four datasets.

Feature type	Features
Basic	Unigrams, bigrams, syntactic and POS-generalized dependencies
Sentiment	LIWC counts, opinion dependencies
Argument	Cue words, repeated punctuation, context

Table 2: Anand et al.'s features.

if explicitly modeled, could improve a learning-based stance classification system. We refer to these two types of inter-post constraints as *user-interaction constraints* and *ideology constraints*. We show how they can be learned from stance-annotated debate posts in Sections 4.1 and 4.2, respectively.

2 Datasets

For our experiments, we collect debate posts from four popular *domains*, Abortion (ABO), Gay Rights (GAY), Obama (OBA), and Marijuana (MAR), from an online debate forum[1]. All debates are two-sided, so each post receives one of two *domain labels*, *for* or *against*, depending on whether the author of the post *supports* or *opposes* abortion, gay rights, Obama, or the legalization of marijuana.

We construct one dataset for each domain (see Table 1 for statistics). The fourth column of the table shows the percentage of posts in each domain that appear in a *thread*. More precisely, a *thread* is a tree with one or more nodes such that (1) each node corresponds to a debate post, and (2) a post y_i is the parent of another post y_j if y_j is a reply to y_i. Given a thread, we can generate *post sequences*, each of which is a path from the root of the thread to one of its leaves.

3 Baseline Systems

We employ as baselines two stance classification systems, Anand et al.'s (2011) approach and an enhanced version of it, as described below.

Our first baseline, Anand et al.'s approach is a supervised method that trains a stance classifier for determining whether the stance expressed in a debate post is *for* or *against* the topic. Hence, we create one training instance from each post in the training set, using the stance it expresses as its class label. Following Anand et al., we represent a training instance using three types of lexico-syntactic features, which are briefly summarized in Table 2. In our implementation, we train the stance classifier using SVM^{light} (Joachims, 1999). After training, we can apply the classifier to classify the test instances, which are generated in the same way as the training instances.

Related work on stance classification of *congressional debates* has found that enforcing *author constraints* (ACs) can improve classification performance (e.g., Thomas et al. (2006), Bansal et al. (2008), Burfoot et al. (2011), Lu et al. (2012), Walker et al. (2012)). ACs are a type of inter-post constraints that specify that two posts written by the same author for the same debate domain should have the same stance. We hypothesize that ACs could similarly be used to improve stance classification of ideological debates, and therefore propose a second baseline where we enhance the first baseline with ACs. Enforcing ACs is simple. We first use the learned stance classifier to classify the test posts as in the first baseline, and then *post-process* the labels of the test posts. Specifically, we sum up the confidence values[2] assigned to the set of test posts written by the same author for the same debate domain. If the sum is positive, then we label *all* the posts in this set as *for*; otherwise we label them as *against*.

4 Extra-Linguistic Constraints

In this section, we introduce two types of inter-post constraints on debate stance classification.

4.1 User-Interaction Constraints

We call the first type of constraints *user-interaction constraints* (UCs). UCs are motivated by the observation that the stance labels of the posts in a post sequence are not independent of each other. Consider the post sequence in Figure 2, where each post is a response to the preceding post. It shows an opening anti-abortion post (P1), followed by a pro-abortion comment (P2), which is in turn followed by another anti-abortion view (P3). While this sequence contains alternating posts from opposing stances, in general there is no hard constraint on the stance of a post given

[1] `http://www.createdebate.com/`

[2] We use as the confidence value the signed distance of the associated test point from the SVM hyperplane.

[P1: Anti-abortion] There are thousands of people who want to take these children because they cannot have their own. If you do not want a child, have it and put it up for adoption. At least you will be preserving a human life rather than killing one.

> **[P2: Pro-abortion]** I agree that if people don't want their babies, they should have the choice of putting it up for adoption. But it should not be made compulsory, which is essentially what happens if you ban abortion.
>
> > **[P3: Anti-abortion]** Why should it not be made compulsory? Those children have as much right to live as you and I. Besides, no one loses with adoption, so why wouldn't you utilize it?

Figure 2: A sample post sequence. P2 and P3 are replies to P1 and P2, respectively.

the preceding sequence of posts. Nevertheless, we found that in our training data, a *for* (*against*) post is followed by a *against* (*for*) post 80% of the time.

UCs aim to model the regularities in how users interact with each other in a post sequence as soft constraints. These kinds of soft constraints can be naturally encoded as *factors* over adjacent posts in a post sequence (see Kschischang et al. (2001)), which can in turn be learned by recasting stance classification as a *sequence labeling* task. In our experiments, we seek to derive the best sequence of stance labels for each post sequence of length $\geq$ 1 using a Conditional Random Field (CRF) (Lafferty et al., 2001).

We train the CRF model using the CRF implementation in Mallet (McCallum, 2002). Each training sequence corresponds to a post sequence. Each post in a sequence is represented using the same set of features as in the baselines.

After training, the resulting CRF model can be used to assign a stance sequence to each test post sequence. There is a caveat, however. Since a given test post may appear in more than one sequence, different occurrences of it may be assigned different stance labels by the CRF. To determine the final stance label for the post, we average the probabilities assigned to the *for* stance over all its occurrences; if the average is $\geq$ 0.5, then its final label is *for*; otherwise, its label is *against*.

4.2 Ideology Constraints

Next, we introduce our second type of inter-post constraints, *ideology constraints* (ICs). ICs are *cross-domain*, *author-based* constraints: they are only applicable to debate posts written by the same author in different domains. ICs model the fact that for some authors, their stances on various issues are determined in part by their ideological values, and in particular, their stances on different issues may be correlated. For example, someone who opposes abortion is likely to be a conservative and has a good chance of opposing gay rights. ICs aim to capture this kind of inter-domain correlation of stances. Below we describe how we implement ICs and show how they can be integrated with ACs.

4.2.1 Implementing Ideology Constraints

We first compute a set of conditional probabilities, P(stance(d_q)=s_d|stance(d_p)=s_c), where (1) d_p, d_q $\in$ *Domains* (i.e., the set of four domains), (2) s_c, s_d $\in$ {*for*, *against*}, and (3) $d_p \neq d_q$. To compute P(stance(d_q)=s_d|stance(d_p)=s_c), we (1) determine for each author a in the training set and each domain d_p the stance of a in d_p (denoted by author-stance(d_p,a)), where author-stance(d_p,a) is computed as the majority stance labels associated with the debate posts in the training set that a wrote for d_p; and (2) compute P(stance(d_q)=s_d|stance(d_p)=s_c) as the ratio of $\sum_{a\in A}$ Count(author-stance(d_p,a)=s_c, author-stance(d_q,a)=s_d) to $\sum_{a\in A}$ Count(author-stance(d_p,a)=s_c), where A is the set of authors in the training set who posted in both d_p and d_q. It should be fairly easy to see that these conditional probabilities measure the degree of correlation between the stances in different domains.

4.2.2 Inference Using ILP

Recall that in our second baseline, we employ ACs to postprocess the output of the stance classifier simply by summing up the confidence values assigned to the posts written by the same author for the same debate domain. However, since we now want to enforce two types of inter-post constraints (namely, ACs and ICs), we will have to employ a more sophisticated inference mechanism. Previous work has focused on employing graph minimum cut (MinCut) as the inference algorithm. However, since MinCut suffers from the weakness of not being able to enforce negative constraints (i.e., two posts cannot receive the same label) (Bansal et al., 2008), we propose to use integer linear programming (ILP) as the underlying inference mechanism. Below we show how to implement ACs and ICs within the ILP framework.

Owing to space limitations, we refer the reader to Roth and Yih (2004) for details of the ILP framework. Briefly, ILP seeks to optimize an objective function subject to a set of linear constraints. Below we focus on describing the ILP program and how the ACs and ICs can be encoded.

Let $Y = y_1, \ldots, y_n$ be the set of debate posts. For each y_i, we create one (binary-valued) indicator variable x_i, which will be used in the ILP program. Let $p_i = P(\textit{for}|y_i)$ be the "benefit" of setting x_i to 1, where $P(\textit{for}|y_i)$ is provided by the CRF. Consequently, after optimization, y_i's stance is *for* if its x_i is set to 1. We optimize the following objective function:

$$\max \sum_i p_i x_i + (1 - p_i)(1 - x_i)$$

subject to a set of *linear* constraints, which encode the ACs and the ICs, as described below.

Implementing author constraints. If y_i and y_j are composed by the same author, we ensure that x_i and x_j will be assigned the same value by employing the linear constraint $|x_i - x_j| = 0$.

Implementing ideology constraints. For convenience, below we use the notation introduced in Section 4.2.1, and assume that y_i and y_j are two arbitrary posts written by the same author in domains d_p and d_q, respectively.

Case 1: If P(stance(d_q)=*for*|stance(d_p)=*for*) $\geq t$, we want to ensure that x_i=1 $\implies$ x_j=1.[3] This can be achieved using the constraint $(1-x_j) \leq (1-x_i)$.

Case 2: If P(stance(d_q)=*against*|stance(d_p)=*against*) $\geq t$, we want to ensure that x_i=0 $\implies$ x_j=0. This can be achieved using the constraint $x_j \leq x_i$.

Case 3: If P(stance(d_q)=*against*|stance(d_p)=*for*) $\geq t$, we want to ensure that x_i=1 $\implies$ x_j=0. This can be achieved using the constraint $x_j \leq (1-x_i)$.

Case 4: If P(stance(d_q)=*for*|stance(d_p)=*against*) $\geq t$, we want to ensure that x_i=0 $\implies$ x_j=1. This can be achieved using the constraint $(1-x_j) \leq x_i$.

Two points deserve mention. First, cases 3 and 4 correspond to negative constraints, and unlike in MinCut, they can be implemented easily in ILP. Second, if ICs are used, one ILP program will be created to perform inference over the debate posts in all four domains.

5 Evaluation

5.1 Experimental Setup

Results are expressed in terms of *accuracy* obtained via 5-fold cross validation, where accuracy is the percentage of test instances correctly classified. Since all experiments require the use of development data for parameter tuning, we use three folds for model training, one fold for development, and one fold for testing in each fold experiment.

[3]Intuitively, if this condition is satisfied, it means that there is sufficient evidence that the two nodes from different domains should have the same stance, and so we convert the soft ICs into (hard) linear constraints in ILP. Note that t is a threshold to be tuned using development data.

System	ABO	GAY	OBA	MAR
Anand	61.4	62.6	58.1	66.9
Anand+AC	72.0	64.9	62.7	67.8
Anand+AC+UC	73.7	69.9	64.1	**75.4**
Anand+AC+UC+IC	**74.9**	**70.9**	**72.7**	**75.4**

Table 3: 5-fold cross-validation accuracies.

5.2 Results

Results are shown in Table 3. Row 1 shows the results of the Anand et al. (2011) baseline (see Section 3) on the four datasets, obtained by training a SVM stance classifier using the SVM^{light} software.[4] Row 2 shows the results of the second baseline, Anand et al.'s system enhanced with ACs. As we can see, incorporating ACs into Anand et al.'s system improves its performance significantly on all datasets and yields a system that achieves an average improvement of 4.6 accuracy points.[5]

Next, we incorporate our first type of constraints, UCs, into the better of the two baselines (i.e., the second baseline). Results of applying the CRF for modeling UCs to the test posts and postprocessing them using the ACs are shown in row 3 of Table 3. As we can see, incorporating UCs into the second baseline significantly improves its performance and yields a system that achieves an average improvement of 3.93 accuracy points.

Finally, we incorporate our second type of constraints, ICs, effectively performing inference over the CRF output using ILP with ACs and ICs as the inter-post constraints. Results of this experiment are shown in row 4 of Table 3. As we can see, incorporating the ICs significantly improves the performance of the system on all but MAR and yields a system that achieves an average improvement of 2.7 accuracy points.

Overall, our inter-post constraints yield a stance classification system that significantly outperforms the better baseline on all four datasets, with an average improvement of 6.63 accuracy points.

[4]For all SVM experiments, the regularization parameter C is tuned using development data, but the remaining learning parameters are set to their default values.

[5]All significance tests are paired t-tests, with $p < 0.05$.

5.3 Discussion

Next, we make some observations on the results of applying ICs to our datasets.

First, ICs do not improve the MAR dataset. An examination of the domains reveals the reason. We find three pairs of ICs involving the other three domains — ABO, GAY, and OBA — in our training data. More specifically, the stances of the posts written by an author for these three domains are all positively co-related. In other words, if an author supports abortion, it is likely that she supports both gay rights and Obama as well. On the other hand, we find no co-relation between MAR and the remaining domains. This means that no ICs can be established between the posts in MAR and those in the remaining domains.

Second, the improvement resulting from the application of ICs is much larger on the OBA dataset than on ABO and GAY. The reason can be attributed to the fact that ICs exist more frequently between OBA and ABO and between OBA and GAY than between ABO and GAY. Specifically, ICs are seen in all five folds of the data in the first two pairs of domains, whereas they are seen in only two folds in the last pair of domains.

6 Related Work

Previous work has investigated the use of extra-linguistic constraints to improve stance classification. Introduced by Thomas et al. (2006), ACs are arguably the most commonly used extra-linguistic constraints. Since then, they have been employed and extended in different ways (see, for example, Bansal et al. (2008), Burfoot et al. (2011), Lu et al. (2012), and Walker et al. (2012)).

ICs are different from ACs in at least two respects. First, ICs are softer than ACs, so accurate modeling of ICs has to be based on stance-annotated data. Although we employ ICs as hard constraints (owing in part to our use of the ILP framework), they can be used directly as soft constraints in other frameworks, such as MinCut. Second, ICs are inter-domain constraints, whereas ACs are intra-domain constraints. To our knowledge, this is the first time inter-domain constraints are employed for stance classification.

There has been work related to the modeling of user interaction in a post sequence. Recall that between two adjacent posts in a post sequence that have opposing stances, there exists a *rebuttal* link. Walker et al. (2012) employ manually identified rebuttal links as hard inter-post constraints during inference. However, since automatic discovery of rebuttal links is a non-trivial problem, employing gold rebuttal links substantially simplifies the stance classification task. Lu et al. (2012), on the other hand, predict whether a link is of type *agreement* or *disagreement* using a bootstrapped classifier. Anand et al. (2011) do not predict links. Instead, hypothesizing that the content of the preceding post in a post sequence would be useful for predicting the stance of the current post, they employ features computed based on the preceding post when training a stance classifier. Hence, unlike us, they classify each post independently of the others, whereas we classify the posts in a sequence in dependent relation to each other.

The ILP framework has been applied to perform joint inference for a variety of stance prediction tasks. Lu et al. (2012) address the task of discovering *opposing opinion networks*, where the goal is to partition the authors in a debate (e.g., gay rights) based on whether they support or oppose the given issue. To this end, they employ ILP to coordinate different sources of information. In our previous work on debate stance classification (Hasan and Ng, 2012), we employ ILP to coordinate the output of *two* classifiers: a *post-stance* classifier, which determines the stance of a debate post written for a domain (e.g., gay rights); and a *topic-stance* classifier, which determines the author's stance on each *topic* mentioned in her post (e.g., gay marriage, gay adoption). In this work, on the other hand, we train only one classifier, but use ILP to coordinate two types of constraints, ACs and ICs.

7 Conclusions

We examined the under-studied task of stance classification of ideological debates. Employing our two types of extra-linguistic constraints yields a system that outperforms an improved version of Anand et al.'s approach by 2.9–10 accuracy points. While the effectiveness of ideology constraints depends to some extent on the "relatedness" of the underlying ideological domains, we believe that the gains they offer will increase with the number of authors posting in different domains and the number of related domains.[6]

[6]Only a small fraction of the authors posted in multiple domains in our datasets: 12% and 5% of them posted in two and three domains, respectively.

References

Rakesh Agrawal, Sridhar Rajagopalan, Ramakrishnan Srikant, and Yirong Xu. 2003. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 529–535.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011)*, pages 1–9.

Alexandra Balahur, Zornitsa Kozareva, and Andrés Montoyo. 2009. Determining the polarity and source of opinions expressed in political debates. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '09, pages 468–480.

Mohit Bansal, Claire Cardie, and Lillian Lee. 2008. The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. In *Proceedings of the 22nd International Conference on Computational Linguistics: Companion volume: Posters*, pages 15–18.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, ICSC '11, pages 162–168.

Clinton Burfoot, Steven Bird, and Timothy 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*, pages 1506–1515.

Kazi Saidul Hasan and Vincent Ng. 2012. Predicting stance in ideological debate with rich linguistic knowledge. In *Proceedings of the 24th International Conference on Computational Linguistics: Posters*, pages 451–460.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*, pages 44–56. MIT Press.

Frank Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. 2001. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47:498–519.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Yue Lu, Hongning Wang, ChengXiang Zhai, and Dan Roth. 2012. Unsupervised discovery of opposing opinion networks from forum discussions. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1642–1646.

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

Akiko Murakami and Rudy 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.

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

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning*, pages 1–8.

Swapna Somasundaran and Janyce 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.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335.

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596.

Yi-Chia Wang and Carolyn 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.

Ainur Yessenalina, Yisong Yue, and Claire 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.