Conversational Flow in Oxford-style Debates

Justine Zhang,[1] Ravi Kumar,[2] Sujith Ravi,[2] Cristian Danescu-Niculescu-Mizil[1]
[1]Cornell University, [2]Google
jz727@cornell.edu, ravi.k53@gmail.com,
ravi.sujith@gmail.com, cristian@cs.cornell.edu

Abstract

Public debates are a common platform for presenting and juxtaposing diverging views on important issues. In this work we propose a methodology for tracking how ideas flow between participants throughout a debate. We use this approach in a case study of Oxford-style debates—a competitive format where the winner is determined by audience votes—and show how the outcome of a debate depends on aspects of conversational flow. In particular, we find that winners tend to make better use of a debate's interactive component than losers, by actively pursuing their opponents' points rather than promoting their own ideas over the course of the conversation.

1 Introduction

Public debates are a common platform for presenting and juxtaposing diverging viewpoints. As opposed to monologues where speakers are limited to expressing their own beliefs, debates allow for participants to interactively attack their opponents' points while defending their own. The resulting flow of ideas is a key feature of this conversation genre.

In this work we introduce a computational framework for characterizing debates in terms of conversational flow. This framework captures two main debating strategies—promoting one's own points and attacking the opponents' points—and tracks their relative usage throughout the debate. By applying this methodology to a setting where debate winners are known, we show that conversational flow patterns are predictive of which debater is more likely to persuade an audience.

Case study: Oxford-style debates. Oxford-style debates provide a setting that is particularly convenient for studying the effects of conversational flow. In this competitive debate format, two teams argue for or against a preset motion in order to persuade a live audience to take their position. The audience votes before and after the debate, and the winning team is the one that sways more of the audience towards its view. This setup allows us to focus on the effects of conversational flow since it disentangles them from the audience's prior leaning.[1]

The debate format involves an opening statement from the two sides, which presents an overview of their arguments before the discussion begins. This allows us to easily identify talking points held by the participants prior to the interaction, and consider them separately from points introduced spontaneously to serve the discussion.

This work is taking steps towards better modeling of conversational dynamics, by: (i) introducing a debate dataset with rich metadata (Section 2), (ii) proposing a framework for tracking the flow of ideas (Section 3), and (iii) showing its effectiveness in a predictive setting (Section 4).

2 Debate Dataset: Intelligence Squared

In this study we use transcripts and results of Oxford-style debates from the public debate series "Intelligence Squared Debates" (IQ2 for short).[2] These debates are recorded live, and contain motions covering a diversity of topics ranging from for-

[1] Other potential confounding factors are mitigated by the tight format and topic enforced by the debate's moderator.

[2] http://www.intelligencesquaredus.org

eign policy issues to the benefits of organic food. Each debate consists of two opposing teams—one for the motion and one against— of two or three experts in the topic of the particular motion, along with a moderator. Each debate follows the Oxford-style format and consists of three rounds. In the *introduction*, each debater is given 7 minutes to lay out their main points. During the *discussion*, debaters take questions from the moderator and audience, and respond to attacks from the other team. This round lasts around 30 minutes and is highly interactive; teams frequently engage in direct conversation with each other. Finally, in the *conclusion*, each debater is given 2 minutes to make final remarks.

Our dataset consists of the transcripts of all debates held by IQ2 in the US from September 2006 up to September 2015; in total, there are 108 debates.[3] Each debate is quite extensive: on average, 12801 words are uttered in 117 turns by members of either side per debate.[4]

Winning side labels. We follow IQ2's criteria for deciding who wins a debate, as follows. Before the debate, the live audience votes on whether they are for, against, or undecided on the motion. A second round of voting occurs after the debate. A side wins the debate if the difference between the percentage of votes they receive post- and pre-debate (the "delta") is greater than that of the other side's. Often the debates are quite tight: for 30% of the debates, the difference between the winning and losing sides' deltas is less than 10%.

Audience feedback. We check that the voting results are meaningful by verifying that audience reactions to the debaters are related to debate outcome. Using laughter and applause received by each side in each round[5] as markers of positive reactions, we note that differences in audience reception of the two sides emerge over the course of the debate. While both sides get similar levels of reaction during the introduction, winning teams tend to receive more laughter during the discussion ($p < 0.001$)[6] and more applause during the conclusion ($p = 0.05$).

Example debate. We will use a debate over the motion "Millennials don't stand a chance" (henceforth *Millennials*) as a running example.[7] The For side won the debate with a delta of 20% of the votes, compared to the Against side which only gained 5%.

3 Modeling Idea Flow

Promoting one's own points and addressing the opponent's points are two primary debating strategies. Here we introduce a methodology to identify these strategies, and use it to investigate their usage and effect on a debate's outcome.[8]

Identifying talking points. We first focus on ideas which form the basis of a side's stance on the motion. We identify such *talking points* by considering words whose frequency of usage differs significantly between the two teams during the introduction, before any interaction takes place. To find these words, we use the method introduced by Monroe et al. (2008) in the context of U.S. Senate speeches. In particular, we estimate the divergence between the two sides' word-usage in the introduction, where word-usage is modeled as multinomial distributions smoothed with a uniform Dirichlet prior, and divergence is given by log-odds ratio. The most discriminating words are those with the highest and lowest z-scores of divergence estimates. For a side X, we define the set of talking points $\mathcal{W}_X$ to be the k words with the highest or lowest z-scores.[9] We distinguish between X's *own* talking points $\mathcal{W}_X$, and the *opposing* talking points $\mathcal{W}_Y$ belonging to its opponent Y. These are examples of talking points for the "Millennials" debate:

Side	Talking points
For	debt, boomer, college, reality
Against	economy, volunteer, home, engage

The flow of talking points. A side can either promote its own talking points, address its opponent's points, or steer away from these initially salient

[3] We omitted one debate due to pdf parsing errors.

[4] The processed data is available at http://www.cs.cornell.edu/~cristian/debates/.

[5] Laughter and applause are indicated in the transcripts.

[6] Unless otherwise indicated, all reported p-values are calculated using the Wilcoxon signed-rank test.

[7] http://www.intelligencesquaredus.org/debates/past-debates/item/1019-millennials-dont-stand-a-chance

[8] In the subsequent discussion, we treat all utterances of a particular side as coming from a single speaker and defer modeling interactions within teams to future work.

[9] In order to focus on concepts central to the sides' arguments, we discard stopwords, perform stemming on the text, and take $k = 20$. We set these parameters by examining one subsequently discarded debate.

Talking point	volunteer	boomer
Introduction	AGAINST: [millennials] **volunteer** more than any generation. 73 percent of millennials **volunteered** for a nonprofit in 2012. And the percentage of [students] believing that it's [...] important to help people in need is [at the highest level] in 40 years.	FOR: [*referring to college completion rate*] the **boomer** generation is now [at] 32 percent. [Millennials] are currently at [...] 33 percent. So this notion that [millennials] have more education at this point in time than anybody else is not actually true.
Discussion	FOR: I'd make the argument [that] **volunteering** [is done] for exntrinsic [sic] reasons. So, it's done for college applications, or it's done because it's a requirement in high school.	FOR: It stinks to be young, having gone through what your generation [*referring to millennials*] has gone through. But keep in mind that [...] the **boomers** [...] have gone through the same.

Table 1: Example talking points used throughout the "Millennials" debate. Each talking point belongs to the side uttering the first excerpt, taken from the introduction; the second excerpt is from the discussion section. In the first example, the For side addresses the opposing talking point **volunteer** during the discussion; in the second example the For side refers to their own talking point **boomer** and recalls it later in the discussion.

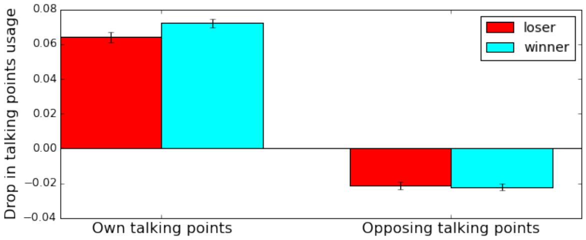

Figure 1: The start of the debate's interactive stage triggers a drop in self-coverage (> 0, indicated by leftmost two bars) and a rise in opponent-coverage (< 0, indicated by rightmost bars), with eventual winners showing a more pronounced drop in self-coverage (comparing the two bars on the left).

ideas altogether. We quantify the use of these strategies by comparing the airtime debaters devote to talking points. For a side X, let the *self-coverage* $f_r(X, X)$ be the fraction of content words uttered by X in round r that are among their own talking points $\mathcal{W}_X$; and the *opponent-coverage* $f_r(X, Y)$ be the fraction of its content words covering opposing talking points $\mathcal{W}_Y$.

Not surprisingly, we find that self-coverage dominates during the discussion ($f_{\text{Disc}}(X, X) > f_{\text{Disc}}(X, Y)$, $p < 0.001$). However, this does not mean debaters are simply giving monologues and ignoring each other: the effect of the interaction is reflected in a sharp drop in self-coverage and a rise in opponent-coverage once the discussion round begins. Respectively, $f_{\text{Disc}}(X, X) < f_{\text{Intro}}(X, X)$ and $f_{\text{Disc}}(X, Y) > f_{\text{Intro}}(X, Y)$, both $p < 0.001$. Examples of self- and opponent-coverage of two talking points in the "Millennials" debate from the introduction and discussion are given in Table 1.

Does the change in focus translate to any strategic advantages? Figure 1 suggests this is the case: the drop in self-coverage is slightly larger for the side that eventually wins the debate ($p = 0.08$). The drop in the sum of self- and opponent-coverage is also larger for winning teams, suggesting that they are more likely to steer away from discussing *any* talking points from either side ($p = 0.05$).

Identifying discussion points. Having seen that debaters can benefit by shifting away from talking points that were salient during the introduction, we now examine the ideas that spontaneously arise to serve the discussion. We model such *discussion points* as words introduced to the debate during the discussion by a debater and adopted by his opponents at least twice.[10] This allows us to focus on words that become relevant to the conversation; only 3% of all newly introduced words qualify, amounting to about 10 discussion points per debate.

The flow of discussion points. The adoption of discussion points plays an important role in persuading the audience: during the discussion, eventual winners adopt more discussion points introduced by their opponents than eventual losers ($p < 0.01$). Two possible strategic interpretations emerge. From a topic control angle (Nguyen et al., 2014), perhaps losers are more successful at imposing their discussion points to gain control of the discussion. This view appears counterintuitive given work linking topic control to influence in other settings (Planalp and Tracy, 1980; Rienks et al., 2006).

[10] Ignoring single repetitions discards simple echoing of words used by the previous speaker.

AGAINST: I would say [millennials] are effectively moving towards goals [...] it might seem like immaturity if you don't actually talk to millennials and look at the **statistics**. FOR: –actually, the numbers are showing [...] that it's worsening [...] Same **statistics**, dreadful **statistics**.
AGAINST: [...] there's a incredible [sic] advantage that millennials have when it comes to social media [...] because we have an understanding of that landscape as **digital** natives [...] FOR: Generation X [...] is also known as the **digital** generation. The companies [...] that make you **digital** natives were all founded by [...] people in generation X. It's simply inaccurate every time somebody says that the millennial generation is the only generation [...]

Table 2: Example discussion points introduced by the Against side in the "Millennials" debate. For each point, the first excerpt is the context in which the point was first mentioned by the Against side in the discussion, and the second excerpt shows the For side challenging the point later on.

An alternative interpretation could be that winners are more active than losers in contesting their opponents' points, a strategy that might play out favorably to the audience. A post-hoc manual examination supports this interpretation: 78% of the valid discussion points are picked up by the opposing side in order to be challenged;[11] this strategy is exemplified in Table 2. Overall, these observations tying the flow of discussion points to the debate's outcome suggest that winners are more successful at using the interaction to engage with their opponents' ideas.

4 Predictive Power

We evaluate the predictive power of our flow features in a binary classification setting: predict whether the For or Against side wins the debate.[12] This is a challenging task even for humans, thus the dramatic reveal at the end of each IQ2 debate that partly explains the popularity of the show. Our goal here is limited to understanding which of the flow features that we developed carry predictive power.

Conversation flow features. We use all conversational features discussed above. For each side X we include $f_{\text{Disc}}(X, X)$, $f_{\text{Disc}}(X, Y)$, and their sum. We also use the drop in self-coverage given by subtracting corresponding values for $f_{\text{Intro}}(\cdot, \cdot)$, and the number of discussion points adopted by each side. We call these the *Flow* features.

Baseline features. To discard the possibility that our results are simply explained by debater verbosity, we use the number of words uttered and number of turns taken by each side (*length*) as baselines. We also compare to a unigram baseline (*BOW*).

Audience features. We use the counts of applause and laughter received by each side (described in Section 2) as rough indicators of how well the audience can foresee a debate's outcome.

Prediction accuracy is evaluated using a leave-one-out (LOO) approach. We use logistic regression; model parameters for each LOO train-test split are selected via 3-fold cross-validation on the training set. To find particularly predictive flow features, we also try using univariate feature selection on the flow features before the model is fitted in each split; we refer to this setting as *Flow**.[13]

We find that conversation flow features obtain the best accuracy among all listed feature types (Flow: 63%; Flow*: 65%), performing significantly higher than a 50% random baseline (binomial test $p < 0.05$), and comparable to audience features (60%). In contrast, the length and BOW baselines do not perform better than chance. We note that *Flow* features perform competitively despite being the only ones that do not factor in the concluding round.

The features selected most often in the Flow* task are: the number of discussion points adopted (with positive regression coefficients), the recall of talking points during the discussion round (negative coefficients), and the drop in usage of own talking points from introduction to discussion (positive coefficients). The relative importance of these features, which focus on the interaction between teams, suggests that audiences tend to favor debating strategies which emphasize the discussion.

[11] Three annotators (including one author) informally annotated a random sample of 50 discussion points in the context of all dialogue excerpts where the point was used. According to a majority vote, in 26 cases the opponents challenged the point, in 7 cases the point was supported, 4 cases were unclear, and in 13 cases the annotators deemed the discussion point invalid. We discuss the last category in Section 6.

[12] The task is balanced: after removing three debates ending in a tie, we have 52 debates won by For and 53 by Against.

[13] We optimize the regularizer (ℓ_1 or ℓ_2), and the value of the regularization parameter C (between 10^{-5} and 10^5). For Flow* we also optimize the number of features selected.

5 Further Related Work

Previous work on conversational structure has proposed approaches to model dialogue acts (Samuel et al., 1998; Ritter et al., 2010; Ferschke et al., 2012) or disentangle interleaved conversations (Elsner and Charniak, 2010; Elsner and Charniak, 2011). Other research has considered the problem of detecting conversation-level traits such as the presence of disagreements (Allen et al., 2014; Wang and Cardie, 2014) or the likelihood of relation dissolution (Niculae et al., 2015). At the participant level, several studies present approaches to identify ideological stances (Somasundaran and Wiebe, 2010; Rosenthal and McKeown, 2015), using features based on participant interactions (Thomas et al., 2006; Sridhar et al., 2015), or extracting words and reasons characterizing a stance (Monroe et al., 2008; Nguyen et al., 2010; Hasan and Ng, 2014). In our setting, both the stances and the turn structure of a debate are known, allowing us to instead focus on the debate's outcome.

Existing research on argumentation strategies has largely focused on exploiting the structure of monologic arguments (Mochales and Moens, 2011), like those of persuasive essays (Feng and Hirst, 2011; Stab and Gurevych, 2014). In addition, Tan et al. (2016) has examined the effectiveness of arguments in the context of a forum where people invite others to challenge their opinions. We complement this line of work by looking at the relative persuasiveness of participants in extended conversations as they exchange arguments over multiple turns.

Previous studies of influence in extended conversations have largely dealt with the political domain, examining moderated but relatively unstructured settings such as talk shows or presidential debates, and suggesting features like topic control (Nguyen et al., 2014), linguistic style matching (Romero et al., 2015) and turn-taking (Prabhakaran et al., 2013). With persuasion in mind, our work extends these studies to explore a new dynamic, the flow of ideas between speakers, in a highly structured setting that controls for confounding factors.

6 Limitations and Future Work

This study opens several avenues for future research. One could explore more complex representations of talking points and discussion points, for instance using topic models or word embeddings. Furthermore, augmenting the flow of content in a conversation with the speakers' linguistic choices could better capture their intentions. In addition, it would be interesting to study the interplay between our conversational flow features and relatively monologic features that consider the argumentative and rhetorical traits of each side separately. More explicitly comparing and contrasting monologic and interactive dynamics could lead to better models of conversations. Such approaches could also help clarify some of the intuitions about conversations explored in this work, particularly that engaging in dialogue carries different strategic implications from self-promotion.

Our focus in this paper is on capturing and understanding conversational flow. We hence make some simplifying assumptions that could be refined in future work. For instance, by using a basic unigram-based definition of discussion points, we do not account for the context or semantic sense in which these points occur. In particular, our annotators found that a significant proportion of the discussion points under our definition actually referred to differing ideas in the various contexts in which they appeared. We expect that improving our retrieval model will also improve the robustness of our idea flow analysis. A better model of discussion points could also provide more insight into the role of these points in persuading the audience.

While Oxford-style debates are a particularly convenient setting for studying the effects of conversational flow, our dataset is limited in terms of size. It would be worthwhile to examine the flow features we developed in the context of settings with richer incentives beyond persuading an audience, such as in the semi-cooperative environment of Wikipedia talk pages. Finally, our methodology could point to applications in areas such as education and cooperative work, where it is key to establish the link between conversation features and an interlocutor's ability to convey their point (Niculae and Danescu-Niculescu-Mizil, 2016).

Acknowledgements. We thank the reviewers and V. Niculae for their helpful comments, and I. Arawjo and D. Sedra for annotations. This work was supported in part by a Google Faculty Research Award.

References

Kelsey Allen, Giuseppe Carenini, and Raymond T Ng. 2014. Detecting disagreement in conversations using pseudo-monologic rhetorical structure. In *Proceedings of EMNLP*.

Micha Elsner and Eugene Charniak. 2010. Disentangling chat. *Computational Linguistics*, 36(3):389–409.

Micha Elsner and Eugene Charniak. 2011. Disentangling chat with local coherence models. In *Proceedings of ACL*.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of ACL*.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the article: Recognizing dialog acts in Wikipedia talk pages. In *Proceedings of EACL*.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of EMNLP*.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Dong Nguyen, Elijah Mayfield, and Carolyn P Rosé. 2010. An analysis of perspectives in interactive settings. In *Proceedings of the KDD 2010 Workshop on Social Media Analytics*.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, Deborah A Cai, Jennifer E Midberry, and Yuanxin Wang. 2014. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, 95(3):381–421.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational markers of constructive discussions. In *Proceedings of NAACL*.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic harbingers of betrayal: A case study on an online strategy game. In *Proceedings of ACL*.

Sally Planalp and Karen Tracy. 1980. Not to change the subject but: A cognitive approach to the management of conversation. *Communication Yearbook*, 4:680–690.

Vinodkumar Prabhakaran, Ajita John, and Dorée D. Seligmann. 2013. Who had the upper hand? Ranking participants of interactions based on their relative power. In *Proceedings of IJCNLP*.

Rutger Rienks, Dong Zhang, Daniel Gatica-Perez, and Wilfried Post. 2006. Detection and application of influence rankings in small group meetings. In *Proceedings of ICMI*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Proceedings of NAACL*.

Daniel M Romero, Roderick I Swaab, Brian Uzzi, and Adam D Galinsky. 2015. Mimicry is presidential: Linguistic style matching in presidential debates and improved polling numbers. *Personality and Social Psychology Bulletin*, 41(10):1311–1319.

Sara Rosenthal and Kathleen McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of SIGDIAL*.

Ken Samuel, Sandra Carberry, and K. Vijay-Shanker. 1998. Dialogue act tagging with transformation-based learning. In *Proceedings of ACL*.

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*.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of ACL*.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of EMNLP*.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of WWW*.

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

Lu Wang and Claire Cardie. 2014. A piece of my mind: A sentiment analysis approach for online dispute detection. In *Proceedings of ACL*.