Controversy Detection In Conversations Using Rhetorical Parsing

Kelsey Allen **Giuseppe Carenini** **Raymond Ng**
Department of Computer Science, University of British Columbia
Vancouver, BC, V6T 1Z4, Canada
{kelseyra, carenini, rng}@cs.ubc.ca

1 Introduction

How does controversy arise in conversation? Being able to detect controversy has a range of applications. For an online educator, dissent over a newly introduced topic may alert the teacher to fundamental misconceptions about the material. For a business, understanding disputes over features of a product may be helpful in future design iterations. By better understanding how controversy arises and propagates in a conversation, a social scientist may gain insight into how users influence each other through forums.

To date, most work in controversy detection has focused on either synchronous conversations such as meetings (Somasundaran et. al, 2007; Raaijmakers et. al, 2008), monologue conversations (sets of reviews (Popescu et. al, 2005), news articles (Choi et. al, 2010; Awadallah et. al, 2012)) or micro-blogging sites (Lin et. al, 2013). However, there has been relatively little work applied to asynchronous conversations, which cover blog forums such as Reddit and Slashdot, as well as e-mails. Additionally, previous research in domains such as twitter and online reviews has focused on incorporating variations of sentiment features (Lin et. al, 2013), or using known controversial terms/articles to determine the level of controversy in new articles (Dori-Hacohen and Allan, 2013).

In our work we extend previous research in two ways. We study controversy in asynchronous conversations as well as exploiting a new knowledge source for controversy detection: rhetorical structure. Rhetorical relations, such as "Contrast", "Topic-Comment" and "Elaboration" describe how clauses in a sentence are connected, how sentences in a paragraph are connected, and how sections of a document are connected. By applying a previously developed rhetorical parser (Joty et. al, 2013) to the domain of asynchronous conversations, we can build a *relation graph* which shows how posts from different users are connected (fig 1). This work uses features from this graph, including connectivity, relations between users, and relations within one user's post, to investigate controversy detection in the popular blog forum Slashdot. Rhetorical features are additionally compared to sentiment features, to show that they can provide significant improvement in controversy detection.

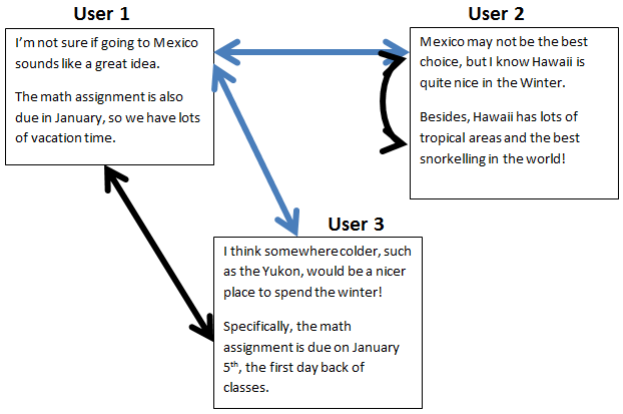

Figure 1: User relation graph for a sample asynchronous conversation. In this example, there are three users. Some of their sentences are connected by rhetorical relations. Blue lines denote a contrast relation, whereas black lines denote an elaboration relation.

As a third key contribution of this work, we have crowd-sourced a new blog corpus annotated for controversy at the topical level. This is described further in the next section.

2 A New Controversiality Blog Corpus

As little work has been performed in the domain of blog forums, we first annotated a controversiality corpus consisting of 95 topics from the popular blog forum Slashdot (www.slashdot.org). In a previous study (Joty et. al, 2013), 20 Slash-

Table 1: Sample of sentiment and rhetorical features to be included

Structural	Sentiment	Rhetorical
Author references	*Range*	*Number of intra-user connections per post*
Average length of post	*Number of negative comments*	*Number of inter-user connections per post*
Rate of posting	*Standard deviation*	*Average depth in inter-user relations*
Questions in first sentence of post	*Average polarity*	*Average branching ratio per post*

Table 2: Preliminary results

Feature set	Structural only (basic)	Basic+Sentiment	Basic+Rhetorical Sentiment	All features
Recall	**0.68**$\pm$0.02	**0.67**$\pm$0.03	**0.74**$\pm$0.02	**0.72**$\pm$0.01
Precision	**0.68**$\pm$0.02	**0.67**$\pm$0.03	**0.74**$\pm$0.02	**0.72**$\pm$0.01
F-score	**0.68**$\pm$0.02	**0.67**$\pm$0.03	**0.74**$\pm$0.02	**0.72**$\pm$0.01

dot articles were annotated for topic segmentation boundaries and labels by expert Slashdot contributors. Our dataset then consisted of the topics as determined by these human annotators. Overall, this consisted of 95 topics after filtering out topics with only one contributing user.

A Human Intelligence Task (HIT) was then developed using the crowdsourcing platform Crowdflower (`www.Crowdflower.com`). The objective of this task was to both develop a corpus for testing our weakly supervised system, as well as to investigate how easy it is to determine controversiality in an asynchronous conversation for human annotators. For training, users were shown 3 sample topics with their controversiality label. In each round, annotators were shown 5 topics, with a set of radio buttons for participants to choose "Yes", "No", or "Not sure" in response to the question "Controversial?". In order to limit the number of spam responses, users were shown test questions, which were obviously controversial or non-controversial topics. We required that users correctly identify 4 of these test topics before they were allowed to continue with the task. Users were also shown test questions throughout the task, which, if answered incorrectly, would reduce the amount of money they received for the task.

For each topic, five different judgements were obtained. Each topic was assigned a confidence score in the controversy level. This was calculated as

$$score = A \sum_{users} (\frac{test_{correct}}{test_{total}})_{user_i} \times (0, 0.5, 1)$$

where 0, 0.5 and 1 represent the answers "No", "Not sure" and "Yes" to the question "Controversial?" in the HIT, and A is a normalization factor. If the score is less than 0.5, its confidence would be $1 - score$ towards "Not controversial", whereas greater than 0.5 would be a confidence of $score$ towards "Controversial". The average confidence score across all topics was 0.73. Our corpus thus consists of 49 non-controversial and 46 controversial topics. Interestingly, 22 topics had confidence scores below 55%, which suggests that subtle controversy detection is a subjective and difficult task.

3 Feature list and Preliminary Results

In the interest of comparing different types of features for controversy analysis, we have explored sentiment features, conversational features, rhetorical features, and structural features. A sampling of these features is shown in table 1.

Previous work has found referrals (Mishne and Glance, 2006), negative polarity, range of sentiment, questions and discourse markers (Somasundaran and Wiebe, 2009) to be of most importance to controversy detection. In addition to the features above, some conversational features taken from the Fragment Quotation Graph (Joty et. al, 2013), as well as "Rhetorical Sentiment" features (which specify sentiment scores for text associated with each rhetorical relation), were added. Using a random forest classifier with 10,000 trees, and 50 features selected in each tree, we find that rhetorical structure does improve controversy detection. Additionally, if only rhetorical structure features are considered, the F-score is **0.72** $\pm$ 0.01, showing that rhetorical structural features alone can allow for relatively good controversy detection.

Overall, this analysis shows that rhetorical structure can significantly improve controversy detection in the conversational domain. Future work will involve trying different classifiers including SVM and logistic regression, as well as modifying the analysis to perform finer grained controversy classification.

References

Swapna Somasundaran, Josef Ruppenhofer, Janyce Wiebe. 2007. Detecting Argument and Sentiment in Meetings. SIGdial Workshop on Discourse and Dialogue.

Stephan Raaijmakers, Khiet Truong, Theresa Wilson. 2008. Multimodal subjectivity analysis of multiparty conversation. In *Proceedings of EMNLP*, pages 466-474.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting Product Features and Opinions from Reviews. In *Proceedings of HLT/EMNLP*, pages 339-346.

Y. Choi, Y. Jung, and S.H. Myaeng. 2010. Identifying controversial issues and their sub-topics in news articles. In *Proceedings of PAISI*, pages 140-153.

Rawia Awadallah, Maya Ramanath, Gerhard Weikum. 2012. Harmony and Dissonance: Organizing the Peoples Voices on Political Controversies. In *Proceedings of WSDM*, pages 523-532.

Ching-Sheng Lin, Samira Shaikh, Jennifer Stromer-Galley, Jennifer Crowley, Tomek Strzalkowski, Veena Ravishankar. 2013. Topical Positioning: A New Method for Predicting Opinion Changes in Conversation. In *Proceedings of LASM*, pages 41-48.

Shiri Dori-Hacohen and James Allan. 2013. Detecting Controversy on the Web. In *Proceedings of CIKM*, pages 1845-1848.

Shafiq Joty, Giuseppe Carenini, Raymond Ng and Yashar Mehdad. 2013. Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis. In *Proceedings of ACL*.

Shafiq Joty, Giuseppe Carenini and Raymond Ng. 2013. Topic Segmentation and Labeling in Asynchronous Conversations. *Journal of AI Research*, Vol. 47, Page 521-573.

Gilad Mishne and Natalie Glance. 2006. Leave a reply: An analysis of weblog comments. In *Proceedings of WWW*.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of ACL*, pages 226 - 234.