A Preliminary Study of Disputation Behavior in Online Debating Forum

Zhongyu Wei[1,2], **Yandi Xia**[1], **Chen Li**[1], **Yang Liu**[1]
Zachary Stallbohm[1], **Yi Li**[1] and **Yang Jin**[1]
[1]Computer Science Department, The University of Texas at Dallas
Richardson, Texas 75080, USA
[2]School of Data Science, Fudan University, Shanghai, P.R.China
{zywei,yandixia,chenli,yangl,stallbohm,yili,yangjin}@hlt.utdallas.edu

Abstract

In this paper, we propose a task for quality evaluation of disputing argument. In order to understand the disputation behavior, we propose three sub-tasks, detecting disagreement hierarchy, refutation method and argumentation strategy respectively. We first manually labeled a real dataset collected from an online debating forum. The dataset includes 45 disputing argument pairs. The annotation scheme is developed by three NLP researchers via annotating all the argument pairs in the dataset. Two under-graduate students are then trained to annotate the same dataset. We report annotation results from both groups. Then, another larger dataset was annotated and we show analysis of the correlation between disputing quality and different disputation behaviors.

1 Introduction

With the popularity of the online debating forum such as idebate[1], convinceme[2] and createdebate[3], researchers have been paying increasing attention to analyze debating content, including identification of arguing expressions in online debate (Trabelsi and Zaıane, 2014), recognition of stance in ideological online debates (Somasundaran and Wiebe, 2010; Hasan and Ng, 2014; Ranade et al., 2013b), and debate summarization (Ranade et al., 2013a). However, there is still little research about quality evaluation of debating content.

Tan et al. (2016) and Wei and Liu (2016) studied the persuasiveness of comments in sub-reddit *change my view* of Reddit.com. They evaluated the effectiveness of different features for the prediction of highly voted comments in terms of delta score and karma score respectively. Although they considered some sorts of argumentation related features, such features are merely based on lexical similarity, without modeling persuasion behaviors.

Figure 1: A disputation example from createddebate.com (The debating topic is "Should the Gorilla have died?")

In this paper, we focus on a particular action in the online debating forum, i.e., *disputation*. Within debate, disputation happens when a user disagrees with a specific comment. Figure 1 gives a disputation example from the online debating forum *createdebate*. It presents an original argument and an argument disputing it. Our study aims to evaluate the quality of a disputing comment given its original argument and the discussed topic. In order to have a deep understanding of disputation, we analyze disputation behavior via three sub-tasks, including disagreement hierarchy identification, refutation method identification and argumentation strategy identification.

We first manually labeled a small amount of data collected from createdebate.com. It includes 8 debate threads related to different topics. We

[1]http://idebate.org/
[2]http://convinceme.net
[3]http://www.createdebate.com/

extracted all the 45 disputing pairs from these threads. Each pair contains two arguments and the second one disputes the first one. Three NLP researchers (the first three authors of the paper) first developed a rough version of annotation scheme and they annotated all the argument pairs. Based on the annotation feedback and discussioins, they modified the scheme. Two native English speakers are then trained to annotate the same dataset. Further, we asked one annotator with better performance in previous step to annotate a larger set of data. We then analyze the correlation between disputing quality and different disputation behaviors. We will introduce annotation schema in Section 2 and then report the annotation result in Section 3. We conclude the paper in Section 4.

2 Annotation Schema

Our annotation is performed on a pair of arguments from opposite sides of a specific topic. In each pair, the second argument disputes the first one. Any of them can hold the "supportive" stance to the discussed topic. We define four annotation tasks: disagreement hierarchy (DH), refutation method (RM), argumentation strategy (AS) and disputing quality (DQ). The first three are proposed to understand the disputation behavior. In the disputing comment, DH indicates how the disagreement is expressed, RM describes which part of the original argument is attacked, and AS shows how the argument is formed.

2.1 Disagreement Hierarchy

In order to identify how users express their disagreement to the opposite argument, we borrowed the disagreement hierarchy from Paul Graham[4]. We modified the original version of the theory by combining some similar categories and proposed a four-level hierarchy. The definition of different types of DH is shown below. Examples of disputing comments with different disagreement hierarchies are shown in Table 1.

a) **DH-LV1: Irrelevance**. The disagreement barely considers the content of the original argument.

b) **DH-LV2: Contradiction**. The disagreement simply states the opposing case, with little or no supporting evidence.

[4] http://paulgraham.com/disagree.html

Table 1: Examples for disagreement hierarchy

original argument
I strongly feel age for smoking and drinking should not be lowered down as it can disturb the hormonal balance of the body!
disputing argument
DH-LV1: Irrelevance
Wat???? You are an idiot! I would definitely give you a down vote!
DH-LV2: Contradiction
I do not think this correct, it is impossible to be accepted.
DH-LV3: Target Losing Argument
So this age 21 thing is really stupid cause like i said minors still get hold to alcoholic beverages. (Age limit is non-sense because teen-age can always have alcohol.)
DH-LV4: Refutation
Getting involved in a war will also hurt your body as drinking and smoking, but the age limit is 18 instead of 21.

c) **DH-LV3: Target Losing Argument**. The disagreement is contradiction plus reasoning and/or evidence. However, it aims at something slightly different from the original argument.

d) **DH-LV4: Refutation**. Refutation is a counter-argument quoting content from the original argument. The quoting can be either explicit or implicit.

2.2 Refutation Method

When a disputing comment is labeled as *refutation*, we will further identify its refutation method. This sub-task is proposed to indicate what aspect of the original argument is attacked by the disputing one. Three categories are given for this sub-task according to the theory of *refutation methods* proposed by Freeley and Steinberg (2013). Examples for disputing comments using different refutation methods are shown in Table 2.

a) **RM-F: refute fallacy**. Refutation is performed by attacking the fallacy of the original argument. This usually happens when the target of the attack is the correctness of the claim itself in the original argument.

b) **RM-R: refute reasoning**. Refutation is performed by attacking the reasoning process demonstrated in the original argument.

c) **RM-E: refute evidence**. Refutation is performed by attacking the correctness of the evidence given in the original argument.

2.3 Argumentation Strategy

To dispute the original argument, the users will form their own argument. Argumentation strategies have been studies in both The Toulmin Model

Table 2: Examples for refutation methods (OA: original argument; DA: disputing argument)

RM-F: refute fallacy
OA: Humans are not animal's and dont say that we evolved from monkeys because we did not DA: dont say that we evolved from monkeys because we did not http://en.wikipedia.org/wiki/Human_evolution And there's a long list of references and further reading down there.
RM-R: refute reasoning
OA: There is supposed to be equal protection under the law. If we give some couples benefits for being together we need to give it to the rest. DA: Talking about the 14th Amendment's Equal Protection Clause? That was talking about slavery.
RM-E: refute evidence
OA: Dont say that we evolved from monkeys because we did not http://en.wikipedia.org/wiki/Human_evolution And there's a long list of references and further reading down there. DA: Evolution is fake God made you retard learn it. Also Wikipedia is soooooooooo wrong random people put stuff in there and the creator does not even care Yah Fools

of Argumentation[5] and the work of Walton et al. (2008). In our research, we employ the classification version from Toulmin because it is much simpler. Six categories are used to indicate the argumentation strategy used in the disputing argument. Note that this label should be given based on user's intention instead of the quality of the argument. For example, users might choose inappropriate evidence to support the disputing claim. We will still treat it as *generalization*. Examples of arguments with different argumentation strategies are shown in Table 3.

a) **Generalization**. Argument by generalization assumes that a number of examples can be applied more generally.

b) **Analogy**. Argument by analogy examines alternative examples in order to prove that what is true in one case is true in the other.

c) **Sign**. Argument by sign asserts that two or more things are so closely related that the presence or absence of one indicates the presence or absence of the other.

d) **Cause**. Argument by cause attempts to establish a cause and effect relationship between two events.

e) **Authority**. Argument by authority relies on the testimony and reasoning of a credible source.

f) **Principle**. Argument by principle locates a principle that is widely regarded as valid and shows that a situation exists in which this principle applies.

g) **Other**. When no above-mentioned argumentation strategy is identified, we label it as other.

[5] http://www-rohan.sdsu.edu/~digger/305/toulmin_model.htm

Table 3: Examples for argumentation strategy

Generalization
Look at alan turing; government data collection on him and his homosexual tendencies led to his suicide.
Analogy
What has worked for drug decriminalization in the Netherlands should work in the United States.
Sign
Where there's fire, there's smoke.
Cause
Beer causes drunkenness, or that drunkenness can be caused by beer.
Authority
As stated by Wikipedia: human is evolved from animal.
Principle
As it says, "there is a will, there is a way".

2.4 Debating quality evaluation

We are also interested in the general quality of the disputing comment. We use three categories: *bad debate*, *reasonable debate* and *good debate*. The label should be assigned based on the content of the disputing argument instead of annotators' personal preference to the topic.

a) **Bad debate**. The disagreement is irrelevant or simply states its attitude without any support; the support or reasoning or fallacy is not reasonable.

b) **Reasonable debate**. The disagreement is complete including contradiction and related supportive evidence or reasoning. However, the argument might be attacked easily.

c) **Good debate**. The disagreement contains contradiction and related supportive evidence or reasoning. Besides, this argument is good and persuasive to some extent.

3 Annotation Result

The annotation is performed on the variantA dataset[6] provided by the 3rd workshop on argumentation mining collected from createdebate.com. In such forum, each debating thread is about a particular topic and users can initialize a comment with a specific stance. Besides starting a comment, users can also reply to a comment with an intention of supporting, disputing or clarifying.

We first work on one subset of the data, namely *dev* to develop our annotation scheme and analyze the annotation performance of two laymen annotators. The statistics of the original *dev* set are given in Table 4. As we can see, more than half of the comments are disputing ones. We extract all disputing comments together with their original comment to form argument pairs as the first batch of

[6] Please contact authors for the annotated dataset.

Table 4: Statistics of the dev dataset in VariantA from createdebate.com

Thread #	8
avg comment #	10.25
avg initial comment #	3.00
avg disputation comment #	5.63
avg support comment #	1.38
avg clarify comment #	0.25
unique user #	6.7
avg length of initial comments	87.16
avg length of disputation comments	67.02
avg length of comments	69.24

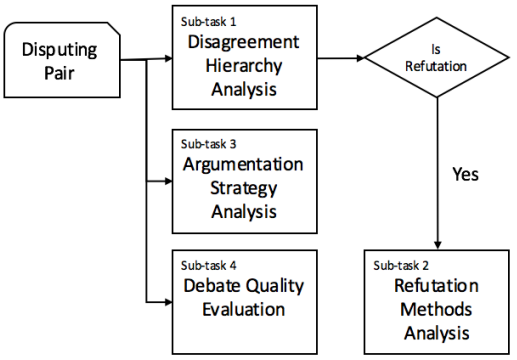

Figure 2: Workflow of the annotation task

our experiment dataset *batch-1*, 8 threads and 45 pairs of arguments in total.

We also analyze the relationship between different disputation behaviors and the quality of the disputing argument. To make this correlation analysis more convincible and also to motivate follow-up research for disputation analysis in the online forum, we collected another batch of annotation on the larger dataset *batch-2* from another two sub-sets of variantA (i.e., *test* and *crowdsourcing*). This batch contains 20 new topics including 93 pairs of disputing arguments. The correlation analysis is then performed on the combination of *batch-1* and *batch-2*.

3.1 Annotation Result of Expert and Layman on Batch-1

Three NLP researchers work together to define the annotation scheme via annotating all the argument pairs in *batch-1*. Two undergraduate students are then hired to annotate the same set of data given two days to finish all the annotation task. A half an hour training session is used for introducing the annotation scheme and demonstrating the annotation process via two samples. The work flow of the annotation is shown in Figure 2. Annotators are given the entire thread of the debating to have a background of the discussion related to this topic.

We first look at the label distribution on all the four annotation tasks based on experts' opinion on *batch-1*. The annotation scheme changes during the annotation process via discussion, we thus are not able to provide agreement between experts. For the three disputation behavior annotation tasks, experts finalize the label after discussion. For the disputing quality evaluation, experts agree on the label for *bad debate* but had different opinions about *good* and *reasonable* ones, since these are subjective. Therefore, for general quality annotation we take the majority. Table 5 shows the detail of the annotation results. For disagreement hierarchy, 36 out 45 (80%) disputation are *refutation*. For refutation method, 20 (44%) disputing comments refute fallacy directly, while 7 (18%) and 9 (20%) refute evidence and reasoning respectively. For argumentation strategy, 20 (44%) disputing comments do not use any specified methods. *Generalization* is the most popular one while no *sign* and *principle* are found. For the disputing quality, more than half of the comments are labeled as reasonable. Only 10 (22%) are labeled as good.

We then analyze the annotation result for two laymen annotators using experts' opinion as ground truth on *batch-1*. Generally speaking, the disputation behavior annotation is difficult for laymen. With only half an hour training, the performance of both annotators is not very good for labeling the four tasks. For disagreement hierarchy, annotators seem to have problems to distinguish *target losing argument* and *refutation*. Annotator-1 mis-labels too many instances as *ttarget losing argument* while annotator-2 gives only 1 such annotation. The lowest accuracy comes from refutation method identification. This is because the task requires deep understanding and analysis of argument. For disputing quality evaluation, it is easier for annotators to identify the *bad* argument. Distinguishing *good* and *reasonable* disputing is much more difficult. This is because the difference between them is very subjective.

3.2 Correlation of Disputation Behavior and Disputing Quality

With the same strategy, we further construct and annotate the second batch of experiment dataset *batch-2*. *Annotator-1* worked for this. Before the annotation, we review the error annotation

Table 5: Annotation results

Annotation Type		Batch-1									Batch-2
		Expert	Annotator-1				Annotator-2				Annotator-1
		#	#	precision	recall	F-1	#	precision	recall	F-1	#
DH	DH-LV1	2 (1%)	3	0.667	1.000	0.800	1	1.000	0.500	0.667	4 (4%)
	DH-LV2	1 (2%)	2	0.500	1.000	0.667	2	0.500	1.000	0.667	5 (5%)
	DH-LV3	6 (13%)	12	0.417	0.833	0.556	1	0.000	0.000	0.000	12 (13%)
	DH-LV4	36 (80%)	27	1.000	0.750	0.857	41	0.829	0.944	0.883	72 (77%)
RM	RM-E	7 (18%)	2	1.000	0.286	0.444	6	0.333	0.286	0.308	4 (4%)
	RM-R	9 (20%)	19	0.263	0.556	0.357	11	0.364	0.444	0.400	27 (29%)
	RM-F	20 (44%)	6	1.000	0.300	0.462	24	0.500	0.600	0.545	41 (44%)
AS	generalization	9 (20%)	6	0.667	0.667	0.667	6	0.167	0.167	0.167	5 (5%)
	analogy	5 (11%)	4	0.500	0.400	0.444	7	0.714	1.000	0.833	8 (9%)
	sign	0 (0%)	4	0.000	0.000	0.000	0	0.000	0.000	0.000	0 (0%)
	cause	6 (13%)	12	0.500	0.667	0.571	10	0.600	0.667	0.632	33 (35%)
	authority	5 (11%)	3	0.667	0.400	0.500	7	0.714	1.000	0.833	7 (8%)
	principle	0 (0%)	0	0.000	0.000	0.000	0	0.000	0.000	0.000	0 (0%)
	other	20 (44%)	16	0.813	0.650	0.722	15	0.800	0.600	0.686	40 (43%)
DQ	bad	12 (27%)	21	0.523	0.917	0.667	11	0.818	0.750	0.783	19 (20%)
	reasonable	23 (51%)	21	0.667	0.609	0.636	9	0.667	0.261	0.375	58 (62%)
	good	10 (22%)	3	0.000	0.000	0.000	25	0.280	0.700	0.400	16 (17%)

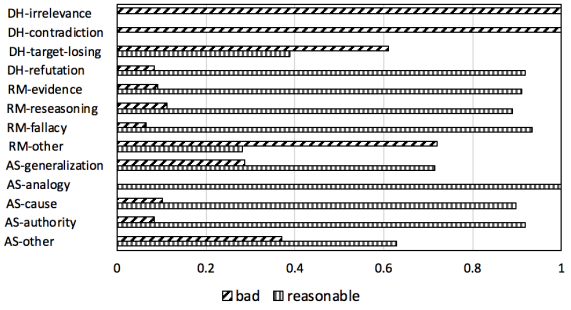

Figure 3: The correlation between disputation behaviors and disputing quality (binary setting) on *batch-1*+*batch-2*.

with the annotator to enhance his understanding about the annotation task. The annotation result of *batch-2* can be seen in Table 5. We then report the correlation result between disputation behaviors and disputing quality of the arguments on the combination of *batch-1* and *batch-2*.

For the correlation analysis, we report the label distribution in terms of disputing quality for arguments with different disputation labels. Considering the difference between a "good disputing" and a "reasonable disputing" is hard to decide, we treat both *reasonable* and *good* as *reasonable* to form a binary setting. Figure 3 shows the correlation between disputation behaviors and disputing quality. As we can see, all the arguments labeled as *DH-irrelevance* and *DH-contradiction* are *bad* ones, and 91.7% of *DH-refutation* arguments are *reasonable*. For argumentation strategy, *analogy* (100%), *cause* (89.7%) and *authority* (91.7%) are good indicators for *reasonable* arguments.

3.3 Discussion

We identified two major reasons for annotation errors after result analysis on *batch-1*. First, some categories within sub-tasks are difficult to distinguish in nature (e.g. target losing argument and refutation). Second, some disputing comments contain multiple claims and premises. This makes it difficult to identify the essential claim of the disputation. We believe we can improve the annotation performance in future work by: a) extend the time for training session and pick some representative samples for demonstration; b) modify the annotation scheme to avoid the ambiguity between categories; c) preprocess the disputing comment to identify the essential argument for better annotation.

4 Conclusion

In this paper, we analyzed the disputation action in the online debate. Four sub-tasks were proposed including disagreement hierarchy identification, refutation method identification, argumentation strategy identification and disputing quality evaluation. We labeled a set of disputing argument pairs extracted from a real dataset collected in createddebate.com and showed annotation results.

Acknowledgments

The work is partially supported by DARPA Contract No. FA8750-13-2-0041 and AFOSR award No. FA9550-15-1-0346.

References

Austin Freeley and David Steinberg. 2013. *Argumentation and debate*. Cengage Learning.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *EMNLP*, pages 751–762.

Sarvesh Ranade, Jayant Gupta, Vasudeva Varma, and Radhika Mamidi. 2013a. Online debate summarization using topic directed sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM.

Sarvesh Ranade, Rajeev Sangal, and Radhika Mamidi. 2013b. Stance classification in online debates by recognizing users? Intentions. In *SigDial*, pages 61–69.

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. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. *arXiv preprint arXiv:1602.01103*.

Amine Trabelsi and Osmar R Zaıane. 2014. Finding arguing expressions of divergent viewpoints in online debates. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL*, pages 35–43.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

Zhongyu Wei and Yang Liu. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *ACL*.