Structural Analysis of Rebuttals to Evaluate Argumentative Interaction in Parliamentary Debates

Masahiro Fukui[1*][0009−0000−6715−2509] and Satoshi Nakamura[1][0000−0003−3492−7093]

[1]Meiji University, 4-21-1 Nakano, Nakano-ku, Tokyo 164-8525, Japan
*onedanijo11@gmail.com

Abstract. This study introduces a structural framework for evaluating the quality of argumentative interaction in parliamentary debate. We proposed four hypotheses about rebuttal structures and defined corresponding features (Distance, Interval, Order, Rally). From a corpus of 20 English debate rounds with 1,573 ADUs and 679 rebuttal relations, we compared these features with human and LLM ratings. Regression analysis revealed a moderate correlation ($r = 0.609$), with Rally emerging as the most important predictor of interaction quality, followed by Distance and Interval, while Order showed limited explanatory power. To apply these insights in practice, we developed DebaTube, a visualization system that maps rebuttal structures to debate videos. A user study with experienced debaters confirmed that the system helps identify effective rebuttal patterns and improves exploration efficiency.

Keywords: Parliamentary debate · Rebuttal structure · Dialogue

1 Introduction

Parliamentary debate is a turn-based, impromptu format in which two teams argue for (proposition) or against (opposition) a motion to persuade judges. It is not only a competition but also a valuable educational practice where participants develop critical thinking and learn how to argue constructively [4]. However, due to the complexity of parliamentary debate, debaters often focus excessively on details of individual rebuttal and talk past each other. Nevertheless, most existing methods for debate analysis only evaluate the quality of individual rebuttals. Ruiz-Dolz et al. [6] pointed out that most prior research has focused on short-text debates and tends to oversimplify argumentative dynamics by isolating individual arguments. To address this, they and Hsiao et al. [3] have attempted to model rebuttal structures.

However, while these computational approaches have advanced the modeling of debate dynamics, they primarily focus on predicting debate winners, which provides only a limited view of argumentation quality. In parliamentary debate, this approach cannot fully capture the quality, as teams often win due

to opponent mistakes or unconstructive arguments rather than substantive engagement. Furthermore, this winner-prediction-based evaluation overlooks the dialogic and educational nature of parliamentary debate. Therefore, it is essential to evaluate debates not only by which team wins, but also by how effectively debaters engage with and build upon each other's arguments, which we term "argumentative interaction" in this paper.

In this study, we address the following research question: *What structural features of rebuttals indicate high-quality argumentative interaction in parliamentary debate?*

To explore this question, we propose a novel approach for evaluating the quality of argumentative interaction based on the rebuttal structures. Our analysis builds upon the concept of argumentative discourse units (ADUs), which are elemental textual segments that serve specific argumentative functions. We introduce four hypotheses grounded in the structural dynamics of rebuttal relations between ADUs.

The main contributions of this paper are as follows:

1. We constructed a manually annotated corpus of 20 English parliamentary debates with 1,573 ADUs and 679 rebuttal relations.
2. We proposed four structural features and, through expert–LLM combined evaluation, showed they moderately predict interaction quality ($r = 0.609$), with *Rally* as the strongest indicator.
3. We developed *DebaTube*, a visualization system linking rebuttal structures with debate videos, and confirmed its usefulness in a user study.

2 Modeling Parliamentary Debate

2.1 Key Terms and Hypotheses on Rebuttal Structure

To analyze rebuttal structures, we first define key terms. Following prior work in argumentation analysis [8], we adopt the concept of ADUs. Each ADU represents an argumentative unit that contains either a claim with reasons or a standalone claim. Here, Point of Information (POI), which is a question posed during opponents' speeches, is treated as an ADU since they typically present a single claim due to time constraints.

We define a rebuttal as a statement that responds to a specific argument made by the opposing team. Anticipating an opponent's claim or presenting a conflicting stance without reference to a particular statement is not considered a rebuttal. In our model, rebuttals are represented as directed edges between ADUs, forming a graph where nodes are ordered chronologically.

To investigate our research question, we propose four hypotheses grounded in prior argumentation studies, which show that temporal proximity between arguments and argument order [3], and the chains of counter-attacks [1, 6] are important factors in debate quality. While these studies primarily focus on winner prediction for either monologues or short-text debates, we extend their insights by adopting a more dialogical perspective and tailoring our approach

to the specific characteristics of parliamentary debate, where multiple topics are addressed simultaneously and well-organized arguments are particularly crucial. Based on these theoretical foundations, we propose the following hypotheses:

- **H1**: Rebuttals that target arguments from two or more speeches earlier tend to overlook recent content and create fragmented exchanges. A higher frequency of such distant rebuttals suggests lower interaction quality.
- **H2**: When rebuttals to the same point are spread across a speech with large intervals between them, the coherence of the exchange weakens. Larger intervals between repeated rebuttals indicate unproductive interactions.
- **H3**: Rebuttals that follow the original order of the opponent's arguments within a topic indicate more coherent and productive dialogue.
- **H4**: Longer and more frequent rebuttal rallies, sequences of multiple counter-rebuttals, reflect deeper engagement and higher interaction quality.

2.2 Definitions of Structural Features of Rebuttals

Based on the four hypotheses described in Section 2.1, we define structural features that quantify rebuttal characteristics for each debate round: *Distance*, *Interval*, *Order*, and *Rally*. Each corresponds to our hypotheses H1 to H4. We compute these features as follows:

- **Distance (H1)**: This feature calculates the proportion of rebuttals that target ADUs from at least two speakers earlier in the opposing team.
- **Interval (H2)**: For each ADU rebutted multiple times from the same speech, we calculate the interval between the first and last rebuttal, counting both endpoints (e.g., the first to second ADU counts as 2, the first to third ADU counts as 3). We then normalize this by the number of ADUs in the speech minus two (the maximum possible interval), where two is the minimum interval. This feature is the sum of all normalized gaps.
- **Order (H3)**: This feature measures rebuttal pairs from the same speech that either share the same source or have crossing edges. Here, a crossing is defined as when a later rebuttal targets an earlier ADU than the target of a preceding rebuttal (See Figure 1). The feature returns total rebuttals divided by crossing count.
- **Rally (H4)**: This feature counts pairs of rebuttals where one rebuttal's target is another rebuttal's source, then normalizes by dividing the count by the product of total rebuttals and total speeches.

3 Empirical Study: Structural Features and Argumentative Interaction Quality

3.1 Corpus Construction

We collected 20 videos of two-team parliamentary debates featuring experienced student debaters in practice rounds and tournament preliminaries. All speakers were Japanese high school or university students debating in English, following

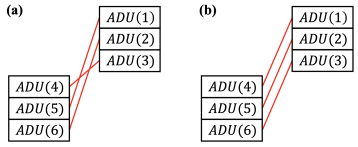

Fig. 1. Illustration of crossing edges in the Order calculation. Black blocks represent ADUs numbered chronologically, with ADUs 4–6 belonging to the same speech, and red lines indicate rebuttal relations. In (a), the rebuttal pairs (3→1, 4→2) and (4→2, 5→1) cross, resulting in a crossing count of 2. In (b), no crossings occur, so the crossing count is 0.

worldwide parliamentary debate rules. We specifically chose intermediate-level debates because they contain both strong and weak arguments. This variety allows us to analyze various argumentation characteristics.

The videos were transcribed using whisper-large-v2. The transcripts were then manually segmented into ADUs and annotated with rebuttal relations by the first author, an experienced parliamentary debater, as automated rebuttal detection methods for competitive debate have not been established. As a result, we constructed a corpus of 1,573 ADUs and 679 rebuttal relations.

3.2 Evaluation Method and Inter-rater Reliability

To evaluate the structural features defined in Section 2.2, we assessed the quality of argumentative interaction in each round with three raters: a human expert (10+ years of judging experience), a human non-expert (3+ years of debate experience without judging experience), and a large language model (LLM). While acknowledging that LLMs have limitations such as biases toward later speeches, recent studies show that LLMs are sufficiently capable for debate evaluation [5], which we consider acceptable given the exploratory nature of our study. For the LLM rater, we used OpenAI o3[1].

All raters reviewed videos of each debate round and evaluated the statement "Both teams demonstrated good argumentative interaction throughout the debate" using a four-point Likert scale: Agree (4), Rather agree (3), Rather disagree (2), and Disagree (1). For LLM ratings, we generated five outputs per round and used the mode as the final rating, with 90% of rounds showing consistency in at least 3 out of 5 ratings. Raters also provided open-ended comments for qualitative insights.

To establish our evaluation method, we assessed inter-rater reliability by calculating Cohen's Kappa using binary classifications (positive: scores 3-4; negative: scores 1-2). This preliminary check revealed moderate agreement between the expert and LLM ($\varkappa = 0.490$, with 60% perfect agreement on the four-point scale), while the non-expert showed poor agreement with both the

[1] Details of the prompts used are available at: `https://osf.io/ceugp`

Table 1. Performance comparison of regression models

Model	RMSE	MAE	Correlation Coefficient
Multiple Linear Regression	0.509	0.409	0.609
Ridge Regression	0.562	0.447	0.508
Lasso Regression	0.511	0.412	0.589

Table 2. Regression coefficients and feature importance

Feature	Multiple Linear	Ridge	Lasso	Importance (%)
Distance	-1.172	-0.530	-0.915	28.8
Interval	-1.030	-0.611	-0.914	25.3
Order	0.216	0.007	0.000	5.3
Rally	1.656	0.671	1.319	40.7

expert ($\varkappa = 0.175$) and LLM ($\varkappa$ = -0.056). Thus, we used the average of expert and LLM rating as our final quality score, excluding the non-expert rating.

3.3 Correlation Between Structural Features and Argument Quality

We verified the hypotheses by calculating features of rebuttal structures in Section 2.2 and examining the relationship between features and raters' evaluation through regression analysis. To ensure robust evaluation of model generalization performance, we employed leave-one-out cross-validation on our dataset of 20 samples. The feature values were normalized by dividing each value by the maximum feature value in the 20 rounds.

As Table 1 shows, Multiple Linear Regression achieved the best performance with RMSE of 0.509 and MAE of 0.409. The correlation coefficient between predicted and rater evaluation was 0.609, indicating a moderate positive correlation. Given an SD of 0.688, an RMSE/SD of 0.74 suggests practically acceptable predictive performance for subjective evaluation tasks.

Regarding feature importance, Table 2 shows that the *Rally* feature demonstrated the highest importance at 40.7%, followed by *Distance* (28.8%), *Interval* (25.3%), and *Order* (5.3%) features. Multiple linear regression showed particularly strong coefficients for *Rally* (1.656), *Distance* (1.172), and *Interval* (1.030). Additionally, both Lasso and Ridge regression eliminated the *Order* feature (Lasso: 0.000, Ridge: 0.007).

3.4 In-depth Analysis on the Accuracy of Models

As for model's accuracy, results described in Section 3.3 show that H4 (related to *Rally*) is strongly supported with the highest importance at 40.7%. Moreover, H1 (related to *Distance*) and H2 (related to *Interval*) are moderately supported with 28.8% and 25.3% importance, respectively. In contrast, H3 (related to *Order*) showed limited support with only 5.3% importance and was completely eliminated in Lasso regression.

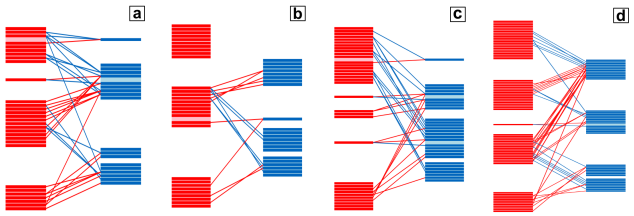

Fig. 2. Visualization of rebuttal structures with the (a) largest *Rally* (b) smallest *Distance* (c) largest *Distance* (d) third-largest *Order* value.

To understand these results, we provide an in-depth analysis of these structural features through visualization of rebuttal structures. Throughout this paper, we employ the visualizing format where block nodes represent ADUs and line edges represent rebuttals, with red indicating proposition's components and blue indicating opposition's components.

First, *Rally* demonstrated the strongest alignment with H4, suggesting a tendency for the hypothesis to be supported. The round in Figure 2a shows a round with the largest *Rally* value that received a rating of 3.0 out of 4, despite having the worst *Interval* and *Order* scores. This suggests that long rally indicates meaningful dialogue even when rebuttals are not well-organized.

Second, *Distance* tends to moderately support H1. Figure 2b presents a notable case with the smallest *Distance* and highest rating (3.5) though first opposition speaker did not rebut at all. This suggests continuous counterarguments are important. Furthermore, the feature also captures argumentative flaws. Figure 2c shows the round with the largest *Distance* value (rated 2.5), where the second proposition speaker barely spoke, forcing the opposition to concentrate rebuttals on the first speaker, collapsing the dialogue. Thus, *Distance* reflects both positive and negative aspects of dialogue.

Third, *Interval* demonstrated moderate alignment with H2. For instance, the round with the smallest *Interval* received good rating (3.0) despite average *Rally* value. This suggests that when teams concentrate their multiple attacks on the same argument, it compensates for unremarkable exchange frequency.

Last, *Order* showed minimal alignment with H3 and was eliminated by regularized methods. The model failed because individual speeches with many crossings disproportionately lowered the overall score. For example, the round in Figure 2d received the highest rating (3.5) but had the third-worst *Order* score, as strong early speeches with few crossings (especially 1st opposition and 2nd proposition) were overshadowed by later speeches with many crossings. Weighting crossings per speech could address this issue.

3.5 Short Summary

Our empirical analysis revealed that structural features of rebuttals, particularly *Rally* and *Distance*, can serve as reliable indicators of argumentative interaction

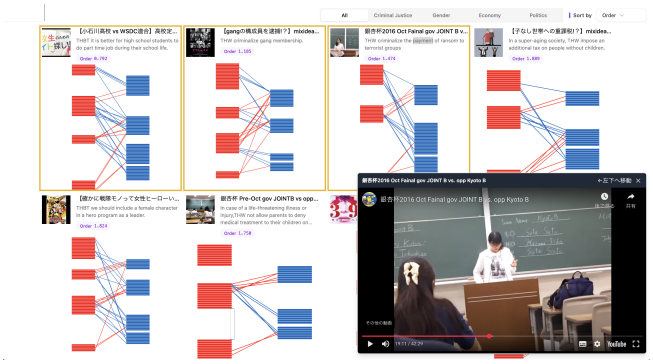

Fig. 3. User Interface of DebaTube. Users can sort rounds by features and pin the round and click any ADU node to jump to the scene.

quality. The predictive performance of our model suggests that debate quality can be captured through quantifiable structural patterns. Interestingly, while argumentative interactions may seem complex to quantify, simple structural metrics proved sufficient for evaluation. This indicates that beyond the detailed content of arguments, structural patterns provide useful cues for assessing debate quality. The quantifiable nature of these patterns raises the possibility of making them more accessible through visualization.

4 Application and User Study

Our empirical findings demonstrated that structural features can effectively indicate argumentative interaction quality. To explore how these insights about rebuttal structure can support debate learning in practice, we developed DebaTube[2], an interactive visualization system that leverages our structural features. While previous work has shown the value of argument visualization for understanding political debates [7] and improving discussion skills [9, 2], our system specifically targets parliamentary debate education by combining rebuttal structure visualization with video exploration.

DebaTube allows users to visually explore debate patterns using the rebuttal graph and navigate directly to specific argumentative exchanges by clicking ADU nodes, enabling debaters to overview features of many debate rounds at once and efficiently identify and study rounds with desirable argumentative patterns before watching the corresponding video segments.

We conducted a user study with five experienced debaters to evaluate how structural visualization aids debate exploration. Participants used our system to find instructional rounds for four scenarios (e.g., covering struggling teammates, balancing rebuttal targets). As a result, it turned out that some participants compared same speaker positions across rounds, while others carefully examined

[2] https://debatube.nkmr.io

the visualization before watching the videos. Results show our visualization helps users recognize effective rebuttal patterns and improves exploration efficiency.

5 Conclusion

This study introduced structural features of rebuttals as indicators of high-quality argumentative interaction in parliamentary debate. We proposed four features that describe how rebuttals are distributed, sequenced, and exchanged, offering an alternative to traditional winner-prediction-based methods. This research establishes a structural framework for evaluating argumentative interaction quality by analyzing all rebuttal relations throughout rounds and capturing dialogic dynamics.

Future work will address current limitations: recruiting multiple judges to establish more reliable ground truth and testing how the findings from this paper can be applied to argumentation in various formats, such as political debates and group discussions. Through these, we aim to not only enhance the educational value of competitive debate but also expand argumentation education more broadly by supporting learners' understanding of rebuttal structures.

References

1. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person matches. Artificial intelligence **77**(2), 321–357 (1995)
2. Guerraoui, Reisert, Inoue, Mim, Naito, Choi, Robbani, Wang, Inui: Teach me how to argue: A survey on NLP feedback systems in argumentation. In: Proceedings of the 10th Workshop on Argument Mining. 19–34 (2023)
3. Hsiao, F.H., Yen, A.Z., Huang, H.H., Chen, H.H.: Modeling inter round attack of online debaters for winner prediction. In: Proceedings of the ACM Web Conference 2022. 2860–2869 (2022)
4. Jodoi, K.: The effects of parliamentary debate as a pedagogy for argumentation in l1 and l2 contexts. Argumentation **39**, 147–163 (2024)
5. Liu, X., Liu, P., He, H.: An empirical analysis on large language models in debate evaluation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Volume 2. 470–487 (2024)
6. Ruiz-Dolz, H.G.: Automatic debate evaluation with argumentation semantics and natural language argument graph networks. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 6030–6040 (2023)
7. South, L., Schwab, M., Beauchamp, N., Wang, L., Wihbey, J., Borkin, M.A.: Debatevis: Visualizing political debates for non-expert users. In: 2020 IEEE Visualization Conference (VIS). 241–245 (2020)
8. Stab, C., Gurevych, I.: Parsing argumentation structures in persuasive essays. Computational Linguistics **43**, 619–659 (2017)
9. Xia, M., Zhu, Q., Wang, X., Nie, F., Qu, H., Ma, X.: Persua: A visual interactive system to enhance the persuasiveness of arguments in online discussion. Proceedings of the ACM on Human-Computer Interaction **6**, 1–30 (2022)