Towards Creating Analytic Dimensions for Evaluating the Quality of Debate Counter-Arguments

Wenzhi Wang[1,2] Paul Reisert[3] Naoya Inoue[4,2]
Shoichi Naito[1,2,5] Camélia Guerraoui[1] Keshav Singh[1] Kentaro Inui[1,2]
[1]Tohoku University [2]RIKEN [3]Beyond Reason [4]JAIST [5]Ricoh Company, Ltd.
{wang.wenzhi.r7,guerraoui.camelia.kenza.q4,singh.keshav.t4}@dc.tohoku.ac.jp
beyond.reason.sp@gmail.com naoya-i@jaist.ac.jp
shohichi.naitoh@jp.ricoh.com kentaro.inui@tohoku.ac.jp

Abstract

Evaluating the quality of argumentative texts is a challenging but exciting research topic which has gained attention over the years. In the context of debates, quality evaluation has been extensively researched and applied to top-level arguments but rarely to counter-arguments due to their complex nature. In this work, we tackle the task of argumentative quality assessment of counter-arguments (CA) in a debate. We first survey a set of debate rubrics and papers to find commonalities applicable to evaluating CAs and create four new analytic dimensions for assessing their quality. To test the feasibility of our dimensions, we employ crowdsourcing to evaluate CAs using our dimensions, and we report our preliminary results.

1 Introduction

Counter-arguments are an important way of constructing an argument, especially in the context of debates. One must first consider their opponent's argument both logically and rhetorically in order to construct an effective counter-argument. Such way of considering an opponent's argument can significantly help one improve their critical thinking skills. One means for constructing counter-arguments includes parliamentary debates which require critical analysis and rhetorical skills [1].

An example of a parliamentary debate is shown in Figure 1. In this debate, the Prime Minister makes their original argument (OA), and the Leader of the Opposition attacks the Prime Minister's point in their counter-argument (CA). After both parties complete their turns, their arguments are then evaluated by a judge who declares a winner.

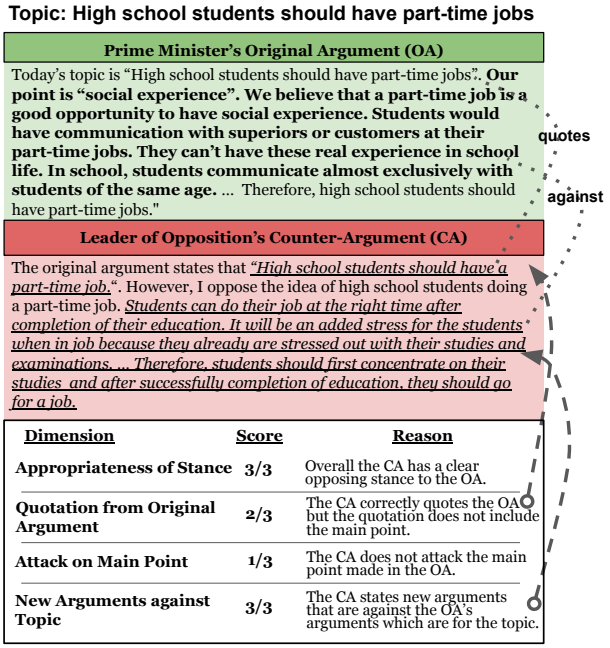

Figure 1 Overview of our analytic dimensions created for capturing the quality of the CA in response to the OA along with their respective scores collected via crowdsourcing.

Not only is this time-consuming for a judge, especially for a teacher in a debate class, but it has the possibility of introducing bias into the evaluation. Therefore, an automated approach to evaluating arguments in a debate is ideal.

Although many works have focused on the evaluation of the original argument (i.e. arguments that argue freely for or against the topic without basing on the logic of another argument) in a debate, little attention has been given to the evaluation of counter-arguments due to their dependency on the original argument during evaluation. In order to automatically evaluate such counter-arguments, a model built on top of a rich dataset of original arguments and counter-arguments consisting of evaluated scores along

with their reasoning is required. However, it remains an open issue as to what evaluation criteria are even required to automate the quality evaluation of counter-arguments.

There are many benefits to creating evaluation criteria for automatic CA evaluation. In the context of education, students learning debates could utilize the criteria scores output by a computational model given to their counter-argument, where the scores themselves could guide the learners towards improving their argumentation given reasons such as those shown in Figure 1. Furthermore, as mentioned a priori, the workload for judges (e.g. educators) and bias could significantly be reduced. To assist learners even further, both scoring the argument while simultaneously providing the spans in the OA and CA, where the arguments could be improved, could even further assist the learning in improving their argumentative skills.

Towards the ultimate goal of automatic evaluation of CAs, in this work, we focus heavily on creating new dimensions for evaluating CAs. We first collect a wide range of debate rubrics and find commonalities among them. Based on our findings, we create four new analytic dimensions (see Figure 1), each of which is assessed with a 3-scale score along with reasons. Finally, we test the feasibility of our dimensions via crowdsourcing.[1)] Overall, we discover that while our dimensions are new, there are still many challenges to overcome.

2 Related Work

Recently, the majority of works in the argumentation field focus on evaluating arguments on general dimensions, such as *persuasiveness* [2, 3, 4, 5, 6] which mostly focuses on whether an argument is strong, clear, supported by decent evidence; *appropriateness and content richness* [7, 8, 9], where *appropriateness* is defined as whether an argument is on-topic and has the right stance and *content richness* is broadly described as how many distinct aspects an argument covers; *plausibility* [7, 10] which is assessed in a sense of whether an argument contradicts the commonsense or in other words, whether a human will plausibly make it; *grammaticality* [6, 8, 7, 9], *coherence* [5], and *bias* [7]. However, those dimensions are too generic in the sense that they solely take into account the relation between an argument and a given prompt (topic) or the relation between argumentative units within an argument.

1) https://www.mturk.com

Table 1 New dimensions created for CA evaluation. Our new dimensions differ from previous dimensions in that they were created for CA in direct response to a topic (prompt) and an OA.

Appropriateness of Stance (AoS)	Whether a counter-argument has an appropriate stance. Overall, the stance should be against the given topic.
Quotation from the original argument (Quo)	Whether a counter-argument correctly quotes the main point from the original argument.
Attack on the main point (Att)	Whether a counter-argument directly attacks the main point in the original argument.
New arguments against the topic (Nat)	Whether a counter-argument provides its own arguments that are against the topic as further evidence to rebut the original argument.

Thus, they do not reflect properties that are inherent in the counter-argument that is in relation to the given topic as well as the original argument.

Wachsmuth et al. [11] proposed a taxonomy with 15 dimensions for argumentation quality assessment and annotated a corpus of stance-argument pairs based on those dimensions. The dimension *Global sufficiency*, which says argumentation is sufficient if it adequately rebuts those counter-arguments that can be anticipated, is most relevant to our work. Nevertheless, it cannot be directly applied to evaluating the counter-argument in the context of debates since it does not address the specific relation between the original argument and the counter-argument with the presence of the actual original argument.

We argue that analytic dimensions of quality for CAs in a setting of debates should consider more the relationship between the original argument and counter-argument. In this work, we attempt to create dimensions that are specific to evaluating CAs in a debate.

3 Creating Analytic Dimensions

As shown in Section 2, several dimensions for CAs exist. However, the dimensions are either too general or do not consider CAs in response to debate-level original arguments. In order to create quality dimensions that are specific to evaluating counter-arguments, we first explore debate rubrics for determining common properties amongst scores which can help in creating new analytic dimensions.

3.1 Existing debate rubrics

To obtain clues about which dimensions to create, we collected 7 publicly available debate rubrics[2)] and surveyed

2) The debate rubrics we surveyed can be found at https://github.com/oubunshitsu/debate_CA_assessment.

Table 2 Scoring rubrics for our newly created dimensions for counter-arguments (CA) in response to an original argument (OA).

Appropriateness of Stance (AoS)		**Quotation from the original argument (Quo)**	
Score	Score Reason	Score	Score Reason
1	Completely has the same stance as the OA, which is in support of the topic.	1	CA misquotes or does not quote the OA.
2	Simultaneously contains arguments that are for the given topic and arguments that are against the topic, which causes a mixed/unclear stance. May feel contradictory.	2	CA correctly quotes the OA (note: paraphrasing or summarizing the point is also acceptable, it does not necessarily have to be the exactly same sentence). However, the quotation does not include the main point made in the OA.
3	May or may not acknowledge points made in the OA as a concession strategy to enhance its own arguments. Overall, the CA has a clear opposing stance to the OA. .	3	The CA correctly quotes the OA (note: paraphrasing or summarizing the point is also acceptable, it does not necessarily have to be the exactly same sentence) and correctly includes the main point from the OA.
Attack on the main point (Att)		**New arguments against the topic (Nat)**	
Score	Score Reason	Score	Score Reason
1	CA does not attack the main point made in the OA.	1	i) CA does not state any new arguments or points that are not mentioned in the OA. or ii) CA states new arguments or points that are not mentioned in the OA, but some or almost all of them are irrelevant to the topic, which may or may not cause the whole CA to be off-topic.
2	CA provides arguments that implicitly/vaguely attack the main point made in the OA.	2	Although the CA states new arguments or points that are not mentioned in the OA and are relevant to the topic, those arguments or points support the topic.
3	CA provides arguments that explicitly/directly attack the main point in the OA. Such attacks make the OA incomplete, weak and/or illogical.	3	CA states new arguments and/or points that are not mentioned in the OA and are indeed against the OA.

them for common properties useful for evaluating counter-arguments in response to debate-level original arguments. We found that a decent CA should directly and effectively respond to or attack the points made by the opposing side, in our case, the OA. Such attacks should be supported with convincing evidence clearly explaining why the arguments from the opposing side are weak or illogical. On top of that, a good CA can also state its own arguments showing the reason that it is stronger than the opposing side, thus enhancing the overall persuasiveness.

3.2 Our new dimensions

Definitions Based on our findings in existing debate rubrics, we create four new dimensions, each of which is scored with a 3-scale value, for evaluating counter-arguments in response to debate-level original arguments. The definitions of the dimensions are shown in Table 1.

Considering the characteristic of parliamentary debates where the OA only mentions one central point (supported with several premises), we set the dimensions of *Quotation from the Original Argument* and *Attack on the Main Point* to be focused on **the main point** in the OA. Although we are aware that a good attack in a CA should be supported by decent evidence, we do not consider dimensions related to evidence in this work since they have already been extensively explored in the literature [2].

Scoring criteria We also provide a scoring rubric for each of the 4 dimensions. The scoring rubrics are shown in Table 2. We set the scoring scheme to be 3-scale to make each score as distinct as possible so that it could serve as constructive feedback to some extent, to show specifically how the argumentation could be improved.

4 Crowdsourcing Experiment

We describe our preliminary crowdsourcing experiment for testing the feasibility of annotating our new dimensions.

4.1 Data

In this work, we use *The Debate Dataset*, a dataset of debates created via crowdsourcing [1)] and an extension of the dataset used in TYPIC [12]. When creating this dataset, crowdworkers were given the topic "Students should have part-time jobs", and an original argument supporting the topic. Workers were then instructed to write a point from the original argument and write their counter-argument to rebut the point. The dataset contains 5 original arguments with 446 counter-arguments in total. In the experiments, we utilize all 5 original arguments with 5 associated counter-arguments for each as a preliminary test. We plan to annotate and publish all the data in the future.

4.2 Experiment: Counter-argument Quality Assessment Task

We first collect workers with a basic understanding of our task and grant them the qualification of *Outstanding worker*.[3)] Following the work [8], for each dimension, we show an example counter-argument for every possible score in addition to the scoring rubric in the crowdsourcing in-

3) For more details on our qualification filtering, see Appendix A.

Table 3 Krippendorff's α with interval distance function for each dimension as well as all dimensions combined.

AoS	Quo	Att	Nat	All
0.511	0.369	0.474	0.075	0.393

terface [4]. Workers are instructed to rate a CA based on the scoring rubrics and also write down their specific reasons for the rating for each dimension. Five unique *Outstanding Workers* are asked to annotate one CA at a time. In total, 125 ($= 5 \times 5 \times 5$) annotations are conducted.

4.3 Results and Analysis

There are seven unique workers in total who participated in the experiment. To measure inter-annotator agreement among workers, we calculated Krippendorff's α [13]. As shown in Table 3, we obtained moderate agreements on *AoS*, *Quo*, and *Att*, however, low agreement on *Nat*.

We investigated the reason for the low agreement rate (e.g. *Nat*) by manually checking some of the scoring reasons written by workers. Our main findings include 1) there are still some expressions used in a nebulous way in the rubrics, which potentially brought in subjectivity, such as "vague attack" in *Att* or "new argument" in *Nat*. 2) for *Nat*, the current rubric cannot cover the situation where there are both "new arguments that are for the topic" and "new arguments that are against the topic" mixed in a single essay, which implies that a more fine-grained level annotation of dividing the counter-argument essay into individual arguments and assessing on top of that is needed. 3) workers' understanding of "the main point made in the original argument" also varies, which indicates the difficulty of this task that one must understand the logic of both sides.

5 Discussion

Towards a more fine-grained annotation, where each of the individual arguments (i.e. claim and its supporting premises) could be assessed separately and aggregated to the overall quality of the counter-argument, we first attempted to capture different granularity of attacks.

Given that, ideally, a strong counter-argument could comprehensively attack all the points made in the original argument (including the central point and all the supporting premises), we investigated various ways of capturing the coverage of attacks using crowdsourcing.

Free selection We first allowed workers to freely select all attack pairs (i.e. free text spans) between the OA and CA. Although we were able to collect reasonable attack pairs, we found that the results largely varied among workers, which made it difficult to calculate the agreement.

Sentence-level selection We attempted to solve the issue by having workers select pre-separated sentences, instead of free spans, from both the OA and CA. We still found difficulties for the annotators, potentially due to the fact that the understanding of the two essays also varies.

Logic-graph Representation To assist workers in better understanding the OA, we tried explicating the logic of the OA by representing it as a logic graph. Each argument, including the central point and all the premises that support the point, is broken down into nodes and relations. Each node is a concept, and the relations mainly include causal relations (*promote* and *suppress* [14]). We had workers select one sentence from the CA and all the nodes or relations attacked by the sentence in the logic graph. We found that although it is possible to collect attack pairs in this way, it is required to know the argumentative structure within the CA in advance when it comes to evaluating the attacks since there might be associations between sentences in different attack pairs (e.g. one sentence might serve as evidence for another in the CA, thus we cannot simply ask the question "Is this attack supported by evidence" for every attack pair).

6 Conclusion and Future Work

In this work, we tackled the challenging task of assessing counter-arguments in debates. We surveyed several debate rubrics and created four new analytic dimensions for evaluating the quality of the counter-argument in relation to the original argument. We conducted a crowdsourcing experiment and found that while the workers understand the dimensions, there are still many challenges to achieving a good agreement among workers.

In our future work, we will refine the scoring rubrics based on workers' scoring reasons and have workers highlight the text segments along with their scores instead of writing the reasons in free text. We will also expand the annotation to the whole corpus. Moreover, we will explore more fine-grained annotations where we could capture the diversity of attacks and also incorporate dimensions related to evidence assessment, based on previous works, into the evaluation procedure of counter-arguments.

4) See Appendix B for more information about our interface.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22H00524.

References

[1] Kate Shuster and John Meany. Introducing parliamentary debate: A resource for teachers and students. **Claremont Colleges Debate Outreach**, January 2016.

[2] Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In **Proceedings of the 56th Annual Meeting of the ACL (Volume 1: Long Papers)**, pp. 621–631, Melbourne, Australia, July 2018. ACL.

[3] Martin Hinton and Jean H M Wagemans. Evaluating reasoning in natural arguments: A procedural approach. **Argumentation**, Vol. 36, No. 1, pp. 61–84, March 2022.

[4] Thiemo Wambsganss and Christina Niklaus. Modeling persuasive discourse to adaptively support students' argumentative writing. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8748–8760, Dublin, Ireland, May 2022. ACL.

[5] Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. AL: An adaptive learning support system for argumentation skills. In **Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems**, CHI '20, pp. 1–14, New York, NY, USA, April 2020. ACM.

[6] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. Aspect-controlled neural argument generation. In **Proceedings of the 2021 Conference of the North American Chapter of the ACL: Human Language Technologies**, pp. 380–396, Online, June 2021. ACL.

[7] Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. Employing argumentation knowledge graphs for neural argument generation. In **Proceedings of the 59th Annual Meeting of the ACL and the 11th IJCNLP (Volume 1: Long Papers)**, pp. 4744–4754, Online, August 2021. ACL.

[8] Xinyu Hua, Zhe Hu, and Lu Wang. Argument generation with retrieval, planning, and realization. June 2019.

[9] Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. Argument undermining: Counter-Argument generation by attacking weak premises. May 2021.

[10] Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. The workweek is the best time to start a family – a study of GPT-2 based claim generation. In **Findings of the ACL: EMNLP 2020**, pp. 528–544, Online, November 2020. ACL.

[11] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In **Proceedings of the 15th Conference of the European Chapter of the ACL: Volume 1, Long Papers**, Stroudsburg, PA, USA, 2017. ACL.

[12] Shoichi Naito, Shintaro Sawada, Chihiro Nakagawa, Naoya Inoue, Kenshi Yamaguchi, Iori Shimizu, Farjana Sultana Mim, Keshav Singh, and Kentaro Inui. TYPIC: A corpus of Template-Based diagnostic comments on argumentation. In **Proceedings of the Thirteenth LREC**, pp. 5916–5928, Marseille, France, June 2022. European Language Resources Association.

[13] K. Krippendorff. Content analysis: An introduction to methodology. Beverly Hills, CA., 1980. Sage Publications, Inc.

[14] Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. Feasible annotation scheme for capturing policy argument reasoning using argument templates. In **Proceedings of the 5th Workshop on Argument Mining**, pp. 79–89, Brussels, Belgium, November 2018. ACL.

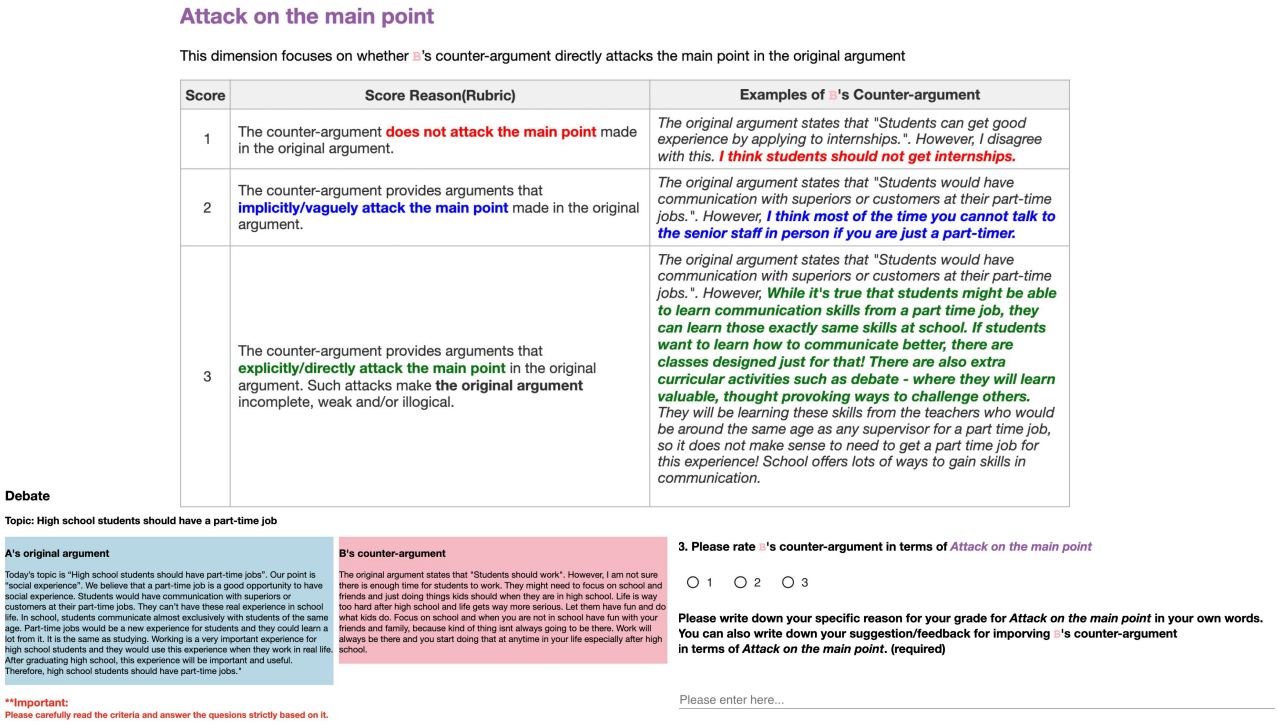

Figure 2 The crowdsourcing interface. For space reasons, we only show one dimension.

A Competent Workers Selection

Given the difficulty of collecting high-level annotations via crowdsourcing, we first create a filtering procedure to collect competent workers for our task through a qualification test and survey. The qualification test contains three debates, each of which is a pair of an original argument and a counter-argument. For each debate, there is one comprehension question asking "Which point from the original argument is the counter-argument attacking?" with four candidate options. Only workers who answered all three questions correctly can get access to the survey and additional comprehension questions. The survey's goal was to learn more about our workers, such as their native language and level of expertise with argumentation. The purpose of the additional comprehension questions was to filter out workers that provided generic responses or exhibited a low level of fluency. Specifically, the workers were given a debate and asked to score the quality of the counter-argument while providing reasons. The reasons were judged by two expert annotators, both authors of this paper, and incompetent workers were filtered out.

B Crowdsourcing Interface

An example of our final crowdsourcing interface is shown in Figure 2. Due to space reasons, we only show one of the dimensions (Attack on the main point) shown in our guidelines, an example debate, as well as the question for the dimension. [5)] Workers were first asked to rate on a scale of 1-3 based on the scoring rubric, and provide a specific reason for their score. We plan to utilize such reasons to further improve our crowdsourcing task in future trials.

5) The complete version of the interface can be found at https://github.com/oubunshitsu/debate_CA_assessment.