DISAPERE: A Dataset for Discourse Structure in Peer Review Discussions

Anonymous ACL submission

Abstract

At the foundation of scientific evaluation is the labor-intensive process of peer review. This critical task requires participants to consume vast amounts of highly technical text. Prior work has annotated different aspects of review argumentation, but discourse relations between reviews and rebuttals have yet to be examined.

We present DISAPERE, a labeled dataset of 20k sentences contained in 506 review-rebuttal pairs in English, annotated by experts. DISAPERE synthesizes label sets from prior work and extends them to include fine-grained annotation of the rebuttal sentences, characterizing their context in the review and the authors' stance towards review arguments. Further, we annotate *every* review and rebuttal sentence.

We show that discourse cues from rebuttals can shed light on the quality and interpretation of reviews. Further, an understanding of the argumentative strategies employed by the reviewers and authors provides useful signal for area chairs and other decision makers.

1 Introduction

Peer review performs the essential role of quality control in the dissemination of scientific knowledge. The rapid increase in academic output places an immense burden on decision makers such as area chairs and editors, as their decisions must take into account not only extensive manuscripts, but enormous additional amounts of technical text including reviews, rebuttals, and other discussions.

One long term goal of research in peer review is to support decision makers in managing their workload by providing tools to help them efficiently absorb the discussions they must read. While machine learning should not be used to produce condensed accounts of the peer review text due to the risk of amplifying biases (Zhao et al., 2017), ML tools could nevertheless help decision makers manage their information overload by identifying patterns in the data, such as argumentative strategies, goals, and intentions.

Any such research requires an extensive labeled dataset. While the OpenReview platform (Soergel et al., 2013) has made it easy to obtain unlabeled public peer review text, labeling this data for supervised NLP requires highly qualified annotators. Correct interpretation of the discourse structure of the text requires the understanding of the technical content, precluding the use of standard crowdsourcing techniques. Prior work on discourse in peer review has made the most of this qualified labor force by focusing its work on labeling arguments extracted from the text. This enables the complete annotation of more examples, with the drawback of precluding research on non-argumentative behaviors in peer review. While there has been extensive research and deep analysis of different aspects of peer review, the taxonomies used to describe review argumentation are disparate and not directly compatible. Finally, there has been limited research into understanding the discourse relations between rebuttals and reviews (Cheng et al., 2020; Bao et al., 2021), and none so far into the discourse structure of the rebuttal itself.

This paper presents **DISAPERE** (**DI**scourse **S**tructure in **A**cademic **PE**er **RE**view), a dataset focusing on the interaction between reviewer and author, and thus gives reviews and rebuttals equal importance, and emphasizing the relations between them. To enable the study of behaviors beyond the core arguments, we also annotate every sentence of both the review and rebuttal, and provide fine-grained labels for non-argumentative types. We annotate at the sentence level not only for completeness but also to avoid the propagation of errors from argument detection. We annotate four properties (REVIEW-ACTION, FINE-REVIEW-ACTION, ASPECT, POLARITY) of each review sentence, where the set of properties and their values was developed by synthesizing taxonomies from

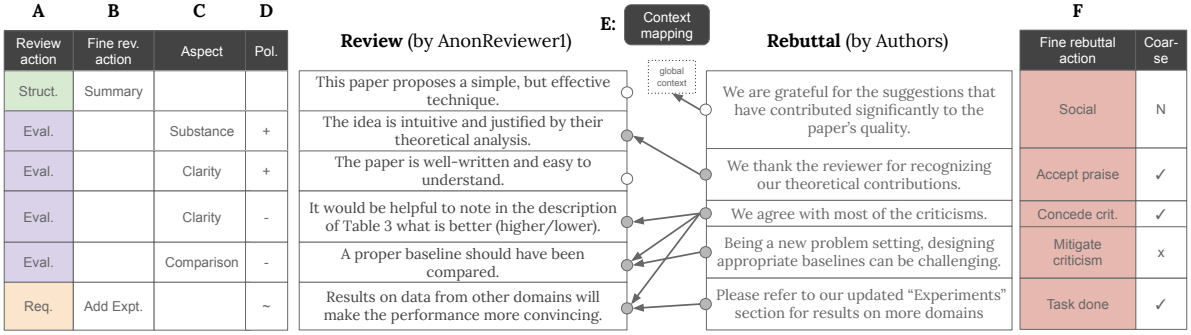

Figure 1: A depiction of our annotation scheme on a minimal, fictional review-rebuttal pair. A: REVIEW-ACTION , including Structuring, Request, Evaluation; B: FINE-REVIEW-ACTION , fine-grained categorization of Structuring and Request sentences; C: ASPECT , indicating the qualities of the manuscript being commented upon D: POLARITY indicating whether these comments are positive or negative in nature. E: Each sentence in the rebuttal is mapped to zero or more sentences in the review, which constitute its context. This is a many-to-many relation. F: The sentences in the rebuttal are labeled with domain-specific discourse acts (REBUTTAL-ACTION); each discourse act may be categorized according to whether it concurs with (✓) or disputes (×) the premise of the context it is responding to.

prior work. We also annotate each sentence of a rebuttal with a fine-grained label indicating the author's intentions and commitment, and a link to the set of review sentences that form its context. Figure 1 shows the DISAPERE annotation scheme on a minimal, fictional example review-rebuttal pair.

DISAPERE is intended as a comprehensive and high-quality test collection, along with training data to fine-tune models. Our annotations are carried out by graduate students in computer science who have undergone training and calibration, amounting to over 850 person-hours of annotation work. Much of the test data is double-annotated, and we report inter-annotator agreement on all aspects of the annotation. We describe the performance of state of the art models on the tasks of predicting labels and contexts, showing that interesting ambiguities in the data provide the NLP community with research challenges. We also show an example preliminary analysis indicating various facets that program chairs should consider when developing policies for conferences or journals (§ 5).

The contributions of this paper are as follows: (1) a new labeled training dataset of 506 review-rebuttal pairs (over 20k sentences) of peer review discussion text in English, where review sentences are annotated with four properties, and rebuttal sentences are annotated with context and labels from a novel scheme to describe discourse structure; (2) a taxonomy of discourse labels synthesizing prior work on discourse in peer review and extending it to add useful subcategories; (3) a summary of the performance of baseline models on this dataset (§ 6); (4) examples of analyses on this dataset that could benefit decision makers in peer review (§ 4), and (5) annotation software and extensive annotation guidelines to support future labeling efforts in this area.

2 Related work

The design of this dataset draws upon extensive, but disparate prior work on this topic. Many works, some addressed below, have taken advantage of the availability of review text hosted on OpenReview.

Argument-level review labeling Prior work has developed label sets that address different phenomena. Hua et al. (2019) introduced the study discourse structure in peer review by annotating argumentative propositions in the AMPERE dataset with a set of labels tailored to the peer review domain (EVALUATION, REQUEST, FACT, REFERENCE, and QUOTE). Similarly, Fromm et al. (2020) describes the AMSR dataset, which frames the problem as an argumentation process, in which the stance of each argument towards the paper's acceptance or rejection is of paramount importance. Each of views peer review as an argumentation, using argument mining techniques to highlight spans of interest.

While its goal is not to examine discourse structure per se, Yuan et al. (2021) uses polarity labels to indicate each argument's support or attack of the authors' bid for acceptance. Besides polarity, these examples follow Chakraborty et al. (2020)

by annotating each argument with the *aspect* of the paper to which the comment is directed.[1] In contrast to Yuan et al. (2021), we do not attempt or recommend generating peer review text, instead focusing on analyzing peer review text generated by humans.

Review-rebuttal interactions We also expand on work by Cheng et al. (2020), who first annotated discourse relations between sentences in reviews and rebuttals. While Cheng et al. (2020) focus on new deep learning architectures, in this paper we focus on the creation and comprehensive annotation of a new dataset, illustrated with results from some less specialized baseline models.

Other research into rebuttals includes Gao et al. (2019). Besides their main finding, that reviewers rarely change their rating in response to rebuttals, they find that more specific, convincing and explicit responses are more likely to elicit a score change. Observations from this paper are formalized into rebuttal action labels in DISAPERE.

Comparison of datasets In DISAPERE we attempted to unify these schemas to form a single hierarchical schema for review discourse structure. We then expanded this hierarchical schema to introduce fine-grained classes for implicit and explicit requests made by the reviewers. The details of the correspondence between DISAPERE labels and those from prior work are summarized in Appendix A. In contrast to prior work, DISAPERE labels discourse phenomena at a sentence level rather than an argument level. This enables more thorough coverage of the text while avoiding the propagation of errors from machine learning models earlier in the annotation pipeline. Other differences with prior work are summarized in Table 1.

	Dataset	AMPERE	AMSR	ASAP-Review	APE	**DISAPERE**
	# examples	400	77	1k	4.7k	506
	# labels	10k	1.4k	5.7k	130k	46k
Review	Arg. stmts.	✓	✓		✓	✓
Review	Arg. types	✓				✓
Review	Polarity		✓	✓		✓
Review	Aspect			✓		✓
Review	Non-arg.					✓
Review	All sents.					✓
Rebuttal	Included?				✓	✓
Rebuttal	Arg. stmts.				✓	✓
Rebuttal	Context				✓	✓
Rebuttal	Arg. types					✓
Rebuttal	Non-arg.					✓
Rebuttal	All sents.					✓

Table 1: Comparison between our dataset and prior work: AMPERE (Hua et al., 2019), AMSR (Fromm et al., 2020), ASAP-Review (Yuan et al., 2021), APE (Cheng et al., 2020). *Arg.stmts.*: are argumentative statements highlighted?; *Arg. types*: Are subtypes of argumentative statements labeled?; *Non-arg*: Are non-argumentative statements labeled?; *All sents.*: Are labels provided for all sentences?; *Context*: Are rebuttal texts' contexts in the review provided? DISAPERE is the only work to annotate every sentence in the review and rebuttal, and the only work that applies discourse labels to the author's actions in the rebuttal.

3 Dataset

Each example in DISAPERE consists of a pair of texts: a review and a rebuttal. Labels for reviews and rebuttal sentences are described below. Review sentence labels are summarized in Table 2, and rebuttal sentence labels in Table 3.

3.1 Review sentence labels

3.1.1 Review actions

REVIEW-ACTION annotations characterize a sentence's intended function in the review. Annotators label each sentence with one of six coarse-grained sentence types including *evaluative* and *fact* sentences, *request* sentences (including questions, which are requests for information), as well as non-argument types: *social*, and *structuring* for organization of the text.

3.1.2 Fine-grained review actions

We also extend two of these review actions with subtypes: *structuring* sentences are distinguished between headers, quotations, or summarization sentences, and *request* sentences are subdivided by the nature of the request, distinguishing between clarification of factual information, requests for new experiments, requests for an explanation (e.g. of motivations or claims), requests for edits, and identification of minor typos.

[1] Aspects are based on the ACL 2018 rubric.

Category		Label	Description	Percentage
Review action		Evaluative	A subjective judgement of an Aspect of the paper	35.14%
		Structuring	Text used to organize an argument	29.65%
		Request	A request for information or change in regards to the paper	21.20%
		Fact	An objective truth, typically used to support a claim	9.15%
		Social	Non-substantive text typically governed by social conventions	1.51%
Aspect		Substance	Are there substantial experiments and/or detailed analyses?	18.30%
		Clarity	Is the paper clear, well-written and well-structured?	11.86%
		Soundness/Correctness	Is the approach sound? Are the claims supported?	10.25%
		Originality	Are there new topics, technique, methodology, or insights?	4.12%
		Motivation/Impact	Does the paper address an important problem?	3.95%
		Meaningful Comparison	Are the comparisons to prior work sufficient and fair?	3.37%
		Replicability	Is it easy to reproduce and verify the correctness of the results?	3.06%
Polarity		Negative	Negatively describes an aspect of the paper (reason to reject)	31.49%
		Neutral	Does not commit to being positive or negative	12.68%
		Positive	Positively describes an aspect of the paper (reason to accept)	11.95%
Fine review action	Struct.	Summary	Reviewer's summary of the manuscript	19.45%
		Heading	Text used to organize sections of the review	9.14%
		Quote	A quote from the manuscript text	1.07%
	Request	Explanation	Request to explain of scientific choices (question)	5.89%
		Experiment	Request for additional experiments or results	5.11%
		Edit	Request to edit the text in the manuscript	4.43%
		Clarification	Request to clarify of the meaning of some text (question)	2.99%
		Typo	Request to fix a typo in the manuscript	2.11%

Table 2: A list of the review sentence labels, their descriptions, and the percentage of review sentences they apply to. Note that the percentages should not add up to 100%, as a single sentence may have labels from multiple categories.

3.1.3 Aspect and polarity

ASPECT annotations follow the ACL review form (Chakraborty et al., 2020; Yuan et al., 2021). These distinguish *clarity*, *originality*, *soundness/correctness*, *replicability*, *substance*, *impact/motivation*, and *meaningful comparison*. Following Yuan et al. (2021), arguments with an ASPECT are also annotated for POLARITY. We include *positive*, *negative*, and *neutral* polarities. ASPECT and POLARITY are applied to sentences whose REVIEW-ACTION value is *evaluative* or *request*.

3.2 Rebuttal sentence labels

We annotate two properties of each rebuttal sentence: a REBUTTAL-ACTION label characterizing its intent, and its context in the review in the form of a subset of review sentences.

3.2.1 Rebuttal actions

The 14 rebuttal actions (Table 3) are divided into three REBUTTAL-STANCE categories (*concur*, *dispute*, *non-arg*) based on the author's stance towards the reviewer's comments.

(1) *concur*: The author concurs with the premise of the context. This includes answering a question or discussing a requested change that has been made to the manuscript, conceding a criticism in an evaluative sentence. (2) *dispute*: The author disputes the premise of the context. The rebuttal sentence may reject a criticism or request, disagree with an underlying fact or assertion, or mitigate criticism (accepting a criticism while, e.g., arguing it to be offset by other properties). (3) *non-arg*: Encompasses rebuttal actions including *social* actions (such as thanking reviewers), *structuring* labels, for sentences that organize the review.

Responses to *request*s are further annotated: if the author *concur*s, we record whether the task has been completed by the time of the rebuttal, or promised by the camera ready deadline; if the author *dispute*s, we record whether the task was deemed to be out of scope for the manuscript.

3.2.2 Rebuttal context

We refer to the set of sentences – if any – which a rebuttal sentence is responding to as the *context* of that rebuttal. The set of sentences may be the entire review (*global context*) or the empty set (*no context*). By not mandating a fixed discourse chunking (e.g. Cheng et al. (2020)), these

Category		Label	Description	Reply to	Percentage
Argumentative	Concur	Answer	Answer a question	Request	33.14%
		Task has been done	Claim that a requested task has been completed	Request	8.68%
		Concede criticism	Concede the validity of a negative eval. statement	Evaluative	2.73%
		Task will be done	Change promised by camera ready deadline	Request	2.03%
		Accept for future work	Express approval for a suggestion, but for future work	Request	1.31%
		Accept praise	Thank reviewer for positive statements	Evaluative	0.36%
	Dispute	Reject criticism	Reject the validity of a negative eval. statement	Evaluative	10.49%
		Mitigate criticism	Mitigate the importance of a negative eval. statement	Evaluative	2.46%
		Refute question	Reject the validity of a question	Request	0.96%
		Contradict assertion	Contradict a statement presented as a fact	Fact	0.87%
Non-arg		Structuring	Text used to organize sections of the review	-	18.02%
		Summary	Summary of the rebuttal text	-	8.04%
		Social	Non-substantive social text	-	6.79%
		Followup question	Clarification question addressed to the reviewer	-	0.32%

Table 3: A list of the rebuttal sentence labels, their descriptions, and the percentage of rebuttal sentences they apply to. The "Reply to" column shows the review action types that a particular rebuttal type would canonically reply to.

labels may handle complex mappings between hierarchies, where some rebuttals refer to larger (possibly non-contiguous) chunks of text and others refer to specific individual sentences.

3.3 Data Sampling and Annotation

DISAPERE uses English text from scientific discussions on OpenReview Soergel et al. (2013). We draw review-rebuttal pairs from the International Conference on Learning Representations (ICLR) in 2019 and 2020. Review-rebuttal pairs are split into train, development and test sets in a 3:1:2 ratio such that all texts associated with any manuscript occur in the same subset. Overall statistics for the dataset are summarized in Table 4.

	Train	Dev	Test
Num. review-rebuttal pairs	251	88	167
Num. manuscripts	94	37	57
Num. adjudicated pairs	0	0	65
Num. review sentences	4846	1358	3087
Num. rebuttal sentences	5805	2015	3283
Avg. sents per review	19.31	15.43	18.49
Avg. sents per rebuttal	23.13	22.9	19.66

Table 4: Statistics for the dataset. Where possible, multiple reviews for the same manuscript were annotated. All reviews for any particular manuscript falll within the same train/dev/test split. Adjudicated pairs are those that were annotated by multiple annotators, and had disagreements resolved by an experienced annotator. All test set pairs are double-annotated.

Each ICLR manuscript is reviewed by three or more reviewers. Authors are able to respond to each review by adding a comment. Although rebuttals are not formally named, we consider direct replies by the author to the initial review comment to constitute a rebuttal. Reviewers and authors may engage in multi-turn interactions beyond this initial response, but we focus on the set of reviews and initial responses, and leave study of extended discussion for future work. The text is separated into sentences using the spaCy (Honnibal et al., 2020) sentence separator.

Annotation was accomplished with a custom annotation tool designed for this task (described in detail in Appendix B), which will be released upon publication. Annotators annotate each sentence of a review, then examine the rebuttal sentences in order, selecting sets of review sentences to form their context. Annotators were permitted to directly propagate context selections to subsequent rebuttal sentences when necessary. While the current annotation focuses on links between sentences, this is not intended to eschew discourse structure. Indeed, in contrast to pipelined approaches where discourse chunks are aligned Cheng et al. (2020), this annotation can describe both single and multiple (not necessarily contiguous) sentence contexts. However, we admit that any resultant clustering of review sentences is a latent structure implied by the context mapping, rather than an explicitly annotated discourse chunking or discourse tree.

3.4 Agreement

We report Cohen's κ (Cohen, 1960) on the IAA of labeling both review and rebuttals, treating each

Label	κ
REVIEW-ACTION	0.605
FINE-REVIEW-ACTION	0.583
ASPECT	0.447
POLARITY	0.561
REBUTTAL-STANCE	0.513
REBUTTAL-ACTION	0.479

Table 5: IAA for review labels (top) and rebuttals (bottom), scored on double annotation. IAA is reported on 65 double-annotated examples, all of which fall in the test set of DISAPERE.

sentence as a labeling unit (Table 5). It shows substantial chance-corrected agreement between annotators, both when using REBUTTAL-ACTION and REBUTTAL-STANCE labels. Details about the overlap of context sentence sets are provided in app:overlap.

4 Analysis

4.1 Context types

We separate the different types of rebuttal contexts in terms of the number and relative position of selected review sentences in Table 6, along with the four cases in which the context cannot be described as a subset of review sentences.

	Context type	Num. reb. sents	% reb. sents
Sents. selected	Multiple contiguous	12375	60.31%
	Single sentence	4313	21.02%
	Mult. non-contiguous	2144	10.45%
No sents. selected	Global context	816	3.98%
	Context in rebuttal	647	3.15%
	No context	112	0.55%
	Context error	101	0.49%
	Cannot be determined	11	0.05%

Table 6: Different types of rebuttal sentence contexts. Top: Over 90% of sentences are linked to a subset of sentences in the review. The remaining sentences are subdivided into four categories. Bottom: sentences not linked to any particular subset of review sentences.

4.2 Alignment

Since authors often respond to reviewers' points in order, one would expect alignment between rebuttal and review sentences to be trivial, yet this is not necessarily the case. In Figure 2, we calculate Spearman's ρ between rebuttal sentence indices and their aligned review sentence indices. Rebuttals responding to each point in order would achieve a ρ of 1.0. However, this is rare – instead, we find that ρ is even negative in some cases. This indicates that while the task of determining rebuttal sentences' contexts may benefit from linear inductive bias, the task is not trivial.

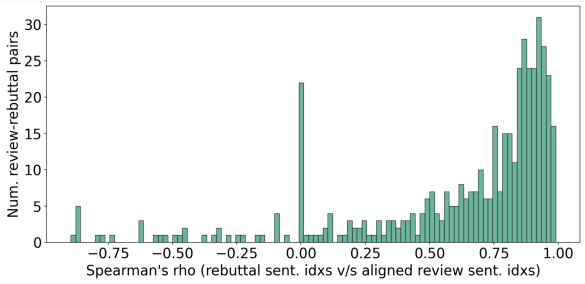

Figure 2: Spearman's ρ between rebuttal sentence indices and aligned review sentence indices.

4.3 Author interpretations of criticism

The REBUTTAL-ACTION labels were designed in the context of REVIEW-ACTION labels, and thus tend to explicitly refer to different responses to criticism (accept, reject, mitigate) or to requests (answer, refute, etc.) However, annotations revealed that authors often interpret a review sentence in ways that support their argumentative goals rather than a general reader's interpretation, such as treating a negative evaluative statement as a request for an improvement. Figure 3 shows the distribution of contexts for three different rebuttal actions.

5 Applications

5.1 Ethics

The outcomes of peer review can have outsize effects on the careers of participating scholars. As machine learning models are known to amplify biases, we strongly recommend against using the outputs of any machine learning system to make decisions about individual cases. A dataset like DISAPERE is best used to survey participants' behavior. Any interventions based on this information should be subjected to studies in order to ensure that they do not introduce or exacerbate bias.

5.2 Agreeability

Gao et al. (2019) showed that reviewers do not appear to act upon the rebuttals responding their reviews. It is possible that this is due to paucity of

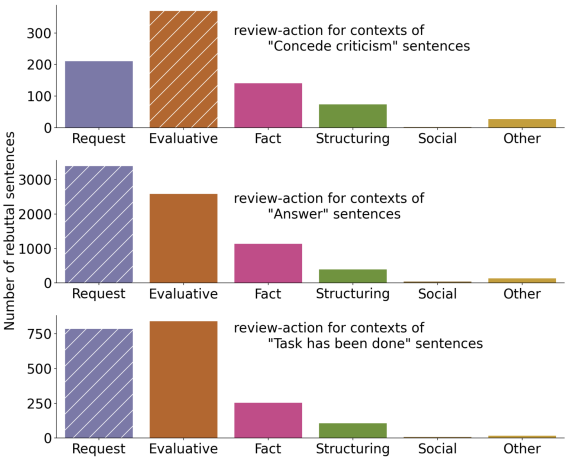

Figure 3: Distribution over REVIEW-ACTION for the context sentences of three rebuttal actions. The canonical REVIEW-ACTION is marked by cross hatching. Note that authors sometimes interpret requests as criticisms ("Concede criticism"); often respond to evaluative sentences as if they are questions ("Answer"), and sometimes treat criticisms in the form of evaluative sentences as requests which they then carry out. ("Task has been done")

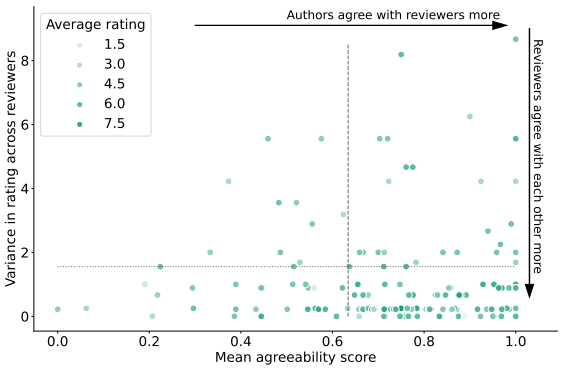

Figure 4: Mean agreeability for a manuscript's reviews v/s reviewer variance. Manuscripts above the dotted line are in the top quartile of rating variance, and are more likely to be reviewed by area chairs. Manuscripts to the left of the dashed line are in the bottom quartile of mean agreeability, in which authors take issue with the premises of reviewers' comments. The color of the dots indicates the mean of the ratings awarded by the three reviewers; the ratings come from the set $\{1 \cdots 9\}$

time on the reviewers' part. It is also common practice for area chairs to use review variance across a manuscript's reviews as a practical heuristic to decide which manuscripts need their attention. We propose that discourse information such as that described by DISAPERE can be used to provide heuristics that are data-driven, yet interpretable, and leverage information from the content of reviews rather than just numerical scores, resulting in better decision making.

One such measure is *agreeability*, which we define as the ratio of CONCUR sentences to argumentative sentences, i.e.: $agreeability = \frac{n_{concur}}{n_{concur}+n_{dispute}}$. We argue that low agreeability can indicate problematic reviews even in cases where the variance in scores does not reveal an issue, as illustrated in Figure 4. Agreeability is only weakly correlated with rating, with Pearson's $r = 0.347$. in Figure 4, 18% (28/159) of manuscripts would not meet the bar for high variance scores (top quartile), although their low agreeability (bottom quartile) indicates that area chairs may want to pay closer attention.

6 Baselines

Two types of machine learning tasks can be defined in DISAPERE. First, a sentence-level classification task for each of the four review labels and the two levels of rebuttal labels. Second, an alignment task in which, given a rebuttal sentence, the set of review sentences that form its context are to be predicted.

The models described below are not intended to introduce innovations in discourse modeling, rather, we intend to show the off-the-shelf performance of state-of-the-art models, and indicate through error analysis the phenomena that are yet to be captured.

6.1 Sentence classification

We report results on two models for classification. First, we use `bert-base` (Devlin et al., 2019) to produce sentence embeddings for each sentence, then classify the representation of the `[CLS]` token using a feedforward network. Second, four labels (REVIEW-ACTION, FINE-REVIEW-ACTION, ASPECT, POLARITY) are compulsorily applied to every sentence. Thus, these four classification tasks are instances of *sequential sentence classification* (SSC), a task introduced by Cohan et al. (2019). We report results using their state-of-the-art SSC model, in which windows over sentences with special separator tokens are passed through BERT, and a feedforward network acts as a classifier on the separator token output representations.

We report F1 scores for both models. The results of these models are shown in Table 7. In general, F1 is lower for tasks with larger label spaces. While the performance is reasonable in most cases, there is still room for improvement. ASPECT achieves a particularly low F1 score, but its κ is within the bounds of moderate agreement, this must be ac-

counted for by the inherent difficulty of the task rather than a deficit in data quality.

Label	Avg F1 (test)	κ	Num. labels
REVIEW-ACTION	62.28%	0.605	6
FINE-REVIEW-ACTION	44.87%	0.583	10
ASPECT	32.82%	0.447	9
POLARITY	63.37%	0.561	4
REBUTTAL-STANCE	44.77%	0.513	4
REBUTTAL-ACTION	31.54%	0.479	17

Table 7: Sequential sentence classification results. Top: review labels; Bottom: rebuttal labels.

We examine the results for REVIEW-ACTION, REBUTTAL-ACTION, and ASPECT.

6.2 Rebuttal context alignment

We model rebuttal context alignment as a ranking task. Ideally, a model should rank all relevant review sentences higher than non-relevant review sentences. As a baseline, we use an information retrieval (IR) model based on BM25 that, given a rebuttal sentence ranks all the corresponding review sentences. We also report results from a neural sentence alignment model based on a two-tower Siamese-BERT (S-BERT) model (Reimers and Gurevych, 2019). Each review and rebuttal sentence are encoded independently using a S-BERT encoder and the similarity between two sentences is computed using cosine similarity. We initialize with a model[2] pre-trained on various sentence-pair datasets. The alignment models are evaluated using mean reciprocal rank (MRR). We add a `NO_MATCH` sentence to the review, to which rebuttal sentences without context sets in the review are aligned.

The neural model outperforms the IR baseline indicating that simple lexical matching models are not enough to achieve good performance on this task. However, both the neural and IR model achieve a relatively low MRR, indicating that there is significant scope for improvement; in particular that *similarity* as encoded by lexical feature or cosine similarity in large language models' latent space is not a good proxy for *relatedness* in the sense encoded in the rebuttal context annotations.

[2]We initialize from a sentence-transformers/all-MiniLM-L6-v2 model

Model	Test MRR
BM-25	0.3544
S-BERT	0.4352

Table 8: Rebuttal Context-alignment results (mean reciprocal rank).

The neural model predicts a contiguous set of sentences as its top choices only 8.36% of the time, although the prevalence of contiguous sentence contexts is 60% (Table 6). This indicates the need for inductive bias in models for this task, and the shortcomings of modeling each context decision independently. Further, the model predicts no context 4.7% of the time, which underestimates the prevalence of such sentences. This may indicate that distinguishing between sentences that are related and sentences that are merely similar remains a challenging task.

7 Conclusion

As the burden of academic peer reviewing grows, it is important for program chairs and editors to act upon data-driven insights rather than heuristics, to make the best possible use of participants' scarce time. Models trained on DISAPERE will allow decision makers to glean deep insights on the interactions occurring during peer review. The detailed annotation guidelines and software provided with this paper support seamless data collection, allowing users to build on DISAPERE, and ensuring that their insights are robust to changes in trends over time and across fields.

References

Jianzhu Bao, Bin Liang, Jingyi Sun, Yice Zhang, Min Yang, and Ruifeng Xu. 2021. Argument pair extraction with mutual guidance and inter-sentence relation graph. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3923–3934, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Souvic Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2020. Aspect-based sentiment analysis of scientific reviews. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 7000–7011, Online. Association for Computational Linguistics.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. 2020. Argument mining driven analysis of peer-reviews. *arXiv preprint arXiv:2012.07743*.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major NLP conference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1274–1290, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

David Soergel, Adam Saunders, and Andrew McCallum. 2013. Open scholarship and peer review: a time for experimentation.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing?

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

A Rationale for taxonomy construction

Our label sets leverage ideas from and commonalities between existing work in this domain, including AMPERE (Hua et al., 2019), AMSR (Fromm et al., 2020) ASAP-Review (Yuan et al., 2021), and Gao et al. (2019):

- ASAP-Review's polarity labels approximately correspond to *arg-pos* and *arg-neg* labels in AMSR
- AMSR and AMPERE each label non-argumentative sentences in a similar manner
- *aspect* labels from ASAP-Review apply only to certain types of sentences; namely *request* and *evaluative* sentences from AMPERE's taxonomy.
- *summary* is an exception among ASAP-Review's *aspect*s, behaving similarly to AMPERE's *quote*. We thus include both of these under a *structuring* category.
- Further, in order to gauge the extent to which authors acquiesced to reviewers' requests, we introduce a fine-grained categorization of the types of requests.
- Gao et al. (2019) enumerates some features of rebuttals, including expressing gratitude, promising revisions, and disagreeing with criticisms. We formalize these observations into our rebuttal label taxonomy.

B Annotation tool

Two modes of annotation are possible. First, annotators can apply labels on a sentence-by-sentence basis. Multiple labeling schemas can be annotated simulatenously, with the option of adding constraints so that certain values govern possible values for other properties. This annotation mode is shown in Figure 5.

The second annotation mode can build on the output of the first annotation mode. Here, sentences of a focus text (the rebuttal) are presented in sequence, and annotators are permitted to select one or more of the sentences in the reference text (the review) which form the context of the sentence of the focus text. Further, a label can be applied to the alignment. This annotation mode is shown in Figure 6 and Figure 7.

C Annotated review-rebuttal pair

Figure 8 shows a truncated version of a review-rebuttal pair from the train set of DISAPERE.

D Context overlap analysis

As a proxy for agreement of rebuttal spans, we show the types of overlap between spans selected by two annotators in 9.

Type of context overlap	Num. rebuttal sentences	% rebuttal sentences
Exact match	1995	49.71%
Agree none	474	11.81%
Partial match	1098	27.36%
Disagree none	215	5.36%
No overlap	231	5.76%

Table 9: Types of context overlap. Full agreement is achieved in the top rows (exact match and 'Agree none', where both annotators agree that there is no appropriate subset of review sentences forming the context. in 'Disagree none', one annotator marks a subset of review sentences, while the other does not.

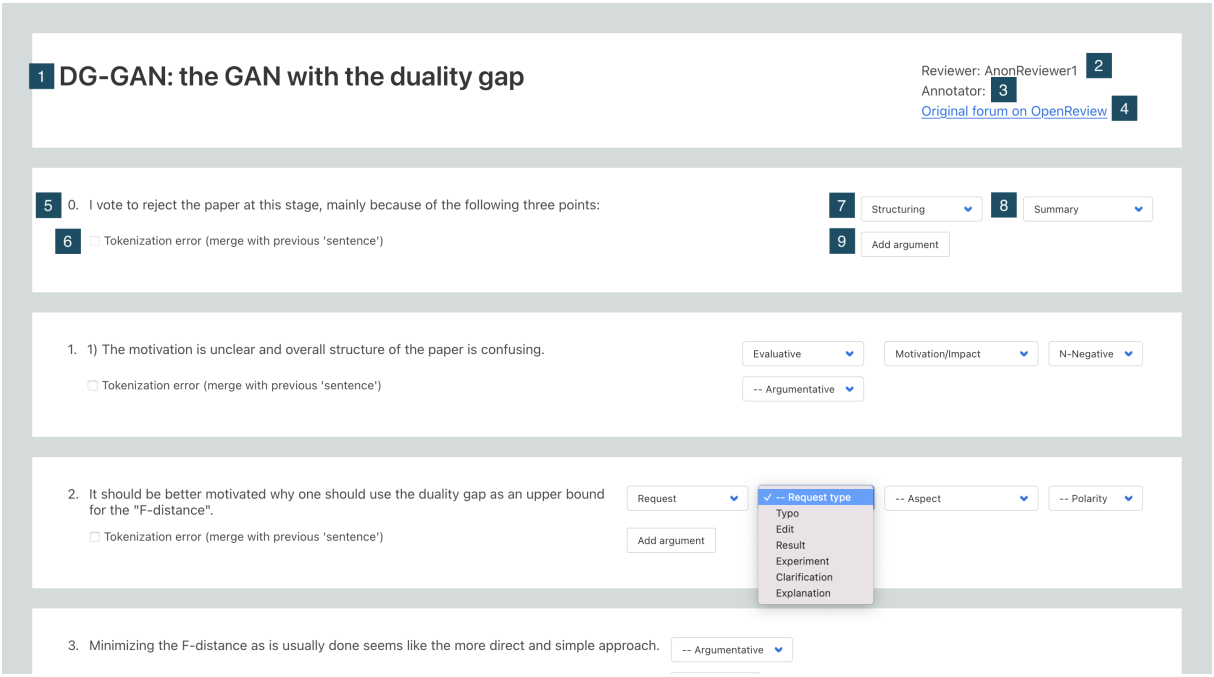

Figure 5: Review annotation interface. Annotators select label values from dropdown menus for each review sentence in turn. (1) Title of the manuscript whose review is being annotated (2) Reviewer identifier (3) Annotator identifier (removed for anonymity) (4) Link to original forum, in case it is needed for context (5) Individual review sentence (6) Option to report sentence splitting error (sentence splitting generally suffered more from precision than recall) (7) Dropdown for REVIEW-ACTION (8) Follow-up dropdownfor FINE-REVIEW-ACTION based on value in 7 (9) Button to add second argument (this was seldom used)

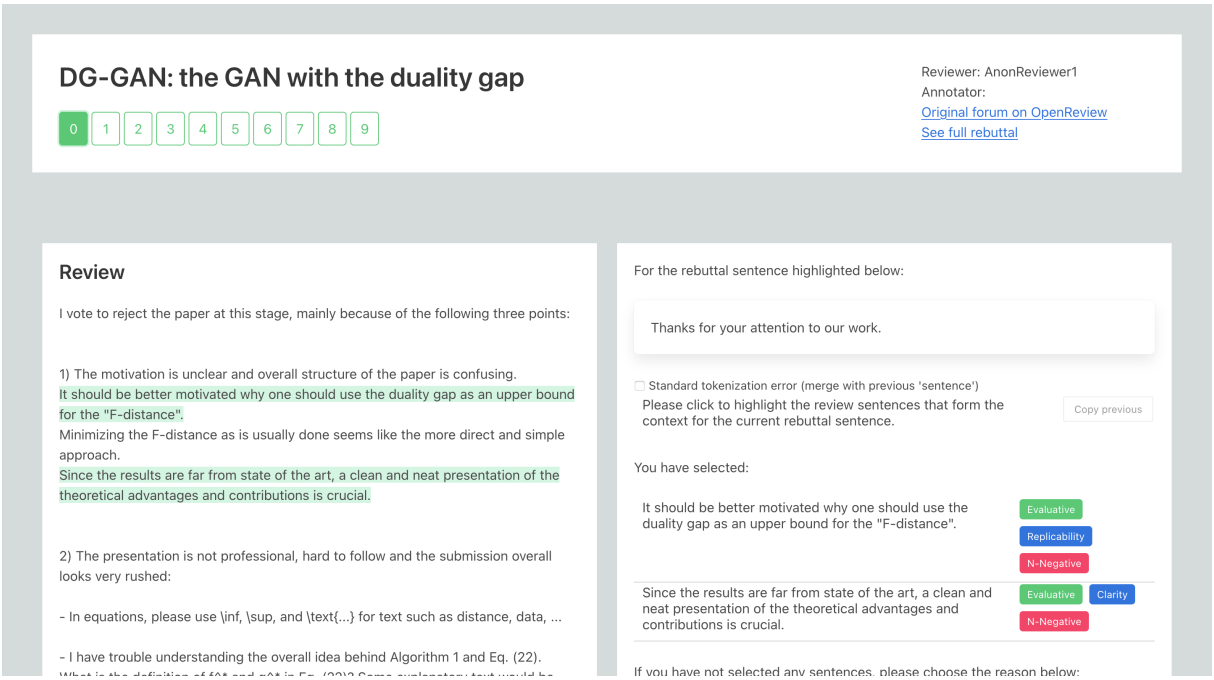

Figure 6: Rebuttal annotation interface. Annotators select review sentences that form the context of the rebuttal sentence being annotated.

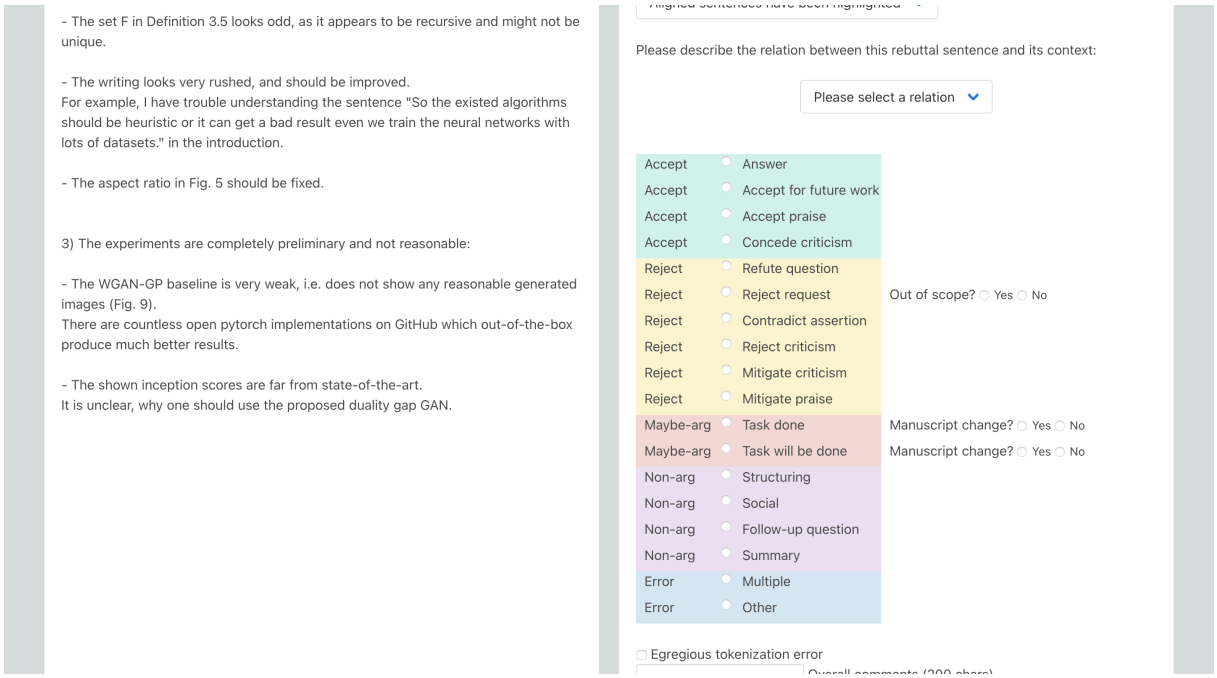

Figure 7: Review annotation interface. Annotators select a label for the rebuttal sentence.

```
{
  "metadata": {
    "forum_id": "ryGWhJBtDB",
    "review_id": "BJgmhEfTcH",
    "rebuttal_id": "rye3zaZ7or",
"title": "Hyperparameter Tuning and Implicit Regularization in Minibatch SGD",
    "reviewer": "AnonReviewer3", "rating": 3, "conference": "ICLR2020",
    "permalink": "https://openreview.net/forum?id=ryGWhJBtDB&noteId=rye3zaZ7or",
    "annotator": "anno10"
  },
  "review_sentences": [
    {
      "review_id": "BJgmhEfTcH",
      "sentence_index": 0,
      "text": "This paper is an empirical contribution regarding SGD arguing that
          it presents two different behaviors which the authors name a noise
          dominated regimen, and a curvature dominated regime.",
      "suffix": "",
      "coarse": "arg_structuring", "fine": "arg-structuring_summary",
      "asp": "none", "pol": "none"
    },
...
    {
      "review_id": "BJgmhEfTcH",
      "sentence_index": 4,
      "text": "I find the observations interesting, but the contribution is
          empirical and not entirely new. It would be nice if there were some
          theoretical results to back up the observations.",
      "suffix": "",
      "coarse": "arg_evaluative", "fine": "none",
      "asp": "asp_originality", "pol": "pol_negative"
    }
  ],
  "rebuttal_sentences": [
    {
      "review_id": "BJgmhEfTcH", "rebuttal_id": "rye3zaZ7or",
      "sentence_index": 0,
      "text": "We thank the reviewer for their comments.",
      "suffix": "\n\n",
      "coarse": "nonarg", "fine": "rebuttal_social",
      "alignment": [ "context_global", null]
    },
    {
      "review_id": "BJgmhEfTcH", "rebuttal_id": "rye3zaZ7or",
      "sentence_index": 1,
      "text": "Although our primary contributions are empirical, we also provided
          a detailed theoretical discussion in section 2, where we give a clear
          and simple account of why the two regimes arise.",
      "suffix": "",
      "coarse": "dispute", "fine": "rebuttal_reject-criticism",
      "alignment": ["context_sentences", [4]]
    },
    ...
  ]
}
```


Figure 8: A (truncated) example from the training set of DISAPERE.