DISAPERE: A Dataset for Discourse Structure in Peer Review Discussions

Neha Nayak Kennard **Tim O'Gorman** **Rajarshi Das**
Akshay Sharma **Chhandak Bagchi** **Matthew Clinton**
Pranay Kumar Yelugam **Hamed Zamani** **Andrew McCallum**
University of Massachusetts Amherst
{kennard, togorman, rajarshi, akshaysharma, cbagchi,
mfclinton, pyelugam, zamani, mccallum}@cs.umass.edu

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

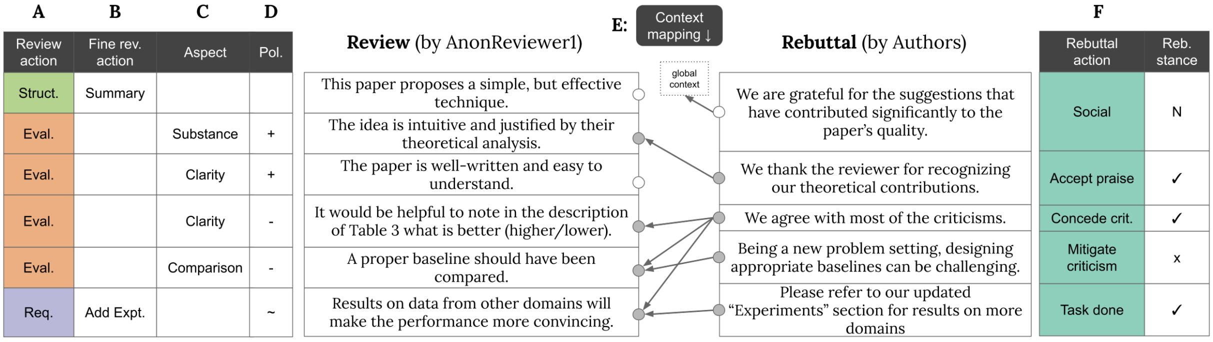
Peer review performs the essential role of quality control in the dissemination of scientific knowledge. The recent rapid increase in academic output places an immense burden on decision makers such as area chairs and editors, as their decisions must take into account not only extensive manuscripts, 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 manage 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 an understanding of the technical content, precluding the use of standard crowdsourcing techniques. Prior work on discourse in peer review has focused this qualified labor force on labeling arguments extracted from the text, which enables the complete annotation of more examples, at the expense of 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 rebuttals.

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[1]. We give reviews and rebuttals equal importance, and emphasize 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 were developed by synthesizing taxonomies from

[1] The dataset, along with its accompanying code and documentation, is available at http://www.github.com/nnkennard/DISAPERE/.



Review action	Fine rev. action	Aspect	Pol.	**Review** (by AnonReviewer1)
Struct.	Summary			This paper proposes a simple, but effective technique.
Eval.		Substance	+	The idea is intuitive and justified by their theoretical analysis.
Eval.		Clarity	+	The paper is well-written and easy to understand.
Eval.		Clarity	-	It would be helpful to note in the description of Table 3 what is better (higher/lower).
Eval.		Comparison	-	A proper baseline should have been compared.
Req.	Add Expt.		~	Results on data from other domains will make the performance more convincing.

Rebuttal (by Authors)	Rebuttal action	Reb. stance
We are grateful for the suggestions that have contributed significantly to the paper's quality.	Social	N
We thank the reviewer for recognizing our theoretical contributions.	Accept praise	✓
We agree with most of the criticisms.	Concede crit.	✓
Being a new problem setting, designing appropriate baselines can be challenging.	Mitigate criticism	x
Please refer to our updated "Experiments" section for results on more domains	Task done	✓

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 that demonstrates how decision makers could use models like these to understand trends and inform policies for future conferences (§ 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 the dataset (§ 6); (4) examples of analyses on the dataset that could benefit peer review decision makers (§§ 4 and 5), and (5) extensive annotation guidelines and software to support future labeling efforts.

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 of 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)'s AMSR dataset 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. Both view peer review as 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 it comments on.[2] In contrast to Yuan et al. (2021), we do not attempt or recommend generat-

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

ing peer review text, instead focusing on *analyzing* human-generated text in peer review.

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, 2021) present 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 the sentence level rather than the 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. While using manually defined discourse units (above or below the sentence level) may more precisely capture some discourse information, a separate pass of discourse segmentation can hinder the use of discourse datasets, as achieving consistent and replicable annotation of argument units is known to be highly challenging (Trautmann et al., 2020), and also because few works actually tackle unit segmentation (Ajjour et al., 2017).

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.

	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.	✓	✓		✓	✓
	Arg. types	✓				✓
	Polarity		✓	✓		✓
	Aspect			✓		✓
	Non-arg.					✓
	All sents.					✓
Rebuttal	Included?				✓	✓
	Arg. stmts.				✓	✓
	Context				✓	✓
	Arg. types					✓
	Non-arg.					✓
	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.

3.1 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 include 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,

Category		Label	Description	Percentage
REVIEW-ACTION		Evaluative	A subjective judgement of an aspect of the paper	32.83%
		Structuring	Text used to organize an argument	27.70%
		Request	A request for information or change in regards to the paper	19.82%
		Fact	An objective truth, typically used to support a claim	8.55%
		Social	Non-substantive text typically governed by social conventions	1.41%
		Other	All other sentences	9.71%
ASPECT		Substance	Are there substantial experiments and/or detailed analyses?	17.09%
		Clarity	Is the paper clear, well-written and well-structured?	11.08%
		Soundness/Correctness	Is the approach sound? Are the claims supported?	9.58%
		Originality	Are there new topics, technique, methodology, or insights?	3.85%
		Motivation/Impact	Does the paper address an important problem?	3.69%
		Meaningful Comparison	Are the comparisons to prior work sufficient and fair?	3.15%
		Replicability	Is it easy to reproduce and verify the correctness of the results?	2.86%
POLARITY		Negative	Negatively describes an aspect of the paper (reason to reject)	29.43%
		Positive	Positively describes an aspect of the paper (reason to accept)	11.16%
FINE-REVIEW-ACTION	Struct.	Summary	Reviewer's summary of the manuscript	18.17%
		Heading	Text used to organize sections of the review	8.54%
		Quote	A quote from the manuscript text	1.00%
	Request	Explanation	Request to explain scientific choices (question)	5.50%
		Experiment	Request for additional experiments or results	4.78%
		Edit	Request to edit the text in the manuscript	4.14%
		Clarification	Request to clarify the meaning of some text (question)	2.80%
		Typo	Request to fix a typo in the manuscript	1.98%

Table 2: A list of the review sentence labels, their descriptions, and the percentage of review sentences they apply to. Labels from all categories besides REVIEW-ACTION are optional, and thus may not add up to 100%.

requests for an explanation (e.g. of motivations or claims), requests for edits, and identification of minor typos.

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 label *positive* and *negative* 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), and *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 which a rebuttal sentence is responding to as the *context* of that

Category		Label	Description	Reply to	Percentage
Argumentative	Concur	Answer	Answer a question	Request	32.76%
		Task has been done	Claim that a requested task has been completed	Request	8.58%
		Concede criticism	Concede the validity of a negative eval. statement	Evaluative	2.70%
		Task will be done	Promise a change by camera ready deadline	Request	2.01%
		Accept for future work	Express approval for a suggestion, but for future work	Request	1.30%
		Accept praise	Thank reviewer for positive statements	Evaluative	0.35%
	Dispute	Reject criticism	Reject the validity of a negative eval. statement	Evaluative	10.37%
		Mitigate criticism	Mitigate the importance of a negative eval. statement	Evaluative	2.43%
		Reject request	Reject a request from a reviewer	Request	1.16%
		Refute question	Reject the validity of a question	Request	0.95%
		Contradict assertion	Contradict a statement presented as a fact	Fact	0.86%
Non-arg		Structuring	Text used to organize sections of the review	-	17.82%
		Summary	Summary of the rebuttal text	-	7.94%
		Social	Non-substantive social text	-	6.71%
		Followup question	Clarification question addressed to the reviewer	-	0.32%
		Other	All other sentences	-	3.75%

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. Each rebuttal sentence has exactly one REBUTTAL-ACTION label, so these percentages add up to 100%.

sentence, with special labels for when referring to the entire review (*global context*) or the empty set (*no context*). By not mandating a fixed discourse chunking, these annotations may handle situations when some rebuttal sentences respond to large sections of text, and other rebuttal sentences respond to specific sentences within those sections.

3.3 Data Source and Annotation

DISAPERE uses English text from scientific discussions on OpenReview (Soergel et al., 2013), which makes peer review reports available for research purposes. We draw review-rebuttal pairs from the International Conference on Learning Representations (ICLR) in 2019 and 2020, resulting in text within the domain of machine learning research. 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.

Authors are able to respond to each ICLR 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. While multi-turn interactions are possible, we focus on reviews and initial responses, and leave study of extended discussion for future work. The text is separated into sentences using the spaCy (Honnibal and Montani, 2017) sentence separator.

Annotation was accomplished with a custom annotation tool designed for this task, which is available as part of the code release accompanying DISAPERE. The tool is described in detail in Appendix B. Annotators annotate each sentence of a review, then examine the rebuttal sentences in order, selecting sets of review sentences to form their context. While this linking between sentences does not explicitly align multi-sentence chunks as in pipelined approaches to discourse alignment (Cheng et al., 2020), we note that since multiple sentences may be aligned to the same set of sen-

	Train	Dev	Test
Num. review-rebuttal pairs	251	88	167
Num. manuscripts	94	37	57
Num. adjudicated pairs	0	0	65
Num. review sentences	5216	1484	3246
Num. rebuttal sentences	5805	2015	3283
Num. review tokens	112k	33k	70k
Num. rebuttal tokens	131k	44k	75k

Table 4: Statistics for the dataset. Where possible, multiple reviews for the same manuscript were annotated. All reviews for any particular manuscript fall 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. While the original sentence boundaries were maintained, tokenization within sentences was carried out using Stanza(Qi et al., 2020).

tences in the review, some discourse structure is nevertheless latently implied.

3.4 Agreement

We report Cohen's κ (Cohen, 1960) on the IAA of labeling both review and rebuttals, treating each sentence as a labeling unit (Table 5). The annotators for each example are selected randomly from the pool of 10 annotators. Cohen's κ is calculated for sentences annotated at least twice. Where more than two annotations were produced, we calculate κ between all pairs and normalize by the number of possible pairs. The results show between moderate and substantial chance-corrected agreement between annotators, for both REBUTTAL-ACTION and REBUTTAL-STANCE labels (Appendix D provides details about agreement on context sentences). While these IAA scores do illustrate the noise of the task, note that this is not highly unusual for discourse labeling tasks – e.g. Habernal and Gurevych (2017) and Miller et al. (2019) both report α_u between 0.4 and 0.5.

Label	Cohen's 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.

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. Notably, 84.81% of sentences are linked to some review context. A small number of sentences refer to other sentences within the rebuttal, rather than any review context, posing a challenge for future work.

	Context type	Rebuttal sents. (Num.)	(%)
Sents. selected	Multiple contiguous	4696	42.29%
	Single sentence	4313	38.85%
	Mult. non-contiguous	407	3.67%
No sents. selected	Global context	816	7.35%
	Context in rebuttal	647	5.83%
	No context	152	1.37%
	Context error	61	0.55%
	Cannot be determined	11	0.10%

Table 6: Different types of rebuttal sentence contexts. Top: Over 84% of sentences are linked to a subset of sentences in the review. Bottom: sentences not linked to any particular subset of review sentences.

4.2 Alignment

One might reasonably hypothesize that the task of alignment between rebuttal and review sentences would be trivial, since authors are likely to respond to each point in the review in order. We can show that this is not 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 $\rho = 1.0$; this case is rare. Many examples with positive $\rho < 1.0$ indicate that authors do respond to points approximately in order, but a simple mapping based on order alone would not capture the correct alignment. Thus, while linear inductive bias may be beneficial to alignment models, the task of determining rebuttal sentences' contexts is not trivial.

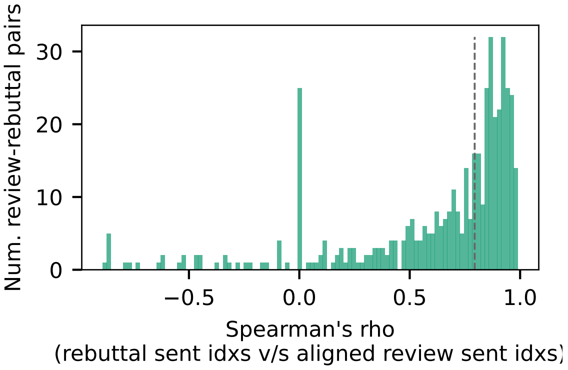


Figure 2: Spearman's ρ between rebuttal sentence indices and aligned review sentence indices. The dashed line indicates the median ρ value, which falls at 0.794.

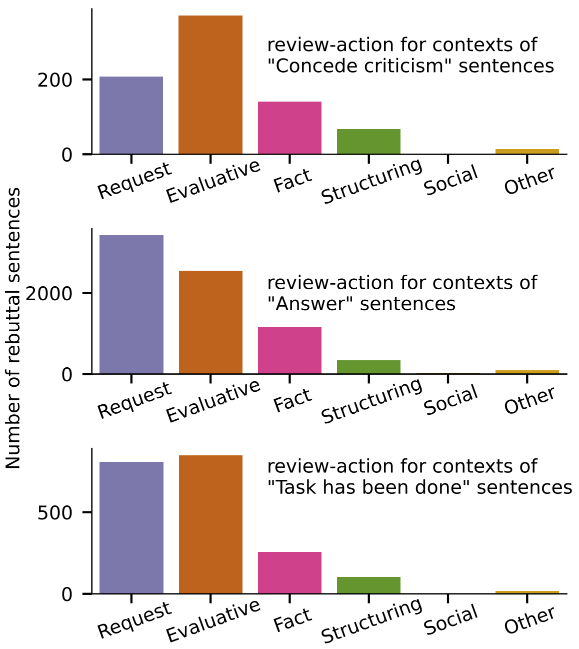


Figure 3: Distribution over REVIEW-ACTION for the context sentences of three REBUTTAL-ACTIONs. The canonical REVIEW-ACTION is marked by cross hatching. Note that authors sometimes interpret requests as criticisms ("Concede criticism"); often respond to evaluative sentences as if they are questions ("Answer"),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

4.3 Author interpretations of criticism

In our taxonomy, each argumentative REBUTTAL-ACTION corresponds to a particular REVIEW-ACTION, which we refer to as its *canonical* REVIEW-ACTION (listed in 'Reply to' column of Table 3). For example, *answers* are generally responses to *requests*, while *conceding criticism* is usually a response to an *evaluative* statement. Annotations revealed that authors often interpreted review sentences as if they embodied REVIEW-ACTIONs besides the canonical one, in a way that furthered the author's argumentative goal. For example, authors often responded to *evaluative* statements as if they were *requests*, perhaps in order to appease a reviewer, although no action was explicitly requested. Figure 3 shows the distribution of contexts for three different REBUTTAL-ACTIONs.

4.4 Relating discourse features to rating

Figure 4 shows one possible analysis taking into account the rating of the review. We show the distribution of FINE-REVIEW-ACTION labels of *requests* with review ratings. It appears that high-scoring manuscripts are rarely asked to add experiments, and are polished enough to not elicit requests to fix typos. Interestingly, low-scoring manuscripts have the second-lowest occurrence of typo requests, which could be due to the preponderance of other requests, but this bears further examination.

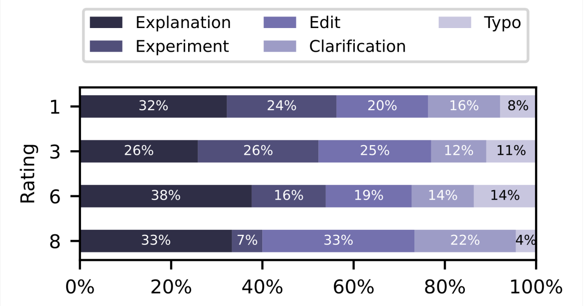


Figure 4: Distribution of REVIEW-ACTION labels, separated by rating

5 Application: 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 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 in a rebuttal, 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 5. Agreeability is only weakly correlated with rating, with Pearson's $r = 0.347$. In Figure 5, 18% (28/159) of manuscripts would not meet the bar for high variance scores (top quartile), although their low agreeability (bottom quartile) indicates that they may merit closer attention from area chairs[3].

[3] Two such examples included in DISAPERE: https://openreview.net/forum?id=r1e74a4twH and https://openreview.net/forum?id=HyMRUiC9YX.

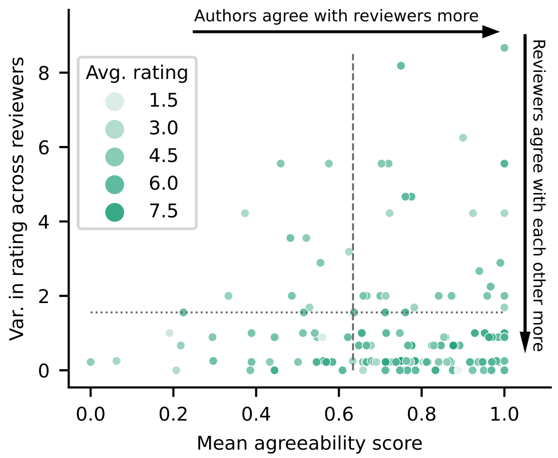


Figure 5: Mean agreeability for a manuscript's reviews v/s reviewer variance. Manuscripts above the dotted line are in the top quartile of rating variance, and are more likely to be reviewed by area chairs. Manuscripts to the left of the dashed line are in the bottom quartile of mean agreeability, in which authors take issue with the premises of reviewers' comments. The color of the dots indicates the mean of the reviewers' ratings.

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

For the six classification tasks, 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.

We report macro-averaged F1 scores, 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. While ASPECT achieves a particularly low F1 score, its κ is within the bounds of moderate agreement; thus, this must be accounted for by the inherent difficulty of the task rather than a deficit in data quality.

Classification task	Macro F1 (test)	Cohen's κ	Num. labels
REVIEW-ACTION	60.42%	0.605	7
FINE-REVIEW-ACTION	44.83%	0.583	10
ASPECT	38.28%	0.447	9
POLARITY	70.88%	0.561	3
REBUTTAL-STANCE	43.36%	0.513	4
REBUTTAL-ACTION	31.23%	0.479	17

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

As one might expect, errors in the classification results largely mirror disagreements in the annotations, which in turn reflect particularly ambiguous utterances. One example is the occurrence of rhetorical questions, such as (1) in Table 8, incorrectly labeled as *request* instead of *evaluative*. In fact, for sentences such as (1), additional context would disambiguate its type: the reviewer answers the question in the next sentence, and hence both sentences were labeled *evaluative*. Similarly, (2) was labeled *fact*, but since it is an integral part of a reviewer's argument against the soundness of the paper, should have been labeled *evaluative*. Certain reviewers also use conventions that do not fit the general schema we observed when developing DISAPERE. For example, (3), an opinionated heading, could be considered both *structuring* and *evaluative*. Finally, certain lexical cues a model may pick up on can be quite subtle. For example, though they share a prefix, sentences (4) and (5) are clearly *evaluative* and *request* respectively.

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). We add a `NO_MATCH` sentence to the review, to which rebuttal sentences without context sets in the review are aligned. Then, each review and rebuttal sentence is encoded independently using a S-BERT encoder and the similarity between two sentences is computed using cosine

		Label (Pred.)
1	Can the proposed [...] function represent all function the authors used in the paper? *Yes.*	E (R)
2	Matrices can have either "horizontal" or "vertical" redundancy (or "other" or neither).	E (F)
3	Solid technical innovation/contribution:	E
4	I am also wondering if the comparison with the baselines is fair.	E
5	I wonder if the authors ever looked at how much [...] determines the performance of the system?	R

Table 8: Example sentences including errors and challenging cases. E, R, F stand for *evaluative*, *request* and *fact* respectively. Letters in parentheses show the incorrect label from the model. Sentence (3) functions both as *evaluative* and *structuring*. Sentences (4) and (5) share a prefix but have different REVIEW-ACTIONS.

similarity. We initialize with a model[4] pre-trained on various sentence-pair datasets. Alignment is evaluated using mean reciprocal rank (MRR) and Mean Average Precision (MAP).

	S-BERT	BM25
MAP	0.4409	0.5174
MRR	0.5022	0.5980

Table 9: Rebuttal context alignment results. The results of both models indicate significant scope for improvement.

Surprisingly, the BM25 model outperforms a neural model (Thakur et al., 2021). While this shows that lexical information is a useful signal, both models have significant scope for improvement, and lexical overlap is clearly not sufficient for this task. Importantly, neither of these models account for the context of the rebuttal sentence, and predict each sentence's context independently. Incorporating this information is likely to lead to performance gains; however, we leave this investigation to future work.

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 data like DISAPERE will allow decision makers to glean deep insights on the interactions occurring during peer review.

Almost all publicly available peer review data is from the domain of artificial intelligence, limiting the scope of DISAPERE and any similar project. While this means that models trained on DISAPERE won't necessarily generalize to all new domains, we hope that with the detailed annotation guidelines and seamless data collection using the software provided with this paper support, users can build on our work, and ensure that their insights are robust to differences over time and across fields.

8 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.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation under Grant Numbers IIS-1763618, IIS-1922090, and IIS-1955567, in part by the Defense Advanced Research Projects Agency (DARPA) via Contract No. FA8750-17-C-0106 under Subaward No. 89341790 from the University of Southern California, in part by the Office of Naval Research (ONR) via Contract No. N660011924032 under Subaward No. 123875727 from the University of Southern California, in part by IBM Research AI through the AI Horizons Network, in part by the Chan Zuckerberg Initiative under the project Scientific Knowledge Base Construction, and in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–

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

128, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Liying Cheng, Tianyu Wu, Lidong Bing, and Luo Si. 2021. Argument pair extraction via attention-guided multi-layer multi-cross encoding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6341–6353, Online. 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.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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.

Tristan Miller, Maria Sukhareva, and Iryna Gurevych. 2019. A streamlined method for sourcing discourse-level argumentation annotations from the crowd. 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 1790–1796, Minneapolis, Minnesota. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9048–9056.

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

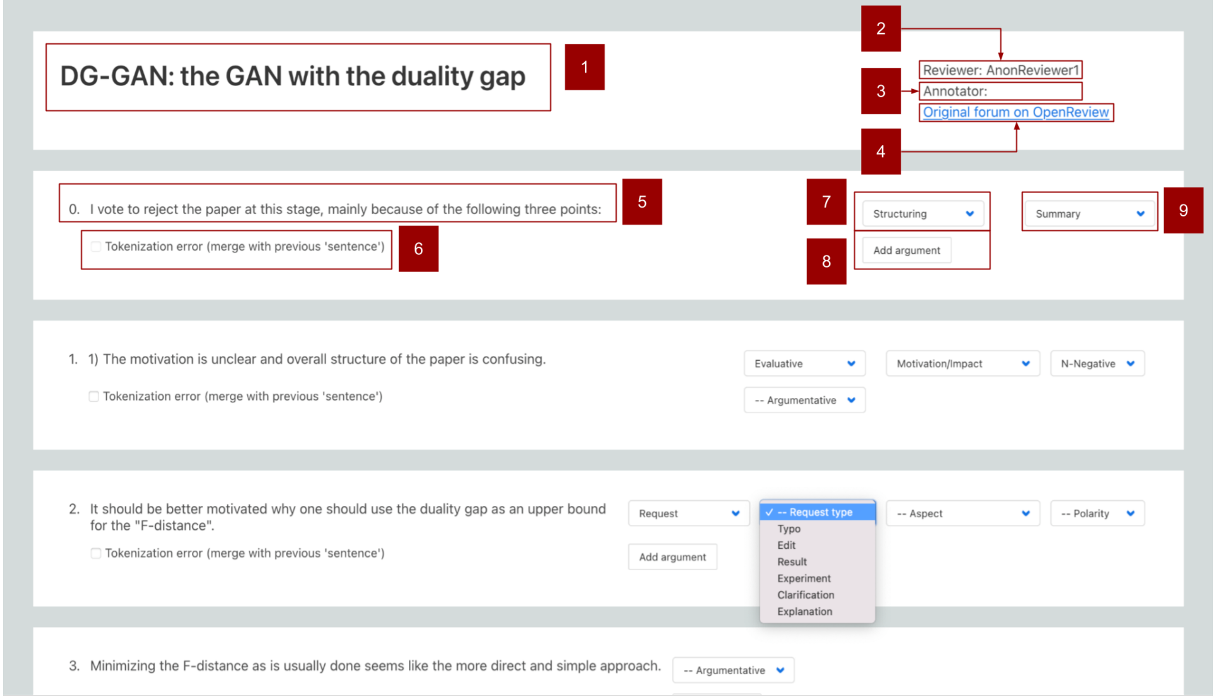
Our label sets leverage ideas from and commonalities between existing work in this domain, including AMPERE (Hua et al., 2019), AMSR (Fromm et al., 2020) ASAP-Review (Yuan et al., 2021), and Gao et al. (2019):

- 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 6.

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 7 and Figure 8.



DG-GAN: the GAN with the duality gap

Reviewer: AnonReviewer1
Annotator:
Original forum on OpenReview

0. I vote to reject the paper at this stage, mainly because of the following three points:
□ Tokenization error (merge with previous 'sentence')

1. 1) The motivation is unclear and overall structure of the paper is confusing.
□ Tokenization error (merge with previous 'sentence')

2. It should be better motivated why one should use the duality gap as an upper bound for the "F-distance".
□ Tokenization error (merge with previous 'sentence')

3. Minimizing the F-distance as is usually done seems like the more direct and simple approach.

Figure 6: 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 from false positives) [7] Dropdown for REVIEW-ACTION [8] Follow-up dropdownfor FINE-REVIEW-ACTION populated based on value in (7) [9] Button to add second REVIEW-ACTION if necessary (this was seldom used)

C Annotated review-rebuttal pair

Figure 9 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 on rebuttal sentences from 81 examples annotated by two annotators in Table 10.

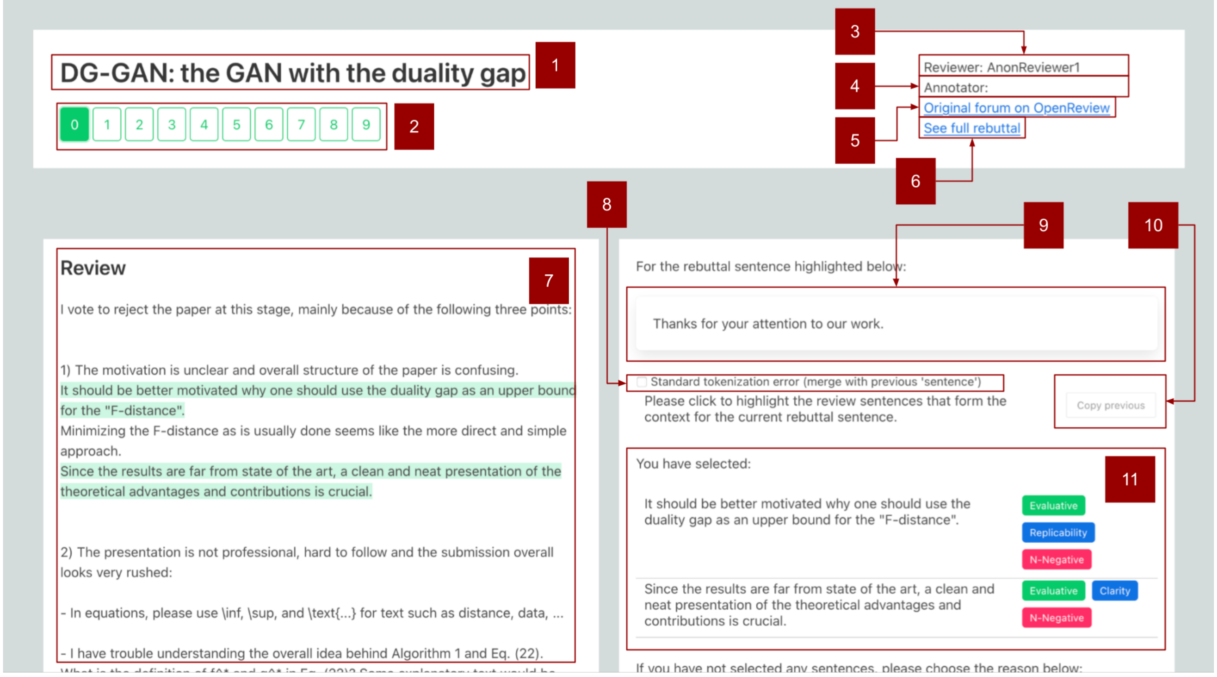
Type of context overlap	Num. rebuttal sentences	% rebuttal sentences
Exact match	914	53.11%
Partial match	492	28.59%
Agree none	122	7.09%
Disagree none	100	5.81%
No overlap	93	5.40%

Table 10: 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.

E Additional Agreement Analysis

While some of the IAA scores on annotation are low, we note that the labels used in this task attempt to characterize relatively complex relationships in text. To give more insight into such disagreements, Figure 10 provides a confusion matrix regarding the REBUTTAL-ACTION labels. Recognizing that there are often situations in which users of a dataset will hope to reduce a label set, we provide some guidance as to which such merges may be acceptable and which are not.

Many disagreements come from three labels which might be said to exist upon a continuum – ANSWER, MITIGATE CRITICISM and REJECT CRITICISM. We suggest that in the situation of needing to minimize IAA disagreement, one might consider first merging *mitigate criticism* into *reject criticism*. The kind of disagreements seen between the two are understandable but nuanced: the difference between saying that the reviewer has a point (but that they disagree on the relevance of that point) and disagreeing with the point itself. Out-of-context rebuttal sentences illustrating this are provided below as examples of this kind of ambiguous situation:



DG-GAN: the GAN with the duality gap

0 1 2 3 4 5 6 7 8 9

Reviewer: AnonReviewer1
Annotator:
Original forum on OpenReview
See full rebuttal

Review

I vote to reject the paper at this stage, mainly because of the following three points:

1) The motivation is unclear and overall structure of the paper is confusing.
It should be better motivated why one should use the duality gap as an upper bound for the "F-distance".
Minimizing the F-distance as is usually done seems like the more direct and simple approach.
Since the results are far from state of the art, a clean and neat presentation of the theoretical advantages and contributions is crucial.

2) The presentation is not professional, hard to follow and the submission overall looks very rushed:

- In equations, please use \inf, \sup, and \text{...} for text such as distance, data, ...

- I have trouble understanding the overall idea behind Algorithm 1 and Eq. (22).
What is the definition of f^* and g^* in Eq. (22)? Some explanatory text would be

For the rebuttal sentence highlighted below:

Thanks for your attention to our work.

□ Standard tokenization error (merge with previous 'sentence')

Please click to highlight the review sentences that form the context for the current rebuttal sentence.

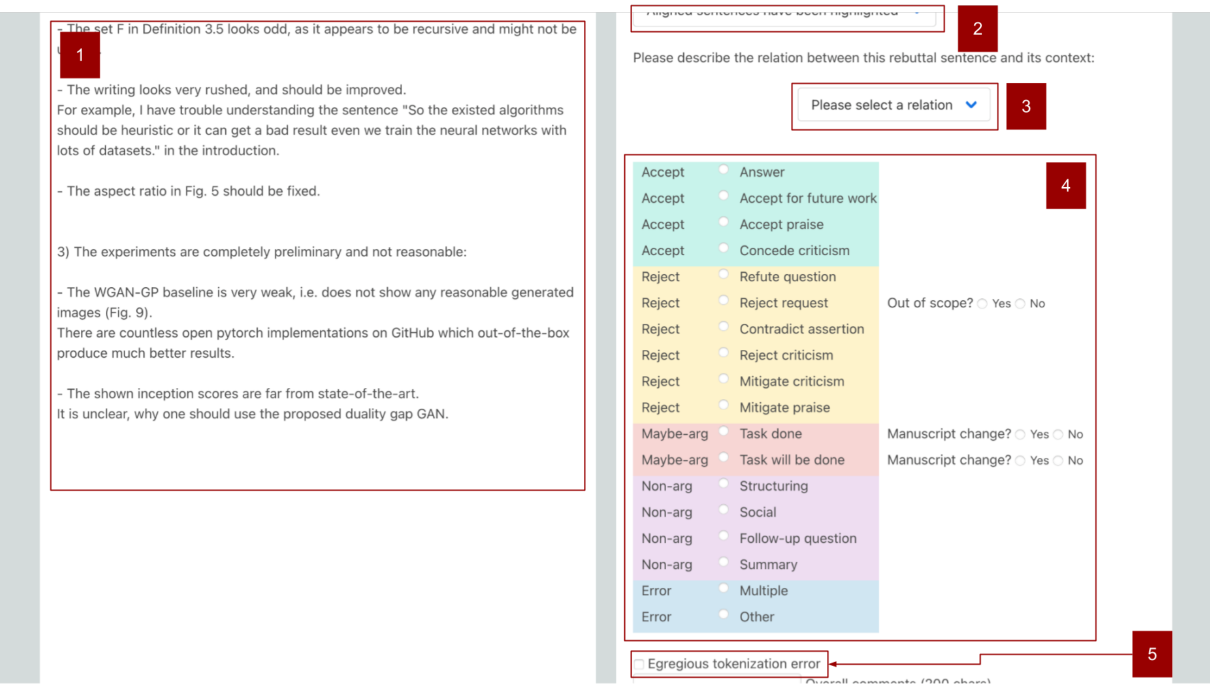
Copy previous

You have selected:

It should be better motivated why one should use the duality gap as an upper bound for the "F-distance".	Evaluative Replicability N-Negative
Since the results are far from state of the art, a clean and neat presentation of the theoretical advantages and contributions is crucial.	Evaluative Clarity N-Negative

If you have not selected any sentences, please choose the reason below:

Figure 7: Rebuttal annotation interface. Annotators examine each rebuttal sentence in turn, selecting sentences as context and specifying REBUTTAL-ACTION. [1] Title of the manuscript whose review is being annotated [2] Buttons to navigate between rebuttal sentences. Each page refers to a single rebuttal sentence (See (9)) [3] Reviewer identifier [4] Annotator identifier (removed for anonymity) [5] Link to original forum, in case it is needed for context [6] Link to open pop-up window with full rebuttal text, in case it is needed for context [7] Full review text. When a review sentence is clicked, it is highlighted and its details appear in (11) [8] Option to report sentence splitting error (false positive) [9] Rebuttal sentence being annotated [10] Button to copy REBUTTAL-ACTION label and context from previous rebuttal sentence [11] Table showing details of selected context sentences from the review, with the labels the annotator provided The screenshot is continued in Figure 8.


1
2
3
4
5
Aligned sentences have been highlighted
Please select a relation
Overall comments (200 chars)

- The set F in Definition 3.5 looks odd, as it appears to be recursive and might not be

- The writing looks very rushed, and should be improved.
For example, I have trouble understanding the sentence "So the existed algorithms should be heuristic or it can get a bad result even we train the neural networks with lots of datasets." in the introduction.

- The aspect ratio in Fig. 5 should be fixed.

3) The experiments are completely preliminary and not reasonable:

- The WGAN-GP baseline is very weak, i.e. does not show any reasonable generated images (Fig. 9).
There are countless open pytorch implementations on GitHub which out-of-the-box produce much better results.

- The shown inception scores are far from state-of-the-art.
It is unclear, why one should use the proposed duality gap GAN.

Please describe the relation between this rebuttal sentence and its context:

Accept	○ Answer	
Accept	○ Accept for future work	
Accept	○ Accept praise	
Accept	○ Concede criticism	
Reject	○ Refute question	
Reject	○ Reject request	Out of scope? ○ Yes ○ No
Reject	○ Contradict assertion	
Reject	○ Reject criticism	
Reject	○ Mitigate criticism	
Reject	○ Mitigate praise	
Maybe-arg	○ Task done	Manuscript change? ○ Yes ○ No
Maybe-arg	○ Task will be done	Manuscript change? ○ Yes ○ No
Non-arg	○ Structuring	
Non-arg	○ Social	
Non-arg	○ Follow-up question	
Non-arg	○ Summary	
Error	○ Multiple	
Error	○ Other	

□ Egregious tokenization error

Figure 8: Rebuttal annotation interface (continued from Figure 7). Annotators examine each rebuttal sentence in turn, selecting sentences as context specifying REBUTTAL-ACTION. [1] Full review text, continued from (7) in Figure 7. [2] Dropdown to select context type, in case context cannot be defined as a subset of review sentences. [3] Dropdown to select REBUTTAL-ACTION (keyboard navigation possible) [4] Buttons to select REBUTTAL-ACTION (in case mouse navigation is preferred) [5] Option to report egregious sentence splitting errors.

- *We note that such rules are indeed limited to some extent, but they still capture a rather expressive fragment of answer set programs with restricted forms of external computations.*

- *The use of C_p^{val} for hyperparameter tuning was incidental and not a central point of our paper.*

- *We agree that the measure theoretic approach is not always necessary (indeed for angular actions, it is not needed), but it is necessary for a very common scenario – clipped actions.*

Furthermore, we note that (as illustrated in the confusion matrix) a wide range of disagreements are hard to distinguish from “answer” labels, as authors often attempt to frame disagreements as simple answers to questions.

```
{
  "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": "",
      "review_action": "arg_structuring", "fine_review_action": "arg-
          structuring_summary",
      "aspect": "none", "polarity": "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": "",
      "review_action": "arg_evaluative", "fine_review_action": "none",
      "aspect": "asp_originality", "polarity": "pol_negative"
    }
  ],
  "rebuttal_sentences": [
    {
      "review_id": "BJgmhEfTcH", "rebuttal_id": "rye3zaZ7or",
      "sentence_index": 0,
      "text": "We thank the reviewer for their comments.",
      "suffix": "\n\n",
      "rebuttal_stance": "nonarg", "rebuttal_action": "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": "",
      "rebuttal_stance": "dispute", "rebuttal_action": "rebuttal_reject-criticism
          ",
      "alignment": ["context_sentences", [4]]
    },
    ...
  ]
}
```

Figure 9: A (truncated) example from the training set of DISAPERE.

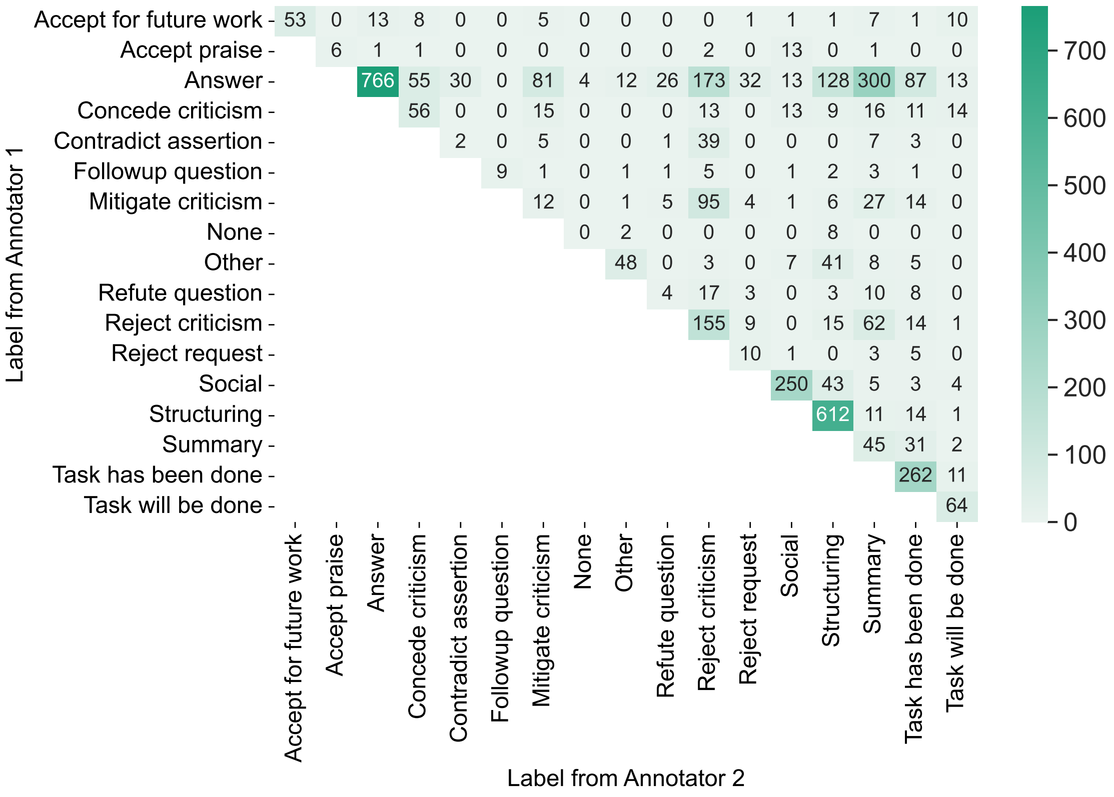


Figure 10: Confusion matrix showing agreement between annotators for REBUTTAL-ACTION labels.