Data Statement for DISAPERE

Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das,
Akshay Sharma, Chhandak Bagchi, Matthew Clinton,
Pranay Kumar Yelugam, Hamed Zamani, Andrew McCallum

May 3 2022

1 Header

Dataset Title: *DISAPERE (**DI**scourse **S**tructure in **A**cademic **PE**er **RE**view)*

Dataset Curator(s): Authors of DISAPERE manuscript

Dataset Version: 1.0

Dataset Citation: n/a

Data Statement Authors: Authors of DISAPERE manuscript

Data Statement Version: 1.0

Data Statement Citation and DOI: n/a

Links to versions of this data statement in other languages: n/a

2 Executive Summary

This dataset annotates discourse structure between 506 peer reviews and their rebuttals in English, taken from the International Conference on Learning Representations in 2019 and 2020. It is intended to assist decision makers in the peer review process to analyze peer review discussions in order to better understand their dynamics.

3 Curation Rationale

This dataset was create to study interactions between reviewers and authors in academic peer review. By annotating discourse structure, we hope to assist decision makers in the peer review process to analyze peer review discussions in order to better understand their dynamics. The dataset includes peer review reports in English taken from the International Conference on Learning Representations (2019 and 2020), whose peer review process was hosted on OpenReview[1]. Each instance consists of a peer review and the authors' comment in response on OpenReview (considered a rebuttal). We selected 506 instances to form the main dataset, and these are split into train, development and test sets. Any other instances annotated are also included in the dataset for completeness, but were not used in experiments.

4 Documentation for Source Datasets

Peer review texts were gathered from OpenReview.

5 Language Varieties

DISAPERE consists of text in English. The texts are produced by members of the international academic machine learning community. Due to the anonymity conditions of double-blind peer review, the identities and demographics of speakers are not available, and we cannot further determine the language variety of each instance.

[1] http://www.openreview.net

6 Speaker Demographic

Due to the anonymity conditions of double-blind peer review, the demographics of speakers are not available. However, we can state that all participants are machine learning researchers. Reviewers further tend to be researchers who have published in ICLR or related conferences in the past.

7 Annotator Demographic

17 annotators participated in annotation. Of these, 10 were students hired for an hourly annotation position, and 7 account for authors who participated in small amounts of annotation, as well as adjudication.

1. First language
    - Bengali: 1 annotator
    - English: 5 annotators
    - Hindi: 5 annotators
    - Marathi: 1 annotator
    - Tamil: 1 annotator
    - Telugu: 3 annotators
2. Last degree attained
    - B.S., B.A, B.Eng. or equivalent in Computer science (major): 11 annotators
    - B.S., B.A, B.Eng. or equivalent in Computer science (minor): 1 annotator
    - M.S., M.A, M.Eng. or equivalent in Computer science: 4 annotators
    - Ph.D. or equivalent in Linguistics: 1 annotator
3. Age
    - Under 20: 1 annotator
    - 20-25: 11 annotators
    - 26-30: 3 annotators
    - 31-35: 2 annotators
4. Gender[2]
    - Man: 12 annotators
    - Woman: 5 annotators
5. Racial identity
    - Asian: 11 annotators
    - South Asian: 4 annotators
    - White: 2 annotators

[2]Gender was collected as free form text in our form.

8 Speech Situation and Text Characteristics

The text used in DISAPERE was collected through web forms on the OpenReview website, as part of discussions for peer review for the ICLR conference. The intended audience of reviews is the authors and ICLR area chairs, and that of rebuttals is the reviewers and the ICLR area chairs. Participants may have composed the text elsewhere, and edited it before entering it into the form. The topic of the texts is machine learning manuscripts, and both reviewers and authors are referring to a particular manuscript in any instance.

9 Preprocessing and Data Formatting

Data was required to be anonymous due to ICLR's double blind format. The data was separated into sentences using spaCy[Honnibal and Montani(2017)].

10 Capture Quality

Limits on comment length on OpenReview may have an effect on capture quality. We used heuristics to determine if a string of comments was intended to be a single comment with continuations. The data was collected in ASCII format, and so contains some idiosyncratic uses of punctuation in lieu of formatting.

11 Limitations

DISAPERE is limited in scope due to all reviews being drawn from a single conference. It is possible that results on DISAPERE will not generalize well to other domains in science..

12 Metadata

License: The dataset is released under the Creative Commons Attribution-NonCommercial 4.0 International license.

Annotation Guidelines: Guidelines included as a pdf with the dataset, and can also be found at https://github.com/nnkennard/DISAPERE/blob/main/DISAPERE/documentation/annotation_guidelines.pdf

Annotation Process: Screenshots of the annotation tool are provided in Appendix B of the paper. The code for the annotation server is available in the associated Github repository at http://www.github.com/nnkennard/DISAPERE/.

Dataset Quality Metrics: Dataset quality metrics are summarized in Section 3 of the paper.

Errata: n/a

13 Disclosures and Ethical Review

The annotation process was determined to fall into the category of *Not human subjects research* by the relevant body. The determination form is included as a pdf with the dataset, and can also be found at https://github.com/nnkennard/DISAPERE/blob/main/DISAPERE/documentation/irb_determination.pdf

Annotators were compensated USD 20 per hour. This amount was based on the average rate for annotators with similar qualifications going through the same hiring process. Annotator demographics were collected through a form that did not collect any identifiable information. The mapping from annotator initials to identifiers is not made public.

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.

About this document

A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software.

This data statement was written based on the template for the Data Statements Version 2 Schema. The template was prepared by Angelina McMillan-Major, Emily M. Bender, and Batya Friedman and can be found at http://techpolicylab.uw.edu/data-statements.

References

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