Something something scientific discourse openReview

Anonymous EACL submission

Abstract

An abstract bla bla

1 OpenReview dataset

1.1 Source and structure

todo: paragraphify

This paper uses data from discussions on OpenReview (**?**). OpenReview is a platform for scholarly peer review, used primarily in computer science, which allows anonymized double-blind peer reviewing. The platform makes anonymized preprints, reviews, rebuttals, meta-reviews and other comments available to the public during the review period of a conference or workshop, with the goal of encouraging more open scientific discussion. In particular, we use data from the 2019 and 2020 iterations of the International Conference for Learning Representations (ICLR). As an out of domain test set, we use comments from the Learning for Dynamics and Control (L4DC) conference.

OpenReview uses a system of tree-connected *notes*, each of which represents a paper or a comment. Each submission is assigned to a *forum*, which is modeled as a note without text. In each forum, the discussion takes the form of a tree of notes. Reviewers generally add their reviews as direct children of the forum note, as do area chairs with their meta-reviews. Authors are able to add comments as children of any notes already in the tree, as are outside commenters, who may do so either with their signature attached or anonymously.

The analysis in this section is on data from ICLR 2019 reviews. In total, 1419 papers were submitted to ICLR 2019, so the API offers 1419 forums. We split these forums into train, development and test sets in a 8:1:1 train-dev-test. This split is stratified such that forums in the bottom and top quintile by total number of comments in the discussion are split uniformly among train, development, and test sets. Table **??** shows the total number of forums and comments in each of the sets.

Participants in OpenReview discussions fall into five categories: (1) "Conference" – an account that posts the initial, official post that initializes the forum for a paper, (2) "AnonReviewer_n" – the anonymized reviewers in a forum (3) "Authors" – the authors for the paper, who submit comments as one unified entity (4) Area Chairs – who also submit comments as a unified entity (5) "Anonymous" – anonymous commenters and (6) Named commenters – OpenReview users who have elected to sign their comments using their OpenReview accounts. While the goals and activities of the "official" participants – Authors, Reviewers, and Area Chairs are well defined, those of named and anonymous commenters are less so – they may be todo: x, y, or z. In addition, multiple 'anonymous' comments may or may not have been made by the same author. To simplify our modeling decisions, we leave out named and anonymous comments, leaving only official participants. The number of comments retained is also show in Table **??** .

Official reviews are marked as such in the OpenReview API. We consider any response to an official review by the authors of the paper to be a rebuttal, even though there is no official designation of a comment as a rebuttal. The number of reviews and rebuttals in each split of the dataset is shown in Table **??**. Note that some authors do not respond to the reviewers individually, and others provide a joint response to all reviewers in a single 'metaresponse'. As a result, there is not necessarily a 1:1 correspondence between reviews and rebuttals.

1.2 Overall dataset statistics

1.3 Typographical features of reviews

todo: is typographical the right word

A preliminary study of todo: maybe look at more? review-rebuttal pairs revealed various strategies participants used to indicate structure in their comments and relations across comments. First, many authors separated their comments into paragraphs using a double newline. We found that todo: 80% of comments adhered to this convention, and only todo: ??% of the comments that did not adhere to this conversation contained more tokens than the median number of tokens per chunk.

todo: some examples of this We also noted that many commenters interleave segments of their response with text copied and pasted from the comment they are replying to. Although this seemed to indicate at first that the problem of relating subparts of comments to each other might be trivial, further investigation revealed that only todo: 55% of rebuttals shared sequences of 7 or more tokens with the review they were responding to. This observation led us to use the 55% of comments with explicit references to build a training set, such that we could train a model to relate sub-parts of rebuttals that did not use this device. todo: the next sentence really needs rephrasing The examples without explicit references must somehow implicitly indicate the part of the review they are referring to, whereas those with string matches don't necessarily contain implicit references. Thus, the instances with exact string matches can be used to construct a labeled dataset for a problem that is strictly harder than the matching problem for implicit string matches.

todo: plots for how many of the comments have structure, how many have exact matches

todo: describe rules

todo: some examples of how structure is shown without exact matches

1.4 Discourse act affordances

We also noted that many statements made in reviews that would not conventionally be considered questions were nonetheless responded to by authors in their rebuttals. This led us to develop a scale of review comment affordances:

- No reply required (e.g. "Thanks for this interesting submission!")
- Requires a simple (e.g. yes/no answer)
- Requires a clarification of something in the text
- Requires further explanation of something in the text
- Requires additional result tables
- Requires that the authors do additional experiments

todo: talk about which of these categories rebuttal chunks tend to call into?

1.5 Rhetorical Structure Theory analysis

We find that, in the conventional sense typical of Rhetorical Structure Theory research, the rhetorical structure of reviews and other comments tend to be similar and not very complex. They usually consist of an introductory paragraph, followed by a list of grievances against the paper. Sometimes these are divided into pros and cons, sometimes strengths and weaknesses, and sometimes they are just a list of questions.

todo: Results/insights from running Neural RST parser on this data

1.6 Preprocessing

We preprocess the text in order to take advantage of the typographical cues users employ in order to add structure to their comments. Sentence splitting and tokenization are carried out using Stanford CoreNLP (Manning et al., 2014). All comments are separated into chunks that are delineated by double newlines. We find that using additional syntactic annotation did not contribute to the todo: something.

2 Tasks

We define two tasks in this study. The first is within-thread chunk similarity (todo: rename this!!), in which we relate parts of a rebuttal to the parts of the review they are responding to. Given the data for this task, we also add a secondary task of inter-thread chunk similarity. Although comments generally cannot reply to other comments in separate threads, this task is central to developing a view of the discussion as a whole.

References

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual*

Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

A Rules used for extraction

B Annotation interface