Review data collection at *ACL

Iryna Gurevych, Ilia Kuznetsov
UKP Lab, Technical University of Darmstadt

09.03.2021

Executive Summary

In this document, we propose two initiatives aiming at facilitating ethical, consent-driven research in NLP for peer reviewing, and peer reviewing for NLP. The first initiative concerns the collection of peer review and manuscript data from *ACL events. We cover the major challenges in this process and propose solutions, as well as report on a pilot study conducted at COLING-2020, where some of those solutions were implemented. The second initiative aims to provide a structured way to study and compare peer reviewing workflows in the *ACL community. For this we propose to create a centralized repository of anonymized numerical data from ACL conferences. This repository can serve as an anchor point for future conference organizers and provide a bridge to the meta-science and science-of-science communities that have been working on similar issues in other disciplines and research fields.

A. Peer review and NLP

Peer review is the core quality control mechanism in academia. It ensures that publications meet the research standards, prioritizes the scientific outputs, and allows authors to get expert feedback on their unpublished work.

The explosive publication growth puts a strain on traditional peer review and exacerbates the already-present issues: reviewer biases, miscalibration, conflicts of interest, reviewer fatigue, high management workload. This is particularly pronounced in dynamic fields like AI/NLP, as well as in crisis scenarios like the COVID pandemic, which has triggered unprecedented publication activity in the biomedical domain.

Science is communicated through texts, and NLP has made big progress in helping researchers navigate, aggregate and summarize scientific manuscripts. Peer reviewing is only starting to receive attention. Natural language processing can address a variety of issues related to peer reviewing, incl. debiasing, COI resolution, reviewer matching and reviewing assistance. However, current applications of NLP for peer review are limited due to the lack of data. Classic double-blind peer review is opaque by design, and collecting this data is associated with a range of anonymity, privacy and copyright challenges. As global research is transitioning towards open reviewing and publishing models, it is crucial to establish best practices and workflows for handling peer review corpora and metadata, and NLP is the right community to take over this task.

Why this data?

Collecting peer reviewing data in general serves several purposes. From the **research community** perspective, the collected data can be used for annotation and training of models to support peer reviewing processes, incl. reviewing assistance, score calibration, conflict detection and debiasing. From the **pure NLP** perspective, peer review is an important and largely untackled area: most existing datasets and studies (Kang et al. 2018, Hua et al. 2019, Cheng et al. 2020 etc.) are based on OpenReview conference crawls from venues focused on machine learning (mostly ICLR). The copyright status of these datasets is not clear as no consent has been explicitly collected (see below). UKP Lab's contribution to peer review data collection in NLP is to design the workflows, collect, clear and preprocess the data, and give the data in form of ready-to-use datasets back to the research community.

From the **organizers'** perspective, the progress in NLP for peer reviewing triggered by the availability of data would result in novel applications to improve reviewing quality and reduce organizational workload. The data can be used as reference to train junior reviewers and promote fair evaluation. Finally, the meta-information from the numerical data alone (see *F.) would simplify statistical analysis of the reviewing process and help design and test future reviewing policies.

Why *ACL?

Devising a strategy for handling peer review data within the NLP/*ACL community has a number of additional advantages. Being part of the same community, **the contributors will benefit** from the subsequent datasets and applications as reviewers, as authors and as researchers. Starting within the home community gives better control over the reviewing workflows and the data, and reduces the communication overhead. The developed practices can strengthen the data collection methodology in NLP and be transfered to other sensitive text types, e.g. documents in the legal domain and clinical data. The closest existing alternative is OpenReview. As of today, the legal basis of the OpenReview data is unclear, there is no consent procedure and the reviewing workflow is not standardized. OpenReview is working on their policy and we are in touch with them; besides, ICLR - the main data source at OpenReview - is a general machine learning conference, constituting a potential domain shift when applied to papers and reviews for computational linguistics, resource papers, etc.

One of the core assets of the ACL community is the Anthology, which publishes the accepted manuscripts and supplementary materials under a liberal license. Our initiative would complement this effort towards open and responsible handling of research outputs.

B. This document

The goal of this document is to ground the discussion on peer reviewing data collection at *ACL conferences and in general. We discuss the peer reviewing workflow at *ACL conferences and the types of data generated during peer review. We summarize the key challenges in peer review data collection, incl. anonymity and privacy concerns, copyright, consent, and managerial overhead, and propose solutions to these challenges. We report on a pilot data collection project by UKP Lab at COLING-2020, where some of these solutions have been field-tested. Finally, we propose an initiative on collecting and publishing anonymized numerical data from the *ACL conferences, that can serve as rich, consent-free auxiliary signal for NLP models, promote interdisciplinary work between NLP and meta-science, and help address the bias and miscalibration in peer reviewing at *ACL.

C. Reviewing workflow

C1. Participants and Information flow

A generic peer reviewing campaign at *ACL proceeds as follows.

- **Program chairs** assemble the reviewing pool and prepare the infrastructure.
- **Authors** (A) provide anonymized blind submissions. These submissions are reviewed. If the event has a *rebuttal phase*, the **authors** can address some of the **reviewers'** concerns. If the submission is accepted, a camera-ready version is prepared and *published*.
- **Program chairs** assign the **reviewers** to papers.
- **Reviewers** (current reviewer R / all reviewers RR) write the initial review given the blind submission. A review might be given a reviewing guideline. A peer review consists of one or more text fields and one or more numeric fields, determined by the review template. Some fields are confidential and are only shown to the **program chairs**. If the event has a *rebuttal* phase, **reviewers** can communicate with the **authors** and update their reviews. If the event has a *discussion* phase, **reviewers** can exchange opinions about the paper in the discussion forum and update their initial reviews. The final review serves as input to the **program chairs** to make the final decision.
- Apart from numerous managerial activities, **program chairs** moderate the reviewer discussions, produce meta-reviews based on the inputs, and make the *final decision* on paper acceptance.

C2. Data

A single reviewing campaign at any *ACL event produces plenty of rich, heterogeneous, structured data. We outline the core data types involved in the process along with the comments regarding privacy and anonymity of this data. The initial author and the intended audience are given in parentheses. Program chairs have access to full data. For each category we specify if the data is private vs public, and anonymous vs open-identity.

- Blind submissions (A→RR). Private. Anonymous. Sensitive if rejected. Less sensitive if accepted.
- Initial reviews: texts and scores (R→A,RR). Private. Anonymous by design, but there is an anonymity risk due to signed reviews and (potential) author profiling of review texts. Scores are anonymous.
- Discussion board logs (R→RR). Private, sensitive. Anonymous by design. Some venues show reviewer names, anonymity risk if published without post-processing.

- Rebuttals (A→RR). Private. Anonymous.
- Final reviews: texts and scores (R→A). Private. Anonymous by design (but see above).

  Some venues show reviewer names for discussion, potential anonymity risk.
- Meta-reviews (PC→RR,A). Private. Anonymous by design.
- Camera-ready versions (A→Public). Published and distributed under CC-BY.
- Conference metadata. As of now, not published in a structured way. Anonymous. Includes reviewing templates, rebuttal yes/no, discussion yes/no, avg. reviewer workload, demographics, etc.

D. Questions, Challenges and Solutions

D1. Anonymity

Challenge: *ACL conferences implement double-blind peer review in which reviewers and authors are anonymous to each other. Some venues allow reviewers to sign their reviews (thereby disclosing their identity to the authors deliberately), and some venues display reviewer identities during the discussion period.

Solutions: To preserve anonymity, no metadata identifying the reviewers and the authors of blind submissions should be included in the datasets by default. If the identities of the participants are disclosed at any point of the peer reviewing campaign, care must be taken to track the potentially affected texts. Maintaining a stable anonymous identifier for reviewers within one event is crucial for bias and calibration studies and only adds minimal anonymity risk. The contributors must be explicitly informed about the theoretical risk of future de-anonymization via author profiling techniques. The contributors are free to disclose their identity voluntarily, as part of license attribution (D3).

D2. Privacy and Security

Challenge: Most data involved in the peer reviewing process is private and can only be used for dataset construction given explicit consent (see below). Privacy in peer review has two aspects, personal privacy and content privacy: it is crucial to protect not only the personal data of the participants, but also the research ideas presented in the blind submissions. Below we list content-sensitive data types involved in the peer reviewing process, from least to most sensitive:

1. **Blind-submission versions of accepted papers** do not leak unpublished results as they are accompanied by the actual publication. However, removing substantial passages in the camera-ready is - while not encouraged - still possible. For such cases, the authors must be given an opportunity to keep the blind-submission versions private.
2. **Reviews and rebuttals for accepted papers**. Since the papers are published, there is no risk of leaking unpublished results in the review texts. However, the reviewers must be given an opportunity to keep the reviews private. Technically, peer review texts do not *belong* to the authors of the reviewed manuscripts. However, it might be desirable to give the authors an opportunity to opt-out the reviews for their papers from

public access. If collected, rebuttals have to be opted in by the authors. Numerical data (scores) is safe as long as it is not associated with specific submissions.
3. **Discussion forums**. Discussion forum messages might contain pointers to the sensitive content in the papers, as well as pointers to reviewer identities in case they are open.
4. **Reviews and rebuttals for rejected papers**. Sensitive due to dissemination of ideas prior to publication. Less sensitive if the submission is published on arXiv.
5. **Blind submissions of rejected papers**. Sensitive due to dissemination of ideas prior to publication. Less sensitive if published on arXiv.

Solutions:

- Only accepted papers are to be considered.
- No personal data in the datasets, apart from the aggregate statistics, e.g. self-reported demographics and experience level.
- If rejected papers are ever to be considered, a privacy period of **minimum two years** allows dissemination of the unpublished results and mitigates the risks associated with rejected papers.
- Explicit informed consent from all involved parties listing possible risks, incl. the potential for deanonymization via author profiling.
- Secure storage and protected environment for running NLP experiments (bringing code to the data, similar to shared tasks e.g. the TIRA platform[1]). Must be coupled with open datasets to enable training the models.
- Sharing derivatives of the data, e.g. deep pre-trained models, poses minimal privacy risks while enabling use in applications and as auxiliary signals. It must be noted that recovering some information from pre-trained models is *still* theoretically possible, although with substantial effort, unless a principled privacy-preserving solution like differential privacy is implemented.

D3. Copyright

Challenge: To maximize the benefit for the community, it is highly desirable to make the data accessible under a liberal license like CC-BY, which allows free sharing and adaptation of the content, instrumental to producing new levels of annotation and training the models. However this objective conflicts with the anonymity of the data, as anonymous peer review reports and blind submissions cannot be attributed to their authors.

Solution: The content contributor may grant ACL a license to sub-license the content for public access according to the Creative Commons license terms, and be identified as the Licensor on behalf of the authors/original contributors[2]. Each content contributor can indicate how the individual contributor should be attributed. The license agreement will also include the option for the content contributor to remain anonymous and not be individually attributed. In the event a contributor wishes to become anonymous at a later time, the ACL license agreement can be amended to reflect such a change, and the individual's name can be removed from the source website. The license agreement should make clear and remind each contributor that the CC licenses are irrevocable once granted unless there is a breach. The resulting license statemement could look as follows: *Copyright © 2021 administered by the Association for Computational Linguistics (ACL) on behalf of ACL content contributors:*

---

[1] http://universaldependencies.org/conll17/tira.html
[2] The workflow has been suggested by a legal expert as a solution to CC-licensing and anonymity

Professor John Smith, Dr. Susan Lee, Dr. Michael Jones, and other contributors who wish to remain anonymous. Content displayed on this webpage is made available under a [Creative Commons Attribution 4.0 International License [creativecommons.org]].

D4. Consent

Challenge: To make the data collection process ethical and GDPR-compliant, the contributors must be explicitly notified of the data collection purposes and the associated risks, and given agency over their data. Obtaining consent during submission process creates a stress situation which might influence the responses.
Solution: Data should only be collected given explicit consent from the authors and the reviewers. Collected data should be precisely specified on each occasion. Consent should not affect data from the past events. Authors and reviewers should be notified in advance about the possibility of opting in their data. The consent should only be requested after the acceptance notifications to provide the reviewers and authors with full information and give them an opportunity to make a grounded, reflected choice. In addition, authors and reviewers should be made aware that the process is managed externally and that donating the data is optional even if the invitation message is sent by the conference PC chairs. To enable reproducible research and comply with the CC Licensing, the consent cannot be withdrawn: the license agreement should make clear and remind each contributor that the CC licenses are irrevocable once granted unless there is a breach.

D5. PC Workload

Challenge: Obtaining consent and extracting the data creates additional work for the program chairs.
Solution: External workflow managed by the community/moderator. Consent collection via external forms with subsequent automatic extraction and filtering on the CMS side (requires minimal programming).

D6. Summary

Based on the above considerations and discussions with the community, we propose the following general workflow as a summary:

- Notify all participants about the future data collection in advance.
- **On review submission**: obtain the license on reviews from the reviewers.
- **After the acceptance decisions**:
- If the paper is REJECTED, no further action is needed. This might lead to dataset bias towards good papers and positive reviews; this solution can be revisited at a later point, e.g. by introducing a two-year privacy period until data publication.
- If the paper is ACCEPTED (i.e. will be published)
    - Obtain the license on the blind submission and ask for permission to publish the reviews for the blind submission (if the reviews have been contributed).
    - If the authors AGREE to publishing their blind submission version and the reviews, and the reviewer AGREES to publish their review
        - Collect the blind submission version and the review texts from the conference management system

        - Publish the data under CC license as part of a research dataset; no additional anonymity period necessary as the papers are accepted.
    - If the authors DISAGREE and the reviewer AGREES
        - Collect the review texts from the conference management system
        - Keep for internal purposes and as test data in a protected environment, e.g. TIRA

E. COLING-2020 Pilot

To test some of these ideas in the field, we have conducted a pilot dataset collection at COLING-2020 in Autumn 2020. We have used an earlier version of the workflow proposed above, hence not all measures have been implemented. We started the negotiations in Spring and spent some time refining the consent forms and deciding on the workflow. Since implementing the consent forms via SoftConf turned out to be too expensive, we opted in for external forms via Microsoft Forms. The forms ask authors and reviewers for their SoftConf username and (authors) IDs of the publications they are willing to contribute. This data is used to filter the conference dump and extract the entries for which consent was given.

E1. Workflow

While refining the workflow for the first time has caused some overhead, the core steps for the consent collection did not require much time:

- (UKP) Prepare consent forms, set up the infrastructure, prepare notification texts [~one month due to the lack of standardized consent forms]
- (COLING PC) Send out the one notification e-mail after the acceptance decisions [30 minutes]
- (COLING PC) Authorize the ext. manager to get access to the conference data [10 minutes]
- (SoftConf) Provide conference data [10 minutes]
- (UKP) Filter and anonymize the data according to the consent [one-time several hours effort due to programming]

E2. Implemented measures

- **Anonymity**: no identifying metadata, random event-level identifier for reviewers (e.g. “R154”). Since the discussion involved reviewer identities, we only consider the last review version *before* the discussion started. The status of the contributed messaging board data is currently under discussion, as many board threads are incomplete due to the gaps in consent.
- **Privacy**:
    - No personal data fields
    - Explicit mention of the potential for author profiling in the consent
    - Only approached the authors of accepted papers. However, reviews were collected for all papers.
    - 2-year privacy period for the whole dataset.
    - In the future: secure test environment via TIRA

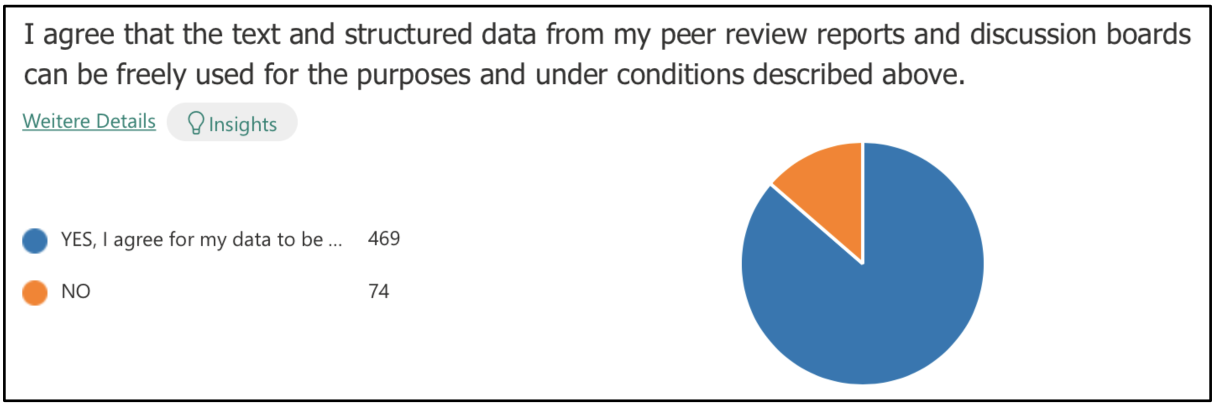
- **Copyright**: CC-0 for anonymous materials; later couple blind submissions with official ACL publications (CC-BY) for credit.
- **Consent**: explicit consent from reviewers and authors after the notification deadline with a two week window to make a decision.
- **PC Workload**: external forms, all programming on UKP side. Most effort spent on negotiating the consent forms.

E3. Consent forms

Reviewers
https://forms.office.com/Pages/ResponsePage.aspx?id=HBOHA8qCXkqX_VQgmzB6s4Y3gOE-fypItlQ_eix2fkBUN1pJUjlFUTBVNTFTUkpVR05GTk1MNkYyNC4u
Authors
https://forms.office.com/Pages/ResponsePage.aspx?id=HBOHA8qCXkqX_VQgmzB6s4Y3gOE-fypItlQ_eix2fkBUMVlENzNHOUVOT0ZJVlFOWDhJVzNPRUtDUS4u

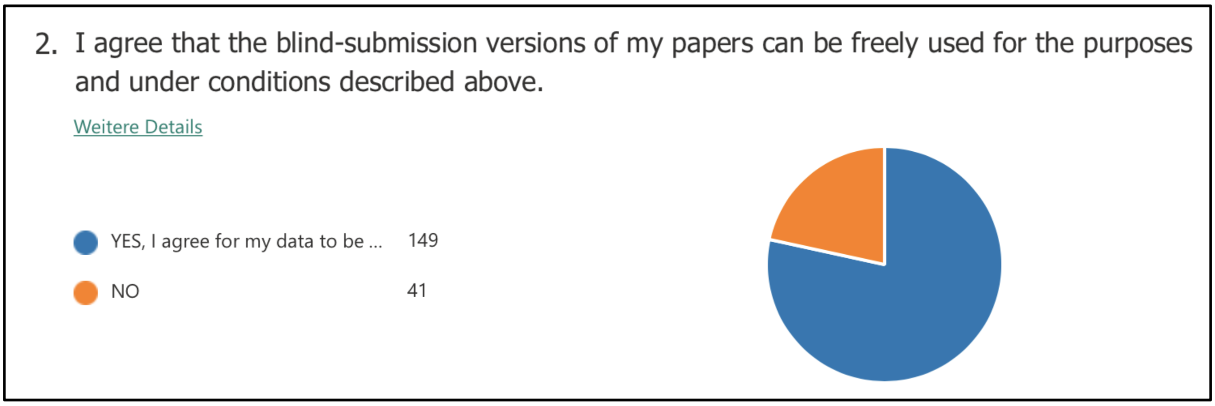
E4. Results

Out of ~1500 reviewers, 530 have participated in our consent survey by the deadline; most participants were willing to share their reviews and structured scores.



Out of ~580 authors of accepted papers, 190 have participated in the survey. Most authors were willing to share their blind submissions, although less enthusiastically than the reviewers. Approx. half of the authors provided TeX sources for their blind submissions; the TeX sources were automatically cleaned with the arxiv-latex-cleaner tool[3] to remove any private comments. The authors were informed of that and given a reference to the specific tool used to make the clean-up on author side possible.

[3] https://github.com/google-research/arxiv-latex-cleaner



As result, we collected explicit consent for 1300 review texts and associated scores for accepted and rejected papers, as well as over 140 accepted paper PDFs, 80 of which have also directly contributed the TeX source of their blind submissions. This is a substantial dataset equivalent to the ICLR-2017 portion of the PeerRead corpus, which did not ask for user consent and used the data directly, i.e. operated under the “opt-in by default”.

E.5 Future Improvements and Community Feedback

Authentication. Since integration with SoftConf is costly, we had to use external forms and a file drop for TeX sources. While this worked well, it would be good to have a way to authenticate the reviewers and authors, i.e. provide them with a unique code in the notification email. For now we just ask for the username. Transitioning to OpenReview as CMS might solve this issue as forms can be easily implemented on the CMS side.

Increasing the turnaround. In the COLING-2020 consent collection campaign, around 30% reviewers and authors have participated in the survey. While this already results in a substantially sized dataset, it would be great if this number were closer to 100%; this way we could get more data and a more precise estimate of the community’s opinion on the topic. For now the remaining 70% fall into a default opt-out, although this might be purely due to people skipping the notification mail. It would be useful to make the choice more explicit by requiring the participants to take a stance; optional feedback for negative cases would allow to collect the concerns from the community that prevent authors and reviewers from contributing their data.

Conference workflow details in advance. The reviewing workflow should be considered in advance; some key factors include the presence of author response and the visibility of reviewer identity during the discussion. While we took the former into account (which made the process easier), the latter came as a surprise during the communication with the users.

Notification before consent collection. To keep the process simple, we have only asked the COLING-2020 PC to send two emails to the conference participants (authors and reviewers, respectively). However, it could be highly beneficial to give the authors and reviewers a heads-up *before* the reviewing starts. This way people who plan to opt-in their data can take additional care in keeping their reports anonymous, and the authors can keep the blind submission TeX intact to contribute it later.

Better communication channel with the community. During the COLING-2020 consent collection campaign we received feedback via mail; while many were enthusiastic and encouraged the idea, there was also some useful constructive feedback and criticism. It would be better to let the community lead a moderated discussion on the topic in a public space, e.g. set up a forum for questions, answers and discussions regarding the consent collection and associated issues.

Adaptation to rolling review

Due to the upcoming transition towards rolling review at ACL, the workflows might need adjustments. However, since data processing (D6) is conditioned on acceptance and subsequent publication, the workflows would still be tied to particular ACL events.

*F. Open meta-repository for *ACL events

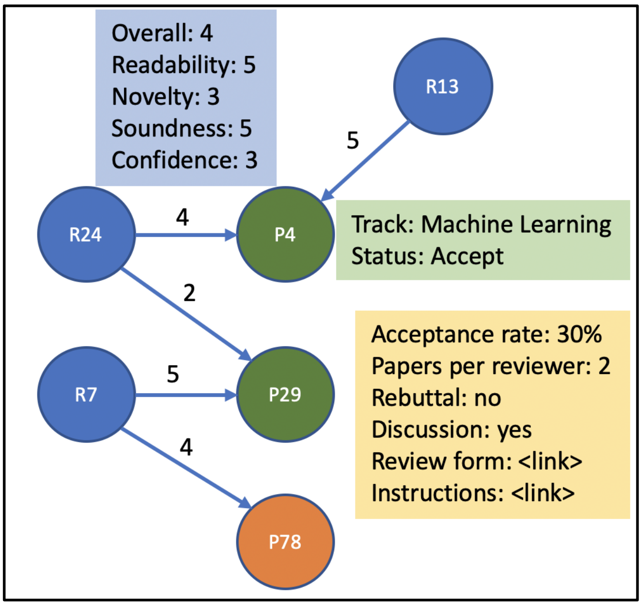
There is a need for structured, principled, meta-analysis of the peer reviewing processes in the *ACL community (Rogers & Augenstein 2020). Similar studies are performed in the neighboring communities, e.g. NIPS and ICML (Stelmakh et al. 2020abc), allowing to test hypotheses and optimize the reviewing workflows, however, these studies are performed on event basis. A centralized repository that would allow easy access to meta-information from previous *ACL events would help to encourage and ground similar efforts in our community. In particular, independent from peer review <u>text</u> collection, one can publish anonymous, <u>scores-only meta-datasets</u> for *ACL events. This would require minimal PC effort, does not pose privacy risks, does not require licensing, and will produce structured numerical data that can help optimize the reviewing processes in our community.

There exists a multitude of structural issues in peer review that can be studied and addressed without text processing, including reviewer miscalibration, commensuration bias, low reviewing quality and the opinion dynamics effects. Apart from the reviews themselves, a typical *ACL event produces a large amount of numerical and structural reviewing data which - coupled with a standardized representation of the venue’s reviewing workflow - can be studied and used to improve peer review in our community.

Most privacy and anonymity issues with peer reviewing data come from the texts. Scores, on the other hand, are perfectly anonymous and convey a lot of information. We propose to publish **meta-datasets for *ACL events**, consisting of:

- Anonymized reviewer-paper graph
- Review scores and confidence scores
- Acceptance decisions
- Standardized conference workflow metadata (author response yes/no, #papers per reviewer, etc.)
- Minimal metadata (long/short, track, etc.)

The diagram below depicts the structure of the envisioned dataset. This non-textual data can be extracted automatically from the CMS with minimal one-time programming effort and stored in a public repository, coupled with standard visualization and data exploration modules available to the community.



A single numerical dataset is a graph connecting anonymized reviewers and accepted/rejected papers coupled with additional review, paper and conference-level metadata.

The envisioned range of meta-datasets will enable comparisons between peer reviewing campaigns of different ACL events, keep track of historical data and answer questions such as:

- Does the reviewer discussion lead to higher reviewing quality?
- Does the contribution of aspect scores to the overall score change over time?
- Does the number of papers assigned per reviewer affect the reviewing quality?
- What are the reviewing style differences between different tracks?
- How does the composition of the reviewing form affect the reviewing results?
- and many others.

The numeric datasets will be made available to the NLP community and advertised in the meta-science community to foster collaboration between the “science of science” and NLP in both directions. Collecting this data over time would allow us to study the effects of different measures on peer reviewing quality and consistency. It will help conference organizers make informed decisions about peer reviewing workflows. The data can serve as rich non-textual auxiliary signal for NLP models. Automatically generating the standard and advanced conference statistics in a unified manner will reduce the PC reporting overhead for each event.

References

- Rogers, A., & Augenstein, I. (2020). What Can We Do to Improve Peer Review in NLP?. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 1256–1262). Association for Computational Linguistics.
- Kang, D., Ammar, W., Dalvi, B., Zuylen, M., Kohlmeier, S., Hovy, E., & Schwartz, R. (2018). A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. In Proceedings of the 2018 Conference of the North American Chapter

of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 1647–1661). Association for Computational Linguistics.

- Hua, X., Nikolov, M., Badugu, N., & Wang, L. (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) (pp. 2131–2137). Association for Computational Linguistics.
- Gao, Y., Eger, S., Kuznetsov, I., Gurevych, I., & Miyao, Y. (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) (pp. 1274–1290). Association for Computational Linguistics.
- Cheng, L., Bing, L., Yu, Q., Lu, W., & Si, L. (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) (pp. 7000–7011). Association for Computational Linguistics.
- Ivan Stelmakh, Nihar B. Shah, Aarti Singh, & Hal Daumé III. (2020a). Prior and Prejudice: The Novice Reviewers' Bias against Resubmissions in Conference Peer Review. arXiv:2011.14646
- Ivan Stelmakh, Nihar B. Shah, Aarti Singh, & Hal Daumé III. (2020b). A Novice-Reviewer Experiment to Address Scarcity of Qualified Reviewers in Large Conferences. arXiv:2011.15050
- Ivan Stelmakh, Charvi Rastogi, Nihar B. Shah, Aarti Singh, & Hal Daumé III. (2020c). A Large Scale Randomized Controlled Trial on Herding in Peer-Review Discussions. arXiv:2011.15083