Categorised open-ended responses

Other issues (in the last experience and historical)

AC / Action Editors

Interface

- It was difficult to find alternative reviewers who were not already assigned the maximal load

- You need to click on the reviewer's name to see anything and go through a list of paper + additionally click on google scholar profile. Note that reviewer's names aren't always clickable! Basically, you need to know all your reviewers or you'll be clicking too much. (...) November ARR reviewing cycle isn't reachable from my profile easily, b/c there's no direct link. It is not on the list of tasks b/c maybe there are no pending tasks, but can we have something like "current reviewing events?

- Another issue is modifying reviewer's assignments. Whenever there's a paper, there should be an opportunity to view/modify reviewers! Finding this tiny link "modify reviewer's assignment" is a fun exercise.

- OpenReview is better than softconf in that you can immediately access reviewer profiles in the reviewer assigning interface.

- Assignment system did not respect reviewers' requests for reduced load, requiring reassignments.

- Openreview system is not the best platform to use. It's hard to get to know how it works, who see your comments, how to communicate with the reviewers, and the chance for discussion is nearly non-existence.

- The qualified reviewers already had a heavy reviewing load, making it necessary to do complex switches between multiple people to ensure that everyone's assignments were improved.
-

Load

- It was difficult to find alternative reviewers who were not already assigned the maximal load
- Assignment system did not respect reviewers' requests for reduced load, requiring reassignments.
- In some occasions, the workload was too much, e.g. 25 papers + emergency load towards deadline

- The qualified reviewers already had a heavy reviewing load, making it necessary to do complex switches between multiple people to ensure that everyone's assignments were improved
- Generally, reviewers overburdened with review-related obligations - SAC/AC/reviewer for multiple conferences
- relevant reviewers already had a heavy load. Finding good replacements required manually checking reviewer profiles on their websites / publication aggregation websites.

ARR

- Currently AEs cannot check the maximum number of papers he/she would like to review in ARR within each specific period of time.
- On a related note: OpenReview's interface (though not horrible) is might confusing. For example, November ARR reviewing cycle isn't reachable from my profile easily, b/c there's no direct link…
- In the context of ARR, without author response or reviewer discussion, I find that I end up having to read the paper myself, making it hard to write a meta-review as I find myself injecting new content beyond what is in the reviews.
- ARR action editors have the chance to change reviewers, whereas in the past ACs don't have that option unless they chase it through SACs or PCs. Therefore that is a problem is a reviewer is not qualified or is not responding. However, still that setup is better that Openreview setup and the short reviewing timelines.

Affinity score

- The "affinity score", however, is garbage, so you have to look at every individual profile from thousands or profiles to identify qualified reviewers.

- Really imbalanced automatic review assignments - some reviewers got max and some 0, unpredictably

- the scores in OpenReview help a lot in moving this beyond people whose names you remember.

Information on reviewers

- To clarify a bit on the lack of information. On one hand, some basic information regarding reviewers is supposed to be present. In practice, there are two issues: 1. It's sparse (even areas of expertise are often empty). It is also can be too general (e.g., this is true for my profile haha) 2. It's not easily accessible: You need to click on the reviewer's name to see anything and go through a list of paper + additionally click on google scholar profile. Note that reviewer's names aren't always clickable! Basically, you need to know all your reviewers or you'll be clicking too much.

- The two issues I had were 1) not knowing reviewing history of reviewers, which could have helped me avoid reviewers that have not submitted on time in the past, and 2) some reviewers didn't provide any information about their background or experience, and searching online for their name didn't return any info, so I didn't know if they were qualified to review.
- relevant reviewers already had a heavy load. Finding good replacements required manually checking reviewer profiles on their websites / publication aggregation websites.
- One major problem in being an AC at *ACL conferences, which does not exist in workshops and smaller venues, is that I had limited involvement in selecting reviewers in the pool, and I did not know many of them. So, I found it hard to assign papers to reviewers whose expertise was essentially unknown to me. (also consequently found it hard to adequately weight the assessment of reviewers)

Bad reviewers

- Reviewers hold a huge amount of power over the process, but calibrating their responses and having them perform work that follows review guidelines is difficult, and often impossible to implement - because this all relies on the reviewers. The only hope is during meta-review, where one can ignore review comments that were inappropriate/irrelevant.
- Reviewers writing unacceptable comments in their reviews. I have messaged them to be more respectful to the authors.
- many papers with middle scores (around 3), many reviewers with low confidence (2 or 3)
- The quality of reviews has dramatically declined in the past two years. Maybe they don't understand the paper, but frequently, it seems like many reviewers are so unfamiliar with our area that they don't know what a good review should look like. There should be some kind of video tutorial and quiz required of people who are inexperienced
- There are also a lot of reviewers that are just lazy. They'll give a paper a 2 and then write two or three sentences saying what the paper is about without pointing out strengths or weaknesses and will ignore requests from the AC to write more. I think we need some way of filtering out these latter sort of reviewers. Could you just remove people from the pool who were outliers in terms of how much text they wrote in previous reviews?
-

Reviewer pool

- For ACL 2022, I am seeing many papers assigned that have, like, 2 junior PhD students and 1 industry person or postdoc who has two workshop papers. In the olden days, when reviewers assignments were supervised, you'd always get at least one senior reviewer per paper.
- Too many new reviewers - not necessarily unqualified, but unknown quantities in terms of knowledge/reliability
- Making good matches for specialized papers was hard -- there were qualified reviewers but not on very specific subtopics

Meta-review

- Reviewers hold a huge amount of power over the process, but calibrating their responses and having them perform work that follows review guidelines is difficult, and often impossible to implement - because this all relies on the reviewers. The only hope is during meta-review, where one can ignore review comments that were inappropriate/irrelevant.
- In the context of ARR, without author response or reviewer discussion, I find that I end up having to read the paper myself, making it hard to write a meta-review as I find myself injecting new content beyond what is in the reviews.

AC expertise

- Papers assigned to me as area chair were not in my own area of expertise, so it was harder to think of appropriate reviewers
- it makes more sense for ACs to have reviewers manually recruited specifically for a particular track. Then when you need to reassign, you have a roster of human-verified qualified reviewers to chose from, and since you're a domain expert, you can easily identify experienced/senior reviewers from the roster.

Reviewer recruiting

- In short, it makes more sense for ACs to have reviewers manually recruited specifically for a particular track. Then when you need to reassign, you have a roster of human-verified qualified reviewers to chose from, and since you're a domain expert, you can easily identify experienced/senior reviewers from the roster.

Communication

- Lack of sufficiently detailed instructions for how to navigate the new ARR system and how/when to communicate with reviewers
- The timeline for reviews and area chair duties was unclear in the system. I was not notified when a reviewer did not answer their invitation in a reasonable time. I was not told I was in charge of finding additional reviewers until very late
- In this case (TACL) the automatically generated emails only show the abstract to the reviewers. When the paper is a little out of the way, I often find myself looking for reviewers based on the body of the paper, not the abstract. But then the reviewers don't see what I saw in terms of what makes it possibly relevant to their interests and refuse... (I got around this by actually adding a sentence or two to the message about why I was asking them, but only after a couple of good reviewers turned the paper down.)
- Area Chairs needing to coordinate with Senior Area Chairs on reviewer assignments has at times led to a clunky / confusing system. I remember ACs recently had a spreadsheet of

reserve reviewers to draw from, with reviewers' numbers of assignments -- but the numbers were not being updated on the spreadsheet, so we were trying to assign papers to reviewers who had already received more reviews.

Declines

- In September ARR, I had around 20 reviewers decline to review. Hopefully that is fixed.

Mismatch in goals

- Reviewers frequently don't follow through with their commitments, which frustrates chairs and authors. It's understandable - there's no reward for reviewing on time, and being flaky as a reviewer has no consequences. We're missing a closed feedback loop: authors and reviewers are drawn from the same pool but have wildly different goals that do not align well.

Emergency reviews

- The most common issue were late reviews and looking for emergency reviewers in a very short timeline

Bias

- biases towards/against certain paper types when bidding is enabled

Bidding

- biases towards/against certain paper types when bidding is enabled

Reviewers

Bidding, overall workflow

- Reviewers should be allowed to bid or have a list of keywords to assign papers to them.
- My last completed assignment was with in ML and used a bidding phase, so the matching was OK.... My latest uncompleted assignment in CL is ARR, and I deeply regret there is no bidding phase.
- I prefer the conferences who offer bidding processes to select the papers to review. I have a lot more fun, I am more enthusiastic to review the papers compared to conferences that assign papers based on what my interests were x years ago.
- I once missed the bidding stage and all papers I had assigned at the end were completely outside my expertise. While bidding stage is a good process, reverting to assignment at random if a person missed the bidding is not good.

Areas of expertise - false positive from publication records

- My public papers/preprints usually covers more than one sub-areas of NLP. (One primary - more specific - and few non-primary - often generic - areas). I often get papers from the non-primary areas but not relevant to my direct area of expertise. E.g. Primary - dialog system, Secondary - representation learning

Areas of expertise - past vs present

- I switched between different NLP areas but still most reviews assigned are from my previous area. Although my publication record already has publications from the new area.
- The paper was generally in my area of expertise, but in a subfield where I just recently started working (<6 months). Thus, my expertise in that subfield is a bit limited, but I could generally understand and judge most of the paper's content.
- ARR assignments have been consistently assigned to a single, narrow area of my expertise. I don't mind reviewing papers in this area, but would like to have more diversity.
- Our current area of research and interest may not be reflected in the current publication list. I prefer the conferences who offer bidding processes to select the papers to review. I have a lot more fun, I am more enthusiastic to review the papers compared to conferences that assign papers based on what my interests were x years ago.

Kinds of mismatch

- When none of the: chosen languages, technologies (e.g. rule-based and hybrid), sub-fields (e.g. morphology or syntax) or other details (e.g. engineering within comp.ling, free/open science) are my specific expertise the reviews are a bit of waste of time.

Methods

- topic is the same, but techniques are something I did not work with.
- I do not work in machine learning or deep learning and am not qualified to review papers which describe these models and techniques but am consistently assigned them at large NLP conferences so I have stopped reviewing for conferences such as ACL or EMNLP

Languages

- Paper focused on a different language outside of the languages I speak.

Time issues

- Too many papers to review in short term

Forced choices

- Papers were added without asking for consent
- I tend to get the papers on "exotic" languages, it looks like nobody else is willing to review those; some conferences aren't careful enough with COIs
- I was placed in a track that I had little background in (against my stated track preference). When I requested to be changed to my preferred track, as I lacked an adequate background in the area to provide high-quality reviews, my request was denied.
- Sometimes you just get assigned papers that don't fit anybody else's bill, but I don't mind if this happens rarely enough.
- The worst situation was when I was asked to review a paper completely outside my area, and when I tried to get it reassigned, the chair told me that they didn't have anyone else and that I'd just have to do my best.

Interesting-ness

- Some of the papers that I was assigned did not use particularly exciting methods and didn't really ask questions that I thought were particularly interesting.
- Many papers are of interest to a relatively small group of experts in that area that are qualified and sympathetic. It used to be that papers would be reliably sent to such people. But that is no longer the case, and therefore, most accepted papers have to be of interest to a broader set of potential reviewers because a paper will be rejected if it is sent to a reviewer that is unqualified and/or unsympathetic.
- The paper had some aspects that were interesting to me (dataset, choice of languages), but the actual research question was not quite very for me.

Generations of NLP research

- The majority of "NLP" nowadays is obsessed wtih deep neural models, that are much more recent than my doctoral dissertation and my contribution as reviewer is often not within the details of AI technology that would probably benefit the authors the most. iow I feel most students have ~same amount of expertise as I do when reviewing these papers so the review does not benefit from my senior experience so much.
- The meaning of certain terms has changed. The area that I would consider myself an expert in has been taken over by papers that I don't really recognise as contributing to it, yet they use the same labels.
- Well, my previous research area was more focus con Corpus Linguistics, so I understand that it difficult to match a perfect assignment under these circumstances.

Bias

- It is really important to send papers to reviewers that are qualified and sympathetic with the area and the basic assumptions. Since so many reviewers know more about machine learning than computational linguistics, it is no longer possible to submit papers on computational linguistics to a conference on computational linguistics.

Authors

- The above problems are a good description of problematic reviews I have received in the past. However, I'd like to point out that reviews on my last submission were unique in combining ALL of these issues in a round of review for ONE single paper (and having meta reviewers acknowledge problems in reviews without dismissing them).

Short reviews

- The biggest issue is that reviews are very short. They aren't often terribly informative other than indicating whether the reviewer things the paper should be accepted or rejected. The worst case is when they say the paper is borderline and have very brief comments (not really clear why borderline). A firm accept or reject without much comment is at least some kind of feedback.
- Reviews being very short while positive are also not always great (e.g. "I really like this paper, topic is relevant!" without any other comment - I would have liked some engaging with the content of the paper, whether arguments or literature).
- ARR October reviews were very brief

Expectation for a certain kind of research

- In a paper specifically focusing on evaluation (of multiple systems), the comment from 2 reviewers… at a NLP workshop was: there is no new algorithm/method and we "just" evaluated existing approaches with new techniques.
- Gatekeeping, saying this work doesn't belong in our field.
- Disagreeing with the motivation of the work as a premise for research. This is a subjective opinion and not useful for consideration of the work.
- One thing I would like to point out is the difficulty with getting survey papers accepted. The last time I did - the comment from two reviewers has been that I should expand the review and submit to a journal to make it comprehensive?. But which journal accepts surveys? It is not clear if TACL does. CL Journal took 7 months to review my 3 page survey proposal only to reject it saying the topic is somewhat narrow for CL Journal. So, where can someone submit such survey papers that can be fit into 8 pages, and need not become 40 pages?
- The reviewer were just fond of some other model, and critized my model for not being of that type.
- Reviews that suggest the paper would be better suited to a different venue even though the topic of the paper was specifically included (e.g. MT for the ACL conference). The review then uses that "suggestion" to justify not providing substantial comments on the actual paper.

Confirmation bias

- Confirmation bias errors leading to low scores due to results disagreeing with prior work in a different task using different models in a different setting.

Mismatch between the score and the text of the review

- The reviewer could not find reasons to criticise the paper (and overall wrote a positive review), yet gave it a low score

- Weaknesses listed were minor issues easily addressed, making the reason for the low scores unclear.
- exaggeration of small mistakes, e.g. a false page number in one single citation (such as "p. 50" instead of "p. 51") out of dozens of citations

Non-constructive criticism

- criticism that is not constructive: "your approach is wrong, but I will not tell you how you could improve it"

Shallow reviews

- Generic criticism for "not enough novelty"
- The feedback was very shallow. The main question from all reviewers was about the application to other languages, which is a completely valid point, but I would have liked to also get some more in-depth feedback on the methods we used, which the reviewers mostly didn't comment on (maybe because of a lack of expertise or interest, but they didn't say this).
- ARR October reviews were very brief, and some of them criticized superficial aspects e.g., "what does terminology X refer to?" -- where I expected terminology X to be common knowledge to readers.

Inattentive reviews

- Reviewer obviously only skimmed the paper and didn't even understand the problem it was tackling
- The reviewer clearly did not read the paper carefully (e.g., ask explanations for things that can be found simply with CTRL+F, does not read captions or column titles in Tables)
- It was obvious the reviewer put very little time and effort to write the review.
- Poorly reading the paper (and asking us to run experiments that we did, in fact, run)
- General misunderstanding of our submission, potentially due to only superficial reading.

Lack of competence

- Did not understand the paper
- This reviewer also misrepresented the task that our paper was addressing (or failed to understand it).

Missing reviews

- it was only 2, not 3
- for one paper, some reviews have still not appeared (now more than a month after we expected them).

Requests for irrelevant comparisons

- asking for comparisons with irrelevant works

Requests for irrelevant citations

- Reviewers suggesting what are probably citations to their own work where it's not that relevant. (For some I think this is the reason they review, to police whether their own work is being cited.)
- a reviewer who demanded I cite several papers by the same author that were only tangentially related to my paper.

Impact

- Wild guesses that work won't have impact.

Unannounced policy changes

- Desk rejection was used by chairs as a way to cut through the number of sumbission. this policy was not made sufficiently clear before submission deadline and the chair did not disclose the content of the reviews anyway. Desk rejection decision was made after senior PC were given 5 minutes of time to evaluate wheter a submission was worth the whole review process or not. This policy was not disclosed before submission deadline.

Suggestions for alternative workflows

ACs

Bidding

- "Bidding isn't perfect but can help a lot. Similarity scores would be nice but ARR's are not good from what I've seen"
- I know bidding is a pain, but it's useful as an area chair and also ensures that the reviewer gets papers where she/he has interest. For cases where nobody has interest, automatic similarity scores are useful.

Similarity + manual

- "Automated similarity scores may be helpful but if a manual check is not *required* I worry that there will be some papers that fall through the cracks."

Bidding+manual

Bidding+similarity

- bidding information & similarity

Bidding+similarity+manual

- Maybe a mixture of similarity-based matching and bidding-based matching plus manual adjustments might be a good try.
- "reviewers should be first assigned automatically based on both similarity scores and biddings, and then manually checked"
- Bidding information is frequently incomplete. I would also use automated similarity scores, manual check…
- Bidding + similarity scores + manual checks. The first two provide different signals and the third is essential.
- "ICLR style: bidding, where the list of papers is ranked by automated similarity scores. Then, manual adjustments by SACs/ACs. "

Keywords

+ bidding

- Authors should be able to tag their paper along an extensive, but still finite, set of tags (e.g. anACL-version of ACM CCS concepts, or FAccT's submission tags). These tags can be used to narrow down papers for reviewer bidding as a second step, particularly if reviewers also tag themselves. This should be more fine-grained than traditional ACL tracks, and also multi-field selection (ACM CCS concepts also allow three levels of priority/relevance).

+ sim
- Reviewers should be able to bid on the papers using two things (as it was done, e.g., in Neurips): 1. keyword search, 2. similarity scores. Editors should be able to check this information manually.

Tracks

- When I've seen my own matches by similarity score and area, area is always better and similarity score is anywhere between random and inversely correlated with appropriateness."

+ bidding
- "Bidding + area selection + reviewer distribution with manual check would help

Other

- "Bidding information is frequently incomplete. I would also use automated similarity scores, manual check, seniority and affiliation (to avoid a paper being reviewed by three very homogeneous group, e.g. all PhD students)."
- "Useful information includes: biding information, similarity scores, knowledge of the reviewer's expertise, COI and quotas. "

Interface

- Ideally, we should have an interface where reviewers are listed together with all the basic summary information + a link to their profile.

ARR

- Similarity scores would be nice but ARR's are not good from what I've seen

Reviewers

Manual

- If we have time to review papers manually, we have time to assign papers by manually. The assignment is more important than reviewing. Area chairs do not have that many papers. They can do the assignments manually.

Similarity + manual

Similarity + bidding

- Auto to suggest + bidding
- Filter based on similarity scores and then bid
- If bidding were to be used, it should be done on a reasonable subset of the papers, either from automated similarity scores or reviewer-specified track.

Bidding+manual

Bidding+similarity+manual

- Bidding with supportive similarity scores and manual checks at the end.
- Combination of similarity and biding (as in ICLR2021) with possibility for manual adjustments.
- Bid on a list of papers ranked by similarity where you only have to check abstracts for 50-100 papers pre filtered to be possibly relevant. Then ACs can fine tune the assignment
- papers should be assigned using all three (not sure why this is not offered as option, only combinations of two): automatic similarity scores, bidding information, and final manual checks

Keywords

- "**have papers provide keywords or key phrases** like many other fields which indicate what the authors believe are the salient points of the paper;

+ sim

- The similarity score should weight keywords for potential areas in which I would like to review instead of just the past publication history.

Tracks

- increase (rather than remove!) tracks and track specificity — see for example Interspeech. **we are not ready for fully automated matching, and unconstrained fully automatic matching is definitely a step too far"**
- Specifying the kind of papers I'd like to review (similar to identifying tracks or subject areas), and using this information in addition to any automatic matching based on my past research.

+ bidding

- If bidding were to be used, it should be done on a reasonable subset of the papers, either from automated similarity scores or reviewer-specified track.

Other

- They should have the option to do what we used to do, **namely send the papers to be people who are cited in the papers**. We should not assign reviewers to area chairs in advance. There should be a process for load balancing higher up the tree. But it is super-important that the reviewers be qualified and sympathetic. That is more likely if the reviewers are likely to be cited in the paper.
- The recency of my own publications should be considered when computing the similarity.
- We are computational linguists, we should be able to come up with a better automatic matching strategy based on prior publications.
- A mix of **bidding + random** assignment as an exploitation-exploration setup.

Authors

Similarity

- "A score could be helpful for area chairs, but it needs to be clear how it is computed. It could be useful to know 1) there is an overlap (predicted or provided) topic, 2) the paper cites the reviewer, 3) the paper uses the same methodology/evaluation/dataset the reviewer has used, etc.
- If a score is used, the danger is that the SACs rely blindly on it, and manual checks might not be enough to undo potential harm if the score doesn't work well."

Similarity + bidding

- Initial automatic assignment followed by optional bidding
- Auto + bidding

Similarity + manual

Bidding+manual

Bidding+similarity+manual

- automatic assignment for a pool of papers, then bidding, finally manual checks
- Combination of similarity scores, bidding and manual checks? Typically, papers that reviewers bid on could be filtered using automatic scores.

Keywords

Random

- Reviewers should be asked what area they would like to review for, and then assigned papers randomly in that area subject to COIs. One of the motivations for reviewing is to learn about work in an area of interest, this may not be the area in which I've done the most previous work, so automated methods tend to give me work I am less interested in seeing sometimes.
- Bidding + some random assignment to ensure diversity in the matching. We don't want reviewers to review only papers they *want* to review. However these random assignments should be clearly indicated to all, and treated accordingly.

Other

- Combination of all signals.

ARR

- I think ARR/OpenReview matching system seems to be working well. I got good matches with that.

- I just want to say that I have noticed much better review assignments since the automatic matching process has begun, and unfortunately that seems not to have been carried over to ARR!

Other information presented to ACs and whether it was useful

(AC responses only)

Reviewer history

- "Information about the experience other area chairs had with this reviewer" would be really helpful.
- Regional location\nWork affiliation type: e.g. industry, academic\nReview history/experience/reliability

Number of assigned papers

- Number of currently assigned papers (better with the maximum number of papers the reviewer would like to review) to judge the appropriateness of additional workload

Being able to ask SACs for advice

- Before ARR I could ask for help or advice to the senior area chairs in my track. This was extremely useful and it is not available in current ARR implementation.

Reviewer affiliation

- Regional location\nWork affiliation type: e.g. industry, academic\nReview history/experience/reliability

Correct area match

- If I am manually checking reviewer assignments for a few papers at a time, the main important factors are (a) the paper is in an area I am familiar with and (b) enough experts in that area are in the pool.

Other

- Items 20 and 21 don't apply, b/c they apparently ask about my area chairs experience with other venues, but I have none.