What Can We Do to Improve Peer Review in NLP?

Anna Rogers
Center for Social Data Science
University of Copenhagen
arogers@sodas.ku.dk

Isabelle Augenstein
Department of Computer Science
University of Copenhagen
augenstein@di.ku.dk

Abstract

Peer review is our best tool for judging the quality of conference submissions, but it is becoming increasingly spurious. We argue that a part of the problem is that the reviewers and area chairs face a poorly defined task forcing apples-to-oranges comparisons. There are several potential ways forward, but the key difficulty is creating the incentives and mechanisms for their consistent implementation in the NLP community.

1 Introduction

Traditionally, peer review is expected to act as a filter for high-quality, impactful work (Wingfield, 2018), but this does not hold in practice:

- *Peer review does not guarantee quality control*, neither for small errors nor for serious methodological flaws – even in biomedical literature, where publishing flawed results does real damage (Smith, 2010).
- *Peer review fails to detect impactful papers.* The correlation between conference rejection rates and conference impact in terms of citations is not strong (Freyne et al., 2010; Ragone et al., 2011, 2013), and rejects from one conference sometimes receive awards[1] at another.

The problem is that both expectations are unrealistic to begin with. A peer reviewer cannot perform real quality control, because that would mean ensuring that a paper is *reproducible*. Not only is that impossible, only having a few hours to review a paper, but it is a general problem for Deep Learning (DL)-based NLP (Crane, 2018; Rogers, 2019). The reproducibility checklist at EMNLP 2020 (Dodge et al., 2019) is the first step in that direction.

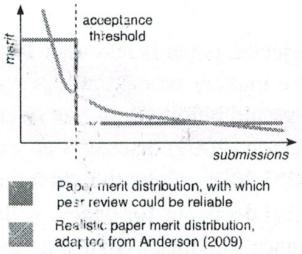

Figure 1: Paper merit distribution

As for paper impact, it is distinct from its scientific merit (Bhattacharya and Packalen, 2020), and strongly depends on completely orthogonal factors: how niche is the topic, how much promotion was done, whether the paper offers room to innovate with a low entry barrier[2] (Anderson, 2009).

What we *could* realistically expect from peer review is to reject the papers with obvious methodology flaws, and turn the spotlight on the ideas which would be beneficial for the field to discuss. However, the current process is not set up to achieve either purpose. Instead, it aims to rank all submissions by their merit so as to identify the top 25%. That task, we argue, is fundamentally impossible.

2 Why Is Peer Review So Difficult?

Peer review would be easy if the paper merit distribution had a clear boundary between good and bad papers (and ideally that boundary would match the conference acceptance threshold). However, that is clearly not the case. Based on citation counts, Anderson (2009) hypothesize that paper merit is Zipf-distributed, as shown in Figure 1. That means that even with the most objective reviewers, the difference between the worst accepted paper and

[1] Mani (2011) discusses the example of a paper by Brana-van et al. that received the award at ACL 2009 after being rejected from NAACL (scored at 2.3/5). More recently, ELMo (Peters et al., 2018) received low scores from ICLR reviewers and was resubmitted to NAACL to win the award there.

[2] Amongst the biggest success stories in DL-based NLP are word2vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019). Note that both of them contributed a transfer learning paradigm with room for incremental modifications.

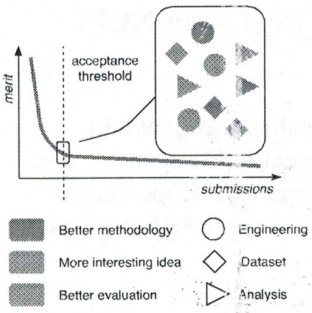

Figure 2: Why it is hard to compare borderline papers

the best rejected paper is less than 1%.

To make matters worse, there are no clear criteria that would help to draw the decision boundary. Anderson (2009) discusses an experiment at SIG-COMM 2006, where they first made the easy accept/reject decisions for papers with low review score variance, and then assigned up to 9 additional reviewers to papers with high variance. The reviewers who had to discuss the difficult cases were reportedly "nearly driven insane" by the apples-to-oranges comparisons, such as incomplete evaluation in one borderline paper vs narrow applicability of another. No matter how long we agonize over such decisions, they will look random. A case in point: at NIPS 2014, 10% submissions were reviewed by two different PCs, who disagreed on 57% of papers (Price, 2014).

For large *ACL conferences, the situation is even worse: we often weigh against each other different *types* of papers with different strengths and weaknesses (Figure 2). There can be no 'correct' answer as to which one has more scientific merit.

3 How Reviewers Cope

Faced with an objectively impossible task, reviewers do what humans generally do to reason under uncertainty: they default to heuristics, which introduces unwanted biases (Korteling et al., 2018). There is an extra incentive to do so because apples-to-oranges comparisons are a slow, deliberate, cognitively expensive process – and peer review is currently invisible work performed for free.

This section lists some of the most problematic reviewer heuristics in NLP.

Writing style. Language errors, non-standard style or rhetorical structure are easy to spot and interpret as sloppiness. This puts almost everybody at a disadvantage against North Americans. Papers with worse English may even be perceived as worse than those with better content (Church, 2020).

Results not surpassing SOTA. An easy heuristic is to check if the paper beats the state of the art (SOTA). While an engineering contribution should demonstrate a significant improvement over prior methods, it does not have to be an improvement in *predictive performance*: advances in compute or data efficiency, interpretability, cognitive plausibility etc. are also valuable (Rogers, 2020a). The focus on predictive performance encourages the 'arms race' for pre-training data and compute, and exacerbates methodological issues[3]. The requirement for comparisons with the latest SOTA model also puts us in the hamster wheel, making experiments outdated already by the submission time.

Narrow topics. It is easier to publish on trendy, 'scientifically sexy' topics (Smith, 2010). In the last two years, there has been little talk of anything other than large pretrained Transformers, with BERT alone becoming the target of over 150 studies proposing analysis and various modifications (Rogers et al., 2020). The 'hot trend' forms the prototype for the kind of paper that should be recommended for acceptance. Niche topics such as historical text normalization are downvoted (unless, of course, BERT could somehow be used for that).

Work not on English. Since prototypical NLP experiments use English as the target language, other languages mark the paper as narrow. This heuristic is indefensible, since approaches tested only on e.g. Estonian are as generalizable as those tested only on English. It also strengthens the 'default' status of English (Bender, 2019).

Already-famous work and work from well-known labs. If reviewers feel that a paper was already accepted by the research community, they do not need to do any more vetting. For example, there was no way for BERT to go through fully anonymous peer review (Cotterell, 2019).

Early preprint citations are arguably a better indicator of paper quality than peer review (Church, 2017, 2020), but they are also influenced by how famous the authors are[4], and how much they pub-

[3]Inter alia: unfair comparisons (Musgrave et al., 2020), dependence on non-architecture-related factors (Dodge et al., 2020; Crane, 2018), no incentives for producing robust systems (Ethayarajh and Jurafsky, 2020), flawed benchmarks (Jia and Liang, 2017; McCoy et al., 2019), which become a tool for producing incremental papers (Reiter, 2020)

[4]Peters and Ceci (1982) resubmitted 12 articles to psychology journals that already *published* these articles, with the author names changed to unknown names. Many were rejected for 'methodology flaws'! See Rogers (2020c) for discussion of anonymity in upcoming ACL peer review reform.

Oxford Internet Institute talk on Snake oil & "hype"

Dictatorship of English.

Bell labs Schön scandal

but then the question is, if we seperate them by category, what's to stop us seperating into sub-categories et ad infinitum?

Ah, is it dangerous or useful to have a standard though? Does the sword cut both ways if we say, "English articles only" for submission?

licize their work. Well-known labs tend to have large online followings or even PR departments, propagating the 'rich get richer' phenomenon (also known as the 'Matthew Effect', Merton (1968)).

The proposed solution seems too simple. Since a prototypical 'acceptable' paper features a sophisticated DL model, simple solutions may look like the authors did not do enough work. This is obviously flawed, as the goal is to solve the problem, rather than to solve it in a complex way.

Non-mainstream approaches. Since a 'mainstream' *ACL paper currently uses DL-based methods, anything else might look like it does not really belong in the main track - even though ACL stands for 'Association for Computational Linguistics'. That puts interdisciplinary efforts at a disadvantage, and continues the trend for intellectual segregation of NLP (Reiter, 2007). E.g., theoretical papers and linguistic resources should not be a priori at a disadvantage just because they do not contain DL experiments.

Resource papers. Surprising as it may seem in a field that relies so much on supervised machine learning, resource papers are routinely rejected simply for being resource papers. Linguistically deeper papers may also receive extra penalties for linguistic details at the cost of DL experiments, for non-English resources, and, by analogy with the SOTA heuristic for engineering papers, for not offering the *largest* resource (Rogers, 2020b).

Novel approaches. This sounds almost absurd, but scientific peer review is systematically biased towards unobjectionable (rather than novel) work (Church, 2005, 2020; Smith, 2010; Bhattacharya and Packalen, 2020). A reviewer faced with evaluating a completely new idea without prior art has to make a more difficult call than one for a paper with clear predecessors and a leaderboard table, and is more likely to fall back to one of the heuristics. The very process based on majority votes necessarily promotes 'safe', incremental, likely boring work, and puts non-mainstream work at a disadvantage.

Substitute questions. The question *"how good is this paper?"* is difficult, because the criteria for scientific merit are vague. What humans often do to answer a difficult question is to unconsciously substitute it with an easier one (Kahneman, 2013). We suspect that one of these substitutes is *"are there any obvious ways to improve this paper?"* This would explain the acceptance rate gap for long and short papers:[5] since the latter include fewer details and experiments, they are easier to find fault with. In our experience, another such substitute question may be *"if I did this study, would I make the same choices?"* The reviewers using this heuristic are not actually responding to the real paper they were assigned, but to an imaginary paper more in line with their interests and methodology – and the real paper compares unfavorably.

EMNLP 2020 explicitly addressed most of the above heuristics in its blog (Liu et al., 2020), but naming and shaming is unlikely to be sufficient. Heuristics *are* the way humans reason under uncertainty, so the only way to fix this is to clarify the very task reviewers are expected to perform.

4 Can We Just Abolish Peer Review?

If the task is fundamentally impossible, should we just give up and look for alternatives to peer review? Each round of conference notifications spurs calls on social media to just abolish the whole system, to increase acceptance rate, to let citations be the metric of the paper quality.

Unfortunately, this is not realistic, even if there were no co-dependence between citation counts and scientific fame or promotion efforts. Fundamentally, low acceptance rates are a proxy for paper quality for non-experts, and that metric is expected by almost every hiring and grant committee. We are not aware of serious proposals on how to change that. And any experiments will require a generation of extremely brave students who are willing to graduate with no 'respectable' publication record.

EMNLP 2020 essentially increased the acceptance rate by creating a second-tier publication named *Findings of EMNLP*, which has "no requirement for high perceived impact, and accordingly solid work in untrendy areas and other more niche works will be eligible".[6] It enabled the organizers to accept 15.5% of extra submissions (including this one), in addition to 22.4% in the main track.

Unfortunately, this approach does not address the fundamental issue (comparing the incomparable), and introduces new problems:

- the very existence of *Findings* is likely to exacerbate reviewer biases: they may give lower scores to 'non-trendy' work to nudge it to-

[5] In 2020. 24.6% long vs 16.7% short papers at EMNLP, 25.4% vs 17.6% at ACL, 35.5% vs 27.7% at COLING.

[6] https://2020.emnlp.org/blog/2020-04-19-findings-of-emnlp/

wards *Findings*, even if not explicitly asked for main track vs *Findings* recommendations);

- no matter what status *Findings* attains in the community, in the academic rankings it will always remain a second-tier outlet, and that will change trajectories of careers and grants of people who engage in 'non-trendy' work;
- *Findings* implicitly caters to 'fast science': rather than improving a paper, authors can publish it as is and move on. In the short run, this helps the authors (particularly those whose SOTA results are likely to 'expire'). In the long run, it means more papers which are less well executed.

Finally, *Findings* also decreases the likelihood that a new top-tier venue would emerge to make the 'untrendy' topics trendy, and potentially change the direction of the field. Ironically, EMNLP itself was born as a home to papers rejected by conservative ACL reviewers (Church, 2005). If ACL had created *Findings* in 1996, there would likely be no EMNLP today – and the whole field might be less empirical.

5 So What Can We Do?

Until there are systemic changes in how researchers are evaluated, peer review remains 'the least bad system available' (Smith, 2010). Still, there is clearly room for improvement.

First, peer review has to become a valued part of academic CVs, and something that employers budget time for. Reviews done by overworked people in their free time will not be top-quality.

Second, we need to reduce the need for reviewers and ACs to reason under high uncertainty. It cannot be fully eliminated, but there are several obvious directions for improvement.

Better reviewer matching. Reviewers are more likely to resort to heuristics when they are not experts in the same narrow area as the paper they are reviewing. A matching should take into account both the tasks and the methods (e.g. a paper on coreference annotation is unlikely to be appreciated by a practitioner who only worked on coreference applications). Since it is not always possible to find perfect matches, reviewers with complementary partial expertise (e.g. someone who speaks the language if the paper is not on English, plus an area expert) could be a fall-back strategy.

More fine-grained tracks: ACs should never have to decide between different types of papers. If surveys, opinion pieces, resource and analysis papers etc. are all welcome, they should have their own acceptance quotas and best paper awards.

Review forms tailored for different paper types: it does not make much sense to evaluate a reproducibility report for novelty, or a resource paper for SOTA results. COLING 2018 developed review forms taking into account different types of contributions, possibly several per paper.

Announcing editorial priorities pre-submission. What is the primary focus of a particular conference: SOTA engineering, diversity of approaches, fresh ideas? What counts more towards acceptance? Stating this clearly would help authors find an appropriate event for their work, and help reviewers and area chairs be more consistent in their recommendations.

Not asking the reviewers for overall recommendation scores. This is where similar papers get seemingly random rankings from different reviewers, because they disagree on whether e.g. originality outweighs weaker evaluation. Even having a clear policy does not help[7] (Noothigattu et al., 2020). The obvious solution is that reviewers should only be asked for specific scores (originality, technical soundness etc.), which would be the basis of the decisions according to the editorial policy.

The above solutions focus on reducing apples-to-oranges comparisons. A fundamentally different approach is to increase reviewer accountability, e.g. by making reviews public. Unfortunately, this does not address the core problem (reasoning under high uncertainty), and would introduce other problems.[8]

6 What Holds Us Back?

At this point, the reader might join the disappointed anonymous reviewers of this paper and say that we are not proposing anything new. This is precisely why the problem is so difficult: we lack the implementation, not the conceptual innovation – and as researchers, we tend to only value the latter.

On the organization side, each *ACL conference is organized by a new set of people each year who set their own policies. Such diversity by it-

[7] AAAI 2013 aimed to select the "exciting but imperfect" papers, and provided the reviewers with instructions about how to compute the overall recommendation based on individual rubric scores. However, they often ignored the instructions.

[8] Fundamentally, public reviews would force reviewers to spend more time to write more careful reviews. This would be great, but unless it is accompanied by systemic changes in how peer review is rewarded (which would not be quick or easy), it is likely to simply reduce participation. Public negative reviews also have repercussions for junior researchers.

self would be fantastic, but often, many things are changed at the same time, no systematic comparisons are drawn, and even the obviously successful innovations might not stay on. Consequently, next year we are no wiser about what works and what does not. We are running continuous experiments on ourselves, and never check[9] the results.

On the community side, we are not aware of any quantitative studies of how peer review is discussed on social media, but as active members of the Twitter #NLProc community, our impression is that this topic mostly gets on the feed during author rebuttals and after acceptance notifications at major conferences, as sketched in Figure 3. At that time, there are bitter complaints and reform suggestions, but few practical initiatives (which ensures that the cycle is repeated at the next conference).

Fundamentally, peer review is an annotation problem, and we can try to tackle it because we know enough about experimental methodology, iterative guideline development, inter-annotator agreement, and biases. Here is where we fail:

- *Organizers:* lack of mechanisms to test if one policy is better than another, and to ensure that successful policies are kept.
- *Authors:* lack of willingness to actively monitor policy changes[10], lack of ability to request reports on them and access the review data to conduct independent analysis.[11]
- *Reviewers:* lack of recognition for meta-research as a valid part of NLP, which, as we learned in writing this paper, makes it difficult to publish on it. In a way, NLP peer review... prevents research on NLP peer review.

To illustrate the latter point: a quick search in the ACL anthology revealed only four conference papers on peer review from a meta-research perspective: a paper-reviewer matching tool (Anjum et al., 2019), a corpus of reviews (Kang et al., 2018), and two experimental studies using NLP to explain the observed reviews (Caragea et al., 2019; Gao et al., 2019). We could not find any ACL-published papers on testing different peer review policies or review form design for the NLP community. Yet without such work nothing will change and no other field will do it for us.

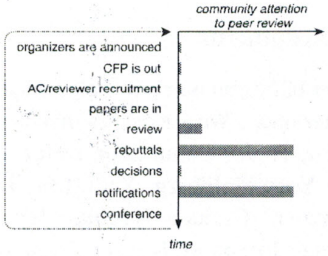

Figure 3: Attention to peer review in NLP community

The work by Gao et al. (2019) offers an actionable insight: ACL reviewers appear to be victims of conformity bias, converging to the mean of reviews. One solution would be to let reviewers interact only with the authors during the rebuttal, but not with each other. The paper was published in NAACL – yet, to the best of our knowledge, there have been no attempts to change any policies accordingly.

7 Conclusion

As a community familiar with annotation, we know that asking people to perform ill-defined tasks is not going to work well. Yet this is exactly what we expect of ourselves as reviewers. We can do better.

There are many known ways to reduce uncertainty in paper merit estimation, such as improving the review forms and reviewer matching. The problem is that *implementing* any of it would take a lot of work beside what ACL is already doing, sometimes counter to its current practices. Big changes in any large organization are difficult (especially in a volunteer-driven one), but this is the only way.

The first step towards turning all the frustration on social media into action is to (a) recognize such work as respectable, main-track-worthy meta-research (so that there are incentives to do it at all), and (b) create new, voted-in ACL roles for systematic development, testing and comparison of review policies, as well as community feedback loops for authors and reviewers. A special ACL committee is working on a rolling review reform[12] to address the increasing *volume* of reviews, but improving their *quality* is a different, long-term project.

[9] E.g. ACL 2020 opted to handle the increased reviewer load by making all authors register as reviewers, and EMNLP 2020 required a senior reviewer who would mentor secondary reviewers. How can we tell which strategy worked better and should be used next year?

[10] For instance, *Findings* was announced on the conference website and social media, but after acceptance notifications there was still confusion about what it meant.

[11] Compulsory data collection opt-in for authors and reviewers is a less radical change than making all reviews public, and it might also reduce the number of one-line reviews

[12] https://www.aclweb.org/adminwiki/index.php?title=ACL_Rolling_Review_Proposal