Resolute Resolutions (a blog), volume 1:

NLP/AI Conference "Peer" Review: A Tragedy of the (Knowledge) Commons?

July 18, 2020

Allen Schmaltz

Abstract

We review the recently proposed ACL long-term reviewing proposal, finding it offers some promising alternatives to the current system. It is a step in the right direction, but we briefly suggest the adoption of two additional practical mechanisms (short of financial compensation), which are at least as important, to encourage self-sustaining incentives in the productive direction: (1) open, non-blind review (where author *and* reviewer identities are known) and (2) structured, non-static post-publication review.

Introduction

At a coarse resolution[1], the ACL community (and the broader ML/AI community) faces the dual challenges of low quality conference reviews and a large number of conference papers that need to be reviewed, both of which (at least anecdotally) are trending in the undesirable direction.[2] Largely due to the first factor, increasingly researchers are incentivized to forgo the NLP conference system and utilize the arXiv+journal route, since conference reviews tend to be at a quality level not particularly productive for the scientific process. Relatedly, the conferences themselves tend not to be particularly useful/efficient forums for consuming new work, given the aforementioned quality issues, several month delay from first appearing on arXiv, and that the number of daily submissions to arXiv cs.CL is still (knock on wood) sufficiently low to go through daily[3]. In short, if conference reviewing quality is not good, a lot of related aspects of the system start to fall apart. We first review the new mechanisms of the proposal, which we would argue do not directly address review quality, but are nonetheless helpful for other reasons, and then propose two mechanisms to improve review quality, beyond financial compensation: (1) open, non-blind review and (2) structured, non-static post-publication review.

The aforementioned concerns do not constitute an extreme view; on the contrary, there is widespread recognition that the current system needs to be revamped (and importantly, is revampable). The recently proposed ACL Rolling Review Proposal, released on June 19, 2020[4], aims to begin to address some of these concerns via a new

[1] This blog is available at https://github.com/allenschmaltz/Resolute_Resolutions.

[2] **Note that the concerns here do not necessarily directly apply to other scientific/engineering fields**, such as medicine. Unlike the natural sciences, NLP and ML/AI have traditionally been conference—rather than journal—driven, at least in recent decades. The combination of a fairly sudden, dramatic increase in the number of papers over the last few years relative to the number of qualified reviewers, and the fast review turn-around required of conferences, creates some unique pathologies.

[3] Typical research papers in NLP have the following characteristics that make this reasonably possible/practical in one's core areas of familiarity (this is, of course, not true for certain other areas, such as medicine, where careful multi-person reviewing is necessary before taking actionable next steps): Most papers consist of no more than a handful of real new ideas (in $\leq$ 10 pages, typically), and determining whether or not an engineering idea is plausible within one's own area is relatively easy in empirical CS (i.e., "recognition" [of plausibility] is cheap). Since reviewers are not re-running/re-implementing experiments, most researchers would need to re-run/re-implement plausible ideas themselves to check/verify on their own data or tasks anyway, so the stamp of approval of the noisy $\approx$ 3 person conference review is comparatively low. Relatedly, if the reviewing quality is low, the filtering of the conference process may also not be desirable in one's own area (i.e., a researcher might as well do their own filtering based on arXiv). The costs, of course, become higher when moving into new subfields, and as the number of papers continues to grow, so **it remains critical that we not abandon reviewing altogether**.

[4] As of writing, the proposal is available at https://www.aclweb.org/adminwiki/index.php?title=ACL_Rolling_Review_Proposal. (The discussion here references the version last modified on June 19, 2020 at 16:06.)

reviewing mechanism.[5] The proposal suggests adopting a rolling, iterative process in which authors would submit papers to a reviewing pool, in which they can request one or more cycles (except in the case of desk rejects) of reviews, after which the paper would be placed into a pool for possible selection for a publication venue (the conferences, or possibly the journal *TACL*). **This is a good idea, ceteris paribus, relative to the current system.** The main advantage of this is that it shortens the time between re-reviews, which **would serve as a brute-force way around low quality reviews by reducing the costs of receiving an ill-informed review**. Additionally, we would suggest that once the review scores are complete, the full history of the reviews could be linked to the arXiv version of the paper, serving as the peer-review stamp of approval, and then the conferences can serve a more perfunctory role as presentational venues and QA sessions (and where applicable, paper tutorial walk-throughs), which makes sense given the fast pace of AI research.

Open, Non-blind Review

However, the proposed system does not directly address review quality, instead opting for an indirect work-around (more efficient re-tries). Additionally, the proposal does not address the peculiar anonymity windows used for NLP (which as far as we are aware, are not used for the conferences of any other subfield of AI). Perhaps counterintuitively, we suspect that eliminating the anonymity windows (and relatedly, modifying the *non-blind blind review* process, as discussed next) will help improve reviewer quality.

There are two types of anonymity involved in NLP's non-blind blind (or "semi-blind") review process: That of reviewer identity and that of author identity. In the idealized version of the current process, reviewers and authors are separated by an opaque wall, over which anonymity preserving messages are passed. In practice, the opaque wall is replaced with a (sometimes cloudy) semi-transparent mirror, in which the authors cannot see the reviewers, but the reviewers can more-or-less see the authors. Unfortunately, other things being equal, either the reverse, or preferably, no anonymity, would be more productive for the scientific process.

More specifically with regard to author identity, in the current process, papers often appear on arXiv before the anonymity windows begin, and in any case, most sub-communities are relatively small in NLP, so choice of methods, data sets, citations, style of writing, etc. all leak identity information. It may not be possible to pin-point exactly who the authors are for papers without a corresponding previous arXiv pre-print, but importantly, it is often within reason to guess who the authors are likely *not to be* within one's own subfield—namely, one can identity within reason papers from research groups one has never previously encountered. This is important, because compelling arguments for blind review are not couched in terms of individual-level anonymity (since direct conflicts of interest are handled by the explicit hand-labeled rules one enters into the online review systems), but rather that it is a system that aims to equalize the publication opportunities of incumbents and earlier stage researchers. However, for the aforementioned reasons, the current NLP conference review system does not preserve anonymity across those two groups to equal fair effect.[6] In fact, because NLP (and AI more generally) research moves so quickly, **arbitrary author anonymity windows greatly favor incum-**

[5] For posterity, or those from other fields, we briefly describe the current xACL conference reviewing system, which has some strange peculiarities relative to other fields. (Recall that the vast majority of published NLP papers go through the conference system.) Papers can be submitted to one of a handful of conferences throughout the year. Rather arbitrarily, there is an anonymity "window" beginning from one month *before* a given conference deadline and lasting until the conference decisions, during which a paper cannot be submitted to a pre-print server, and papers already on a pre-print server may not be modified. (A non-trivial number of papers are posted on pre-print servers before the anonymity window, including by large industry groups, effectively directly nullifying the blind review.) Around three reviewers score and provide feedback. Authors are then typically allocated a small number of sentences to reply to reviewer concerns or questions, but cannot submit (at least in principle), new results or major changes for the reviewers to consider. Reviewers then submit their final decisions and an area chair finalizes the reject or accept decision. Note that in an acceptance, authors are not obligated to make requested reviewer changes. Additionally, the reviewer and area chair identity are not known to the authors, but in typical setups, the reviewers do know the identities of the other reviewers and the area chair. Some slight variations on this theme have been experimented with. For example, in a recent conference, the author replies were dropped, but most follow the above processes. Also, note that most of the ML/AI conferences are run with different mechanisms than that of the core NLP conferences; importantly, most do not use the anonymity window. There are also a small number of NLP journals for longer works that do not use the above system.

[6] To put it another way, for fairness on that attribute, one's ability to predict membership in those two clusters would need to tend to chance.

bents, since incumbents can rely on area chairs (themselves incumbents) to push through papers that receive ill-informed reviews, and the costs of a missed stamp of approval (for a paper *not* on arXiv) by the peer review process from an arbitrary, ill-informed review is much less costly to an incumbent. Put more simply, NLP's current conference anonymity policies for authors serve no coherent, consistent purpose, which is presumably why no other subfield of AI employs such a process.

That said, while the NLP conference anonymity policies toward author identities are well-known as being a peculiar quirk, **likely more important for the scientific process is the state of reviewer identity, as it is tied to review quality**. As noted previously, reviewer identity remains hidden, and unlike author identity, does in fact remain hidden for all practical purposes. In this context, review quality has remained low, and can be attributed to at least two factors: Skilled reviewers may not put in the effort for a particular paper, and secondly, some reviewers may have insufficient background to review a particular paper. The surprisingly simple **solution is to make reviews public with reviewers' names**, no later than the final paper decision. This would encourage reviewers to take the effort to carefully read and review papers, to complete any necessary background research on a particular topic, and to not over-reach in reviewing on topics on which they are not sufficiently familiar. Importantly, the **publicly available reviews can then also serve as templates for new researchers to emulate**.

Structured, Non-static Post-publication Review

We have justified the adoption of open, non-blind review. Building out such a system would also make it possible to add one additional mechanism: Structured, non-static post-publication review. Currently, post-publication review happens, in a sense, in subsequent publications. However, if the issue is more minor, such as to highlight an additional metric, dataset issue, and/or additional work, a full subsequent publication is not needed. Such matters become institutional knowledge for members within a subfield/subarea, discussed via email, social media, or during conferences, but **there is no centralized place for researchers from other areas to quickly get up-to-speed on the latest applicable issues, particularly with regard to *important related work***. (Relatedly, further enabled by low quality reviewing, the NLP literature occasionally suffers from the cycling of stylized facts, and concomitantly, overselling.) With open review this additional mechanism is easy to implement: **Simply keep reviewing open for every paper**, to serve as a forum for post-publication review.[7]

Related Work

The title of this post is a reference to the tragedy of the commons, the common pool resource dilemma, evoking the previous work of Sculley et al. (2018). The race to the bottom with common pool resources is not inevitable, nor does it require top-down, monolithic management to be successful, as shown by the late Elinor Ostrom (Ostrom, 1990). With regard to the machine learning conferences, Sculley et al. (2018) suggests a two-pronged approach: The structured evaluation of review quality by area chairs and senior committee members, and financial compensation for reviewers. Both of these approaches could be employed in conjunction with the mechanisms proposed here. The advantages of making the final review assessments public, as for example, linking the review-quality scores to the articles and reviews in the open review system, are three-fold. First, it can have a pedagogical role for new, or existing, reviewers. Secondly, importantly, it would serve as a further incentive for *those reviewing the reviewers* to carefully fill out the rubric (i.e., provides a means to review the area chairs doing the review of reviewers). Finally, the completed rubrics would serve as additional signal for readers that subsequently come across the paper and its associated reviews.

A policy of compensating reviewers is compelling, as justified by Sculley et al. (2018). That is still possible with the mechanisms proposed here, but we suspect that a large proportion of the main NLP conferences—and certainly, the workshops—would not be sufficiently capitalized to regularly dispense funds at scale, but that is speculation and it is worth considering further. (Variations on this theme, which may still be effec-

[7] Post-publication review in a centralized system tied to the versions of a paper also obviates the need for retraction in all but the most severe cases. In fact, post-publication review is typically more helpful in the long-run than retraction, since there is then a clear paper trail of what went wrong where.

tive, are possible to reduce total costs, such as only compensating the top quality reviews from junior researchers, or perhaps even randomly assigning compensation, subject to some review quality threshold.) The hope is that instead of direct final compensation, having one's name publicly associated with a review will be sufficient incentive to do a good (i.e., effortful) job. **The review can then also become an important scientific artifact** useful for career purposes.[8] We would not go so far as to say that a review is as important as an actual new idea (and thus, paper) for assessing one's contributions to science, but obviously reviews play a critical role in the productive advance of science[9], yet currently most reviews are lost in the ether, without public record, putting further downward pressure (for many, not all) to treat them as nothing more than an afterthought.

Wasserman (2012) proposes a bolder approach: Dispense with formal peer review (as currently practiced) altogether for papers (but perhaps not grants) in fields such as statistics and computer science. For those of us whose own research discovery process revolves at least partly around arXiv rather than only conference proceedings or journals (See Footnote 3), there is an inherent appeal to such an approach of gong all-in on arXiv. However, at least three dangers come to mind in taking such a radical approach. First (and perhaps foremost), the separation between NLP/AI proper and high-risk fields, such as medicine, is increasingly blurry, so at the very least, a separate, parallel system of formal reviewing would need to be established for a non-trivial number of papers in order to provide a countervailing force to assess quality, particularly with regard to fine-grained, domain-specific details. However, then we are back to addressing many of the same issues with review quality and the organization of the reviewing process. Secondly, setting aside high-risk areas, a total free-for-all makes barriers to entry harder—perhaps *much* harder—for non-incumbents to overcome, particularly those not backed by a PR machine (as with membership in a large industry group or a large university), with knock-on effects for career progression. With a formal reviewing system, at least one person (at least casually) reads and reviews every submitted paper. Finally, although we suggested earlier that the filtering yielded by the current conference system is likely flawed, and should be supplemented by daily reading of arXiv, even less effective, noisier filtering systems are possible—and in fact, are in use today. In particular, the one-off **filtering afforded by social media is almost certainly far worse (in terms of bias and coverage)** than reading the daily arXiv email, since social media heavily censors the tails of the distribution of papers (as pre-prints, but also among those that have gone through peer review), and also worse than only reading the conference proceedings. The no-peer-review approach might amplify this, resulting in a relatively peaked distribution of papers that are amenable to hype, which is a property poorly correlated with scientific value.

Conclusion

In this blog post, we have reviewed the ACL long-term reviewing proposal, which was recently put forth in light of widespread concerns that the current ACL conference reviewing system needs to be revamped. The proposed system would be an improvement over the current system in that it would potentially shorten the time between a review cycle and a re-review cycle, serving as a brute-force solution to low review quality. We have additionally proposed that this system should be combined with open review and a post-publication review system.

The issues here are complicated, and no one has a monopoly on the right solutions. Probably any changes will result in at least one cycle of conferences that require much more effort than normal (on the part of reviewers, authors, and conference organizers), but given the current state of affairs, we think that short-term cost would be well-worth paying.

[8] Actually, some interesting, and perhaps surprising, patterns may emerge from having reviews publicly available. Since reviews are typically completed at an individual level, occur with relatively high frequency, and often involve topics slightly outside a researcher's current area of focus, the distribution of informedness may be rather different than when looking at publications and positions alone, including among senior researchers.

[9] They may also be interesting artifacts for the History of Science. In the course of writing this, we found that the Royal Society has digitized some referee reports for classic papers. For example, here is J. B. S. Haldane's referee report for Turing's 1952 paper, "The Chemical Basis of Morphogenesis", in which he notes in the first sentence in his review in no small terms: "Before the paper is accepted, I consider that the whole non-mathematical part should be re-written.": https://makingscience.royalsociety.org/s/rs/items/RR_1950_51_B62_2/01c348.

References

Elinor Ostrom. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Political Economy of Institutions and Decisions. Cambridge University Press.

D Sculley, Jasper Snoek, and Alex Wiltschko. 2018. Avoiding a Tragedy of the Commons in the Peer Review Process.

Larry Wasserman. A World Without Referees [online]. 2012.