Last Words

Improving Our Reviewing Processes

Inderjeet Mani*
The MITRE Corporation

1. Introduction

Our reviewing practices today are failing. With the number of ACL submissions steadily growing over the last several years (for example, ACL 2009 had a 24% increase in submissions over ACL 2008), the need for more reviewers has become more pronounced. However, qualified reviewers are becoming hard to find, and when they are found, they are often hard-pressed for time. As a result, slipshod reviews are becoming commonplace. Allowing this situation to continue as before will result in the deterioration of our ability to recognize excellence in our research. An 'intervention' is therefore needed.

As I see it, there are two distinct problems to tackle: first, a lack of qualified reviewers, and second, a lack of quality control in reviews. After discussing these, I will suggest some solutions that I believe are worth implementing.

2. Problems with Reviewing

2.1 The Lack of Qualified Reviewers

In an earlier *Last Words* piece, Ken Church (Church 2006) pointed out how the ACL conference reviewing process can be derailed by the lack of positive endorsement by reviewers who are not well qualified to review a given paper. He went on to suggest that papers rejected by NAACL are "often strong contenders for the best-paper award at ACL." An instance of this phenomenon was observed in 2009, when a paper rejected from NAACL 2009 with an average acceptance score of 2.3 out of 5 was given a best paper award at ACL 2009 (Branavan et al. 2009).[1]

It is especially hard to find qualified reviewers these days partly because computational linguistics has become increasingly specialized. Papers in fields like parsing and machine translation involve very technical modifications to a few current models. Reviewers for such areas need to be 'insiders', well-versed in the latest developments in the sub-area. This need is likely to become more pronounced as the specialization trend continues.

Reviewers are currently selected based on informal social networks. Unfortunately, most researchers do not have an extensive set of names of reviewers at hand, and relying

* The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730, USA. E-mail: imani@mitre.org.

1 Such discrepancies in judgments across conferences are not confined to computational linguistics; for example, the classic Page Rank paper from WWW 1998 (Brin and Page 1998) had previously been rejected by SIGIR 1998.

on personal connections (not to mention memories of reviewers' prior performance) is limiting and could bias the selection of reviewers to those who share a particular point-of-view. This lack of information as to whom to contact can result in woefully inappropriate selections of reviewers.

2.2 The Lack of Quality Control

Even when qualified reviewers can be found, reviews are often hurried. At the 2009 ACL Business Meeting, Ido Dagan pointed to the growing dissatisfaction with the quality of conference reviewing, adding that the problems seemed to be exacerbated by increasing the number of reviewers. The lack of quality is in part due to the large number of conferences that compete for the reviewer's time. As Fortnow (2009) observes, in the case of computer science conferences the intensive time commitment required for reviewing makes it less likely that more experienced researchers will sign on as reviewers. Such hurried reviewing and decision-making can result in a preference for safe, more incremental papers rather than those that develop new models and research directions (Fortnow 2009).

3. Finding Qualified Reviewers: An ACL Reviewer Database

To encourage better selection of reviewers, improved information management is needed in terms of keeping track of reviewer background and areas of expertise. This goal can be achieved by maintaining an *ACL database of reviewer profiles* that is accessible to the public, as a resource for use in selecting reviewers. It would work as follows: Whenever reviews are produced in an ACL-related forum, the Program Chair or Editor would be responsible for updating the database (assisted in part by the conference or journal management software). S/he would record the reviewer's name and affiliation, the name and type of forum (conference, workshop, journal, etc.), its time and place, the number of papers reviewed, and the reviewer's areas of interest. To protect a reviewer's privacy, information as to which papers were assigned to the reviewer should not be revealed. Note that most conferences already publish a list of reviewers involved, so this is not a huge step beyond what is there today.

Once created, the ACL Reviewer Database would provide for easier selection of reviewers, and would allow a forum organizer or editor to determine, at least semi-automatically, categories of reviewers such as specialist/generalist, prolific/occasional, and skill level (e.g., journal paper vs. workshop reviewing, senior vs. extended program committee experience).

Many reviewers work extremely hard at their reviews. Being recognized as a well-respected reviewer is well worth striving for, and ideally, this reputation would be reflected in part by one's reviewing profile. Just as the impacts of journals and authors are measured based on citations, there is no reason why reviewers should not be assessed in terms of their impacts in guiding and encouraging computational linguistics. Specifically, the database can be used to track how many journal and conference articles a person has reviewed, and how many articles were submitted to those venues. A reviewer's impact factor can be computed by dividing the former by the latter, and including some normalization parameters (such as taking into account the level of publishing activity in the particular computational linguistics subfield).

4. Encouraging High-Quality Reviews

4.1 Open Peer Review: Signing and Publication History

Reviewers often get away with slipshod, low-quality reviews because they can hide under the cloak of anonymity. Other research communities have reduced the level of anonymity by using *open peer review*. In an open peer-review system, reviewers may sometimes become known to the paper authors, and in some instances, reviewers and the content of their reviews may become known to all the readers as well. Journals that have successfully deployed open peer-review include *Atmospheric Chemistry and Physics (ACP)*,[2] the *British Medical Journal (BMJ)*,[3] and all forty-one of the *Biomed Central (BMC)* medical journals.[4]

Open peer-review systems can differ in terms of reviewer transparency: Some venues (e.g., the *BMJ*, and *BMC* medical journals such as *BMC Cancer*) always require that reviews be *signed*, that is, visible to the author, whereas others do not (e.g., the ACP journal), or else they leave it up to the reviewer. Signed reviews can considerably raise the stakes on review quality: It is one thing to palm off a hurried review on a hapless author, quite another to be accountable in front of all one's colleagues for the poor quality of one's review. Although it is theoretically possible that signed reviews might be less frank, in order to avoid alienating particular authors, such a problem tends not to occur in practice. In the case of a randomized trial with medical articles in the *BMJ*, signing did not lower the review quality or recommendations (van Rooyen et al. 1999). Nor has the presence of signed reviews in journals that have used them led to a rise in litigation against those journals. Finally, many journals using signed reviews, such as *BMC Cancer*, continue to thrive and flourish.

Whereas the *BMJ* publishes only the final version of the paper, both the *BMC* medical journals and the *ACP* journal provide public access to the *publication history*, namely, all previous versions of the paper along with their reviews and author responses.[5] A view of such a publication history can be extremely instructive to both prospective authors and reviewers. A further advantage of publication history is that prior submissions within the community can be tracked, providing a collective memory of reviewers' comments. That is far better than the situation today, where regular reviewers can often recall reviewing some version of a paper earlier for some other forum, but may not be able to recall the individual recommendations.

4.2 Reviewer Training

Reviewing is one of the most important activities a researcher carries out, and yet no formal training is provided to reviewers. If we require high-quality reviews, we need to train reviewers as to the best practices in reviewing as carried out by a particular ACL-related forum. This could be organized as an on-line course specific to the journal or conference, which every reviewer for that forum should take. The course, even if

2 http://www.atmos-chem-phys.net/volumes_and_issues.html.
3 http://www.bmj.com/.
4 http://www.biomedcentral.com/.
5 For an example of an accepted paper with signed reviews and author comments, see www.biomedcentral.com/1471-2407/9/348/prepub. For an interesting discussion around a rejected paper, see the following ACP paper: http://preview.tinyurl.com/ycxq65e.

involving only self-study, can be viewed as a more stringent requirement than simply being encouraged to read, as is customary today, the reviewing guidelines. Here, too, the conference or journal management software can check that the course has been completed within the last couple of years before allowing the reviewer to proceed. The course can be updated from time to time.

As an example, journals like the *BMJ* offer materials for training reviewers.[6] These materials include PowerPoint presentations on reviewing best practices ("What we know about peer review," "What editors want") and written exercises (reading and assessing three referee reports for a paper, and comparing the assessments with the editor's critique of those reports, as well as doing a practice review of a paper and comparing it with the published reviews). As shown in a *BMJ* study of the effectiveness of these materials (Schroter et al. 2004), trained reviewers detected significantly more major errors in papers than those who weren't trained.

For computational linguistics journals and conferences, it would be straightforward, under either an open peer review system and/or with permission of authors and reviewers, to collect published reviews of several articles and have potential reviewers go through a similar exercise to that provided by the *BMJ*. In particular, one could focus on practice reviews for a paper. In addition to such on-line courses, improved review quality can be engendered by discussion of reviewing methods in classroom settings, particularly in seminar courses where papers have to be read and jointly discussed.

5. Conclusions

To discover qualified reviewers, I have suggested creating an ACL Reviewer Database. To improve review quality, I have advocated more use of open peer review, with review signing and/or maintaining a history of the publication of each article, as well as specific measures for improved reviewer training. These methods for improving review quality have become common practice in some other fields, and it is high time computational linguists started to explore them.

These improvements will require a higher degree of transparency than has been customary in computational linguistics reviewing, but this transparency will reap considerable benefits in further recognizing and promoting excellence in our research. It would therefore be worthwhile to initiate a discussion group or a workshop to plan a pilot that will try out some of these methods. Over time, it will also be useful to conduct studies of how effective these methods are.

Acknowledgments
This research has been funded by the MITRE Innovation Program (Public Release Case Number 07-0862). I am grateful to Robert Dale for his in-depth comments.

References

Branavan, S. R. K., Harr Chen, Luke Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 82–90, Singapore, August 2–7.

Brin, Sergey and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Proceedings of the Seventh International Conference on the*

6 http://resources.bmj.com/bmj/reviewers/training-materials.

World Wide Web, pages 107–117, Brisbane, Australia, April 14–18.

Church, Kenneth. 2006. Reviewing the reviewers. *Computational Linguistics*, 31(4):575–578.

Fortnow, Lance. 2009. Viewpoint: Time for computer science to grow up. *Communications of the ACM*, 52(8):33–35.

Schroter, Sara, Nick Black, Stephen Evans, James Carpenter, Fiona Godlee, and Richard Smith. 2004. Effects of training on quality of peer review: Randomised controlled trial. *British Medical Journal*, 328:673.

van Rooyen, Susan, Fiona Godlee, Stephen Evans, Nick Black, and Richard Smith. 1999. Effect of open peer review on quality of reviews and on reviewers' recommendations: A randomised trial. *British Medical Journal*, 318:23–27.