Last Words

Reviewing the Reviewers

Kenneth Church
Microsoft Corporation

Recall is More Subtle than Precision

I just returned from the Association for Computational Linguistics' 43rd Annual Meeting (ACL-2005). The acceptance rate was 18%. Is this a good thing or a bad thing?

When the acceptance rate is low, precision tends to be high. The audience can judge precision for itself. If the presentations are good, everyone knows it. And if they aren't, they know that as well. ACL-2005 had great precision.

Recall is more subtle. When there is an issue with recall, it isn't immediately obvious to everyone. If you listen closely, you'll hear some grumbling in the halls. And then the rejects start to appear elsewhere. ACL's low recall has been great for other conferences. The best of the rejects are very good, better than most of the accepted papers, and often strong contenders for the best paper award at EMNLP. I used to be surprised by the quality of these rejects, but after seeing so many great rejects over so many years, I am no longer surprised by anything. The practice of setting EMNLP's submission date immediately after ACL's notification date is a not-so-subtle hint: Please do something about the low recall.

When you read some of the ACL reviews for these top EMNLP papers, you realize what is happening. ACL reviewing is paying too much attention to abstentions (and objections from people outside the area). If a reviewer isn't qualified to say anything on a particular topic, that's okay. An abstention shouldn't kill a paper.

Controversial papers are great; boring unobjectionable incremental papers are not. The only bad paper is a paper without an advocate. A paper with a single advocate should trump a paper with lots of seconds, but no advocates. Don't average votes. The key votes are the advocates. Negative votes matter only if they convince the advocates to change their votes.

Recall is a problem for many conferences, not just ACL; SIGIR, for example, rejected the classic paper on page rank, a hugely successful paper in terms of citations, perhaps more successful than anything SIGIR ever published.

1. A Model

Consider the following model. Suppose we generate the gold standard so that some fraction, $a = 20\%$, of the $s = 400$ submitted papers should be accepted, and $1 - a$, should be rejected. There are $r = 3$ reviews for each paper. A review votes either 1 (accept) or 0 (reject). Papers are scored by summing the votes. Some reviews are good ($g = 70\%$), and some ($1 - g$) are not. A good review votes the same way as the gold standard. A bad review votes randomly, accepting a of the papers and rejecting the rest. The program committee as a whole is evaluated in terms of precision and recall. That is, if they accept a papers with the best votes, how well do those papers match the gold standard?

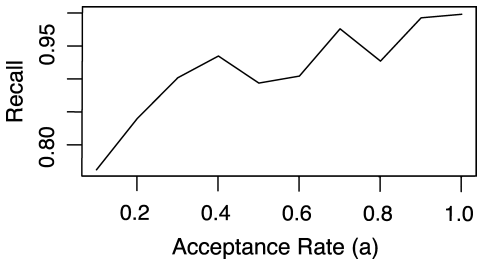

Figure 1
Accepting more papers is an easy way to improve recall. Results vary slightly from run to run because of the non-determinism in the jitter function.

In R (www.r-project.org) notation, we can express this model as

```
a=0.2  # acceptance rate
s=400  # submissions
r=3    # reviews per paper
g=0.7  # mixture of good to random reviews
gold = rbinom(s, 1, a)
good = matrix(rbinom(s*r, 1, g), ncol=r)
score = good*gold +
   (1-good)*matrix(rbinom(s*r,1,a),ncol=r)
accept = rank(jitter(apply(score,1,sum)))>(s*(1-a))
precision = sum(accept * gold)/sum(accept)
recall = sum(accept * gold)/sum(gold)
```

According to this model, recall can be improved in at least three ways:

- Plan A: Increase a (acceptance rate)
- Plan B: Increase r (reviews per paper)
- Plan C: Increase g (mixture of good to random reviews)

We ought to do all of the above, as much as possible. Increasing the acceptance rate is easy. There is no excuse not to. Last century, we kept the acceptance rate low so everyone could hear every paper in a single plenary session. Given the ever increasing submission rates, and other obvious practical modern realities, the old debate over plenary sessions has long since been forgotten. Nevertheless, we still hear a somewhat similar argument, that we can't accept more papers because of some (imagined) constraint involving local arrangements. In fact, ACL-2005 could have accepted more papers than it did; there were empty rooms during much of the meeting. Moreover, local arrangements have obvious financial incentives to come up with creative ways to accept as many papers as possible: more papers → more $$ (conference registrations). If we can accept more papers without hurting (real or perceived) precision too much, we ought to do so.

In any case, acceptance rates should never be allowed to fall below 20%. Even though I can't use the model above to justify the magic threshold of 20%, it has been my experience that whenever acceptance rates fall below that magic threshold, it be-

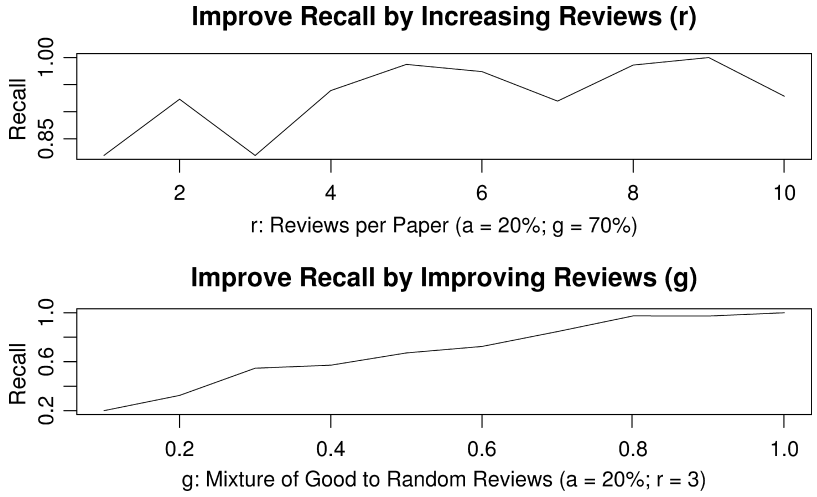

Figure 2
In addition to increasing *a* (acceptance rate), recall can also be improved by increasing *r* (reviews per paper) and *g* (mixture of good to random reviews).

comes too obvious to too many people just how low the recall is. Magic thresholds like 20% change the tone of the grumbing in the halls into an ugly swap meet. Many of the leaders of the field were exchanging tales at ACL-2005 about inappropriate rejections. Pleasant fairy tales like the reviewing-is-perfect myth make people comfortable. It is hard to believe in fairy tales when too many people know the facts. And it's hard to maintain plausible deniability when everyone knows what everyone knows.

Plan B and Plan C are probably unrealistic. Whenever we have extra reviewers, we ought to use them to increase recall by increasing the number of reviews per paper, but realistically, as submissions go up and up and up, we're unlikely to have lots of spare reviewers. However, if the community fails to accept Plan A (increasing acceptance rate), then we could retaliate with Plan B as a deterrent. It the community won't let us to do the right thing (increase acceptance rates), then we can punish them with more and more papers to review until they "appreciate" the merits of Plan A.

As for Plan C, of course, life would be good if reviewers were better than they are. Unfortunately, reviewers do what reviewers do. Some reviewers are conservative and some are really conservative and some are really really conservative. Reviewers love safe (boring) papers, ideally on a topic that has been discussed before (ad nauseam). Precedents are good; novelty is bad. The process discourages growth (contributions from new blood on new topics). But the survival of the organization depends on new blood. We need to liberalize the process, or else.

Meta-reviewers and area chairs should do as much as they can, but it will never be enough. It is up to the meta-reviewers to make sure that a paper with an advocate trumps a paper with three seconds. Controversy is good; boring is bad. The meta-reviewers should maintain a reserve of designated tie-breaker reviewers. The reserve should be highly respected (and highly opinionated). They need to work quickly and decisively, without too many abstentions, selecting interesting novel papers, and avoiding the safe boring unobjectionable papers that tend to do relatively well on the first round with the first tier of reviewers.

2. Recommendations

Whatever you measure, you get. Precision is a good thing, but so is recall. In addition to reporting submissions, and acceptance rates, conferences ought to report estimates of precision and recall. It should be possible to estimate precision and recall using a cross-validation argument. We could give a sample of the papers to another set of reviewers and use their decisions as a gold standard for evaluating the program committee as a whole in terms of precision and recall.

We could also track citations to the papers that we accept, as well as rejects that are published elsewhere; we can hope that the accepts will be more heavily cited than the rejects.

The easiest way to improve recall is Plan A: increase acceptance rates. If the community won't go for that, we can try Plan B (increase the number of reviews per submission) for as long as it takes until they appreciate the merits of Plan A. Meanwhile, the community is adopting a Plan D, a hybrid between Plan A (increase acceptance rates) and Plan B (increase the reviewing burden). Plan D is truly Machiavellian: more conferences.