Natural Language Engineering **23** (3): 473–480.

doi:10.1017/S1351324917000067

Emerging trends: I did it, I did it, I did it, but...

KENNETH WARD CHURCH

IBM, Yorktown Heights, NY, USA
e-mail: kwchurch@us.ibm.com

(*Received 16 January 2017; accepted 26 January 2017*)

Abstract

There has been a trend for publications to report better and better numbers, but less and less insight. The literature is turning into a giant leaderboard, where publication depends on numbers and little else (such as insight and explanation). It is considered a feature that machine learning has become so powerful (and so opaque) that it is no longer necessary (or even relevant) to talk about how it works. Insight is not only not required any more, but perhaps, insight is no longer even considered desirable.

Transparency is good and opacity is bad. A recent best seller, *Weapons of Math Destruction*, is concerned that big data (and WMDs) increase inequality and threaten democracy largely because of opacity. Algorithms are being used to make lots of important decisions like who gets a loan and who goes to jail. If we tell the machine to maximize an objective function like making money, it will do exactly that, for better and for worse. Who is responsible for the consequences? Does it make it ok for machines to do bad things if no one knows what's happening and why, including those of us who created the machines?

1 Papers are reporting better numbers, but...

There has been a trend for publications to report better and better numbers, but less and less insight. Years ago, someone from an industrial lab presented a talk at a conference saying basically *I did it, I did it, I did it, but I'll be damned if I'll tell you how!* (because his employer wouldn't allow him to tell us what we really wanted to hear). Since I also worked for an industrial lab at the time, I had a strong allergic reaction to this talk. I was worried that my employer might ask me to publish similar papers so they could take credit for my results while protecting their intellectual property as trade secret. Since then, I have often argued that we need to reject papers that try to pull this kind of stunt. The better the numbers are, the more important it is to reject the paper. We can't afford papers that report results without insights.

With the rise of neural nets, we are now seeing a new variation on this theme. We are seeing lots of papers, these days, that say, *I did it, I did it, I did it, but I don't know how!* Now even the author of the paper doesn't know how the machine does what it does. It is considered a feature that machine learning has become so powerful (and so opaque) that it is no longer necessary (or even relevant) to talk about how it works. Insight is not only not required any more, but perhaps, insight is no longer even considered desirable.

In my experience, good work tends to be more accessible, and bad work tends to be less accessible. Thus, if we don't know how it works, it probably isn't any good. Transparency is good and opacity is bad.

Cliches often have at least a grain of truth to them: *sunlight is the best disinfectant*, and conversely, by another cliche in government circles, if the work isn't any good, *classify it!*

2 Can a paper exist without an audience?

If the machine gets the job done, and no one knows how it does what it does, is that a good thing, or a bad thing, or just the new world order? We might find ourselves actually living in a science fiction world where future generations have become totally dependent on machines built by previous generations, and there is no one left alive that knows how the machines work, or what they do, or merely, how to perform basic maintenance. There have been science fiction stories where even the very existence of the machines has been a matter of some dispute.[1]

IMHO, a paper should be like a tree falling in a forest. If no one hears it, did it make a sound? So too, if there is a paper in the literature, and no reader remembers reading it, did it make an impact? Can a paper exist without an audience?

Numbers by themselves are unlikely to mean much to most readers unless the readers see some insights that they can use in their own work. If the numbers can't be replicated, they are unlikely to stand up to the test of time.

3 Weapons of math destruction (WMD)

A recent best seller, Weapons of Math Destruction (O'Neil 2016), is concerned about a somewhat similar dystopia, but this one may not be so far off into the future. O'Neil argues that big data (and WMDs) increase inequality and threaten democracy largely because of opacity. O'Neil is concerned that algorithms are being used to make lots of important decisions like who gets a loan and who goes to jail. According to the New York Times,[2]

> Numbers offer the sheen of objectivity; algorithms seem to 'transcend morality', as O'Neil put it, when in fact they only obfuscate the human assumptions that go into creating them... [T]he models include such inputs as whether an inmate's friends and relatives have criminal records, and whether an inmate resides in a 'high-risk' neighborhood.

If we build a machine to maximize a simple metric (like money), and we don't know what it does and why, are we responsible for the consequences? It could well be very profitable to violate various laws (and ethics): guilt by association, profiling,

[1] `https://en.wikipedia.org/wiki/For_the_World_Is_Hollow_and_I_Have_Touched_the_Sky`

[2] `https://www.nytimes.com/2016/10/16/books/review/the-story-behind-this-weeks-best-sellers.html`

https://doi.org/10.1017/S1351324917000067 Published online by Cambridge University Press

discrimination, equal opportunity, redlining,[3] conflicts of interest, etc. If we aren't responsible for the machines we build, who is? The regulators? It is hard to imagine how regulation could be effective if even the experts have no idea what's happening and why.

4 Magic and hard work

I once heard an invited talk by a big name in neural nets suggesting that neural net research was easy. In particular, he argued that it was no longer necessary to think about degrees of freedom. In the bad old days, we used to worry about feature selection. It used to be considered necessary (and desirable) to have more observations than parameters, but these days, the speaker suggested, it is no longer necessary to worry about such details with modern neural nets. Just take your data and toss it into a few standard deep/wide configurations, and if the numbers look promising and top the leaderboard, then poof, just like magic, you have a publication. What could be easier than that?

The discussion in the halls wasn't completely drinking the Kool-Aid. How could machine learning work if there really are more degrees of freedom than observations? Perhaps there aren't as many degrees of freedom as it appears. Perhaps many of the parameters are tied together in ways that are not completely understood, or perhaps the optimization is suboptimal in ways that reduce the risk of overfitting. Modern optimizations are so complicated that it is hard to address traditional questions like degrees of freedom, significance of each parameter and ANOVA.[4] Such questions were well-understood for simple optimization methods such as regression, but the literature has less to say about such questions for more modern optimization methods, though there are a few suggestions such as LeCun (1989).

In any case, everyone knows that research is hard work. Neural nets are great for many tasks, but they haven't yet automated researchers out of a job. Research is harder than just pushing a button and waiting for the optimization to converge on a publishable result. The best results often come from the best labs, suggesting that there must be some insights that separate the best from the rest. The audience ought to expect the best to publish more of these insights in addition to promising numbers.

5 An explanation of my interest in explanation

Explanation played an important role in my decision to work on natural language processing. I didn't start out working in natural language processing. In fact, I was originally part of a medical decision making group in the 1970s, working on expert systems to advise doctors. The doctors made it clear to us that explanation was not merely desirable, but required. There would be no partial credit for advice without explanation. Doctors won't even listen to advice unless it is backed up with explanations that work for them.

[3] https://www.washingtonpost.com/news/wonk/wp/2015/05/28/evidence-that-banks-still-deny-black-borrowers-just-as-they-did-50-years-ago

[4] https://en.wikipedia.org/wiki/Analysis_of_variance

https://doi.org/10.1017/S1351324917000067 Published online by Cambridge University Press

The doctors weren't unreasonable. I remember my elementary school teacher insisting that we show our work. It wasn't good enough to merely get the right answer. We wouldn't get full credit on the quiz unless our answers were backed up with sensible explanations.

What counted as 'explanation' in the 1970s was basically a trace of the stack. Such explanations could be difficult for the audience to follow: e.g., the program started working on x, which caused it to start working on y, which caused it to start working on z. Eventually, it came back with the answer to z, which helped it come back with the answer to y, which helped it come back with the answer to x.

$$x \rightarrow y \rightarrow z \rightarrow \ldots \rightarrow z \rightarrow y \rightarrow x \qquad (1)$$

This kind of stack-nested logic may be perfectly transparent to a computer (or perhaps a computer scientist), but it doesn't work for doctors (and most normal human beings).

This observation caused me to start thinking about the difference between stacks and what we tend to see in natural language, where thoughts tend to flow in a consistent direction from x to y to z, but not back again from z to y to x. This caused me to start working on the differences between stacks and finite state machines, which should have eventually helped me come back with the answer to the original task (explanations for doctors), but apparently, stack nesting is not only a poor model of natural language thought, but also a poor model of careers. We rarely get back to where we started. It is possible that the Big Bang will be followed by a Big Crunch, but not likely.

6 Leaderboards, evaluation and reviewing

I worry that the literature may be turning into a giant leaderboard. As reviewing burdens continue to become more and more onerous, reviewers are looking for easier and easier ways to discharge responsibility. Papers are being rejected for silly reasons like typos, and papers are being accepted for equally silly reasons like topping a leaderboard.

Leaderboards are great, but a paper should do more than merely top a leaderboard. Leaderboards provide a useful service by helping the audience figure out how the proposed solution stacks up to the competition, but that should be just a starting point to motivate a more interesting discussion on why the proposed solution works as well as it does. Such an explanation ought to call out some novel insights that distinguish the proposed solution from the competition.

Similar comments apply to evaluation sections of papers. Reviewers are expecting authors to make reviewing easy. Authors are expected to write a section titled 'Evaluation'. This section should feature a table of numbers. One of the numbers should be highlighted in **bold font**. The caption under the table should assert that the bold number is better than some baseline number (that isn't in bold font). This convention speeds up reviewing. The reviewer knows to skip to the evaluation section, skim the table of numbers and focus on the bold font. Papers that obey this convention are more likely to be accepted than papers that don't.

https://doi.org/10.1017/S1351324917000067 Published online by Cambridge University Press

One could imagine a future possible world where reviewers have little incentive to read the rest of the paper, and authors have little incentive to write the rest of the paper. How hard would it be to automate this type of ‘reviewing’? Machines can probably detect who is at the top of the leaderboard. Actually, reviewers aren’t that good at that, since leaderboards can be a moving target, and reviewers can’t be expected to track changes as often as machines can. Automation is already being used more and more to detect violations of rules such as dual submissions and plagiarism. As reviewing becomes more and more about numbers and rules, one could imagine a day when reviewing is fully automated, leaving little room for subjective considerations (and insights).

7 Responsibilities and motivations

Whose responsibility is it that the literature be insightful? Reviewers? Authors? Venues? The audience?

We can’t blame reviewers for doing what they do. Reviewing is, at best, a thankless chore. Some reviewers take their responsibility more seriously than others, but even the best reviewers can only do so much. Mistakes happen all the time (in both directions). Good papers are rejected and bad papers are accepted. If a paper is submitted to enough different venues, odds of publication are pretty good even for not-so-good papers.

I am not opposed to reviewing, but I see reviewing like government regulation. It is certainly well intentioned, but probably ineffective. That doesn’t mean that we should do away with it (without evidence demonstrating that the costs outweigh the benefits).

We could try to blame authors for writing bad papers, and submitting them (perhaps again and again), but given pressures on authors to publish or perish, it is hard to blame authors too much for their desire to avoid the consequences of not publishing.

Cheating is rarely caught, though there are a few highly publicized exceptions.[5] Perhaps I am naive, but I prefer to believe (whether true or not) that cheating is rarely caught because there is so little to catch. In any case, there are lots of standard tricks that we have all seen way too often like weak baselines, mindless metrics, lack of transparency, etc.

Whatever you measure, you get. If authors are publishing merely to advance their careers, then the work isn’t going to be all that great. The work tends to be better if authors are advocating positions that they care passionately about for reasons that go beyond personal gain.

8 Arxiv.org publishes good work (without reviewing)

The red-faced test is more effective than the reviewing process. Most researchers won’t publish a paper that they would be embarrassed to show to their

[5] https://bits.blogs.nytimes.com/2015/06/11/baidu-fires-researcher-tied-to-contest-disqualification/

https://doi.org/10.1017/S1351324917000067 Published online by Cambridge University Press

colleagues. I think this is largely the reason why there is so much good work on `https://arxiv.org`, despite the lack of reviewing (or perhaps because of the lack of reviewing).

A combination of an open archive with an open citation index such Google Scholar could disrupt traditional academic publishing and pay walls. According to one leaderboard,[6] `https://arxiv.org` is a top venue in our field, well ahead of most journals including *CL*, *TACL* and this journal. Journals rank higher in other fields like speech (signal processing).[7]

Of course, the leaderboard doesn't explain why the numbers come out the way that they do. I might speculate that the leaderboard is reflecting something real and important, like what the audience cares about and what they don't care about. The citation data suggests that journals behind pay walls are at a significant disadvantage, but even journals without paywalls may not be able to compete with top conferences or even the archives. The citation data suggests that a little conference reviewing may be ok, but excessive journal reviewing is not. Conference schedules impose reasonable limits on reviewing delays, but journal reviewing can go on forever. Bowyer[8] offers some excellent mentoring advice, encouraging junior faculty to ignore flames about conferences verse journals (such as this), and publish in great venues, including both conferences and journals.

9 Metrics: Hindex, impact factors and acceptance rates

Admittedly, there are lots of problems with leaderboards including this one. Maybe this leaderboard isn't measuring the right metric; different metrics produce different rankings. Hindex counts quality (papers with lots of citations), with no penalty for quantity, unlike impact factor (citations per publication). One might expect a strong correlation between acceptance rates and impact factors, but the correlation is surprisingly small (Freyne 2010).

Should there be a link between acceptance rates and citations? In Church (2005), I suggested that reviewing should be a leading indicator of future citations, and criticized SIGIR for rejecting (Page 1999),[9] a paper with 10k citations, enough to improve SIGIR's impact factor. Prominent members of the SIGIR community (personal communication) responded by defending the rejection of the Page Rank paper (for inadequate evaluation), and rejecting the link between acceptances and citations.

What counts as a citation? Different indexes produce different rankings. Freyne *et al.* (2010) compare Google Scholar and ISI Web of Knowledge. Google Scholar includes more or less everything their crawler can find, unlike the more selective Web of Knowledge.

[6] `https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics`

[7] `https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_signalprocessing`

[8] `https://www3.nd.edu/ kwb/Mentoring_Conferences_Journals.pdf`

[9] `http://ilpubs.stanford.edu:8090/422/`

https://doi.org/10.1017/S1351324917000067 Published online by Cambridge University Press

One might expect the selection process to add value, but I remain unconvinced (Church 2014). Among other problems, the selection process suffers from an appearance of a conflict, where certain for-profit indexing services may be favoring for-profit publishers. In fact, it can be costly for a new journal to convince commercial indexing services to pick them up. One of the selling points that commercial publishers use when negotiating for a new journal is their ability to out-spend non-profit publishers to make sure that the new journal is picked up quickly by as many indexing services as possible. As a result, commercial indexes include a number of rather obscure journals, while excluding some top conferences sponsored by highly respected non-profits such as ACL, ACM and IEEE.

Freyne *et al.* find reasonable agreement between Google Scholar and Web of Knowledge, at least for top journals (see their Figure 3), but obviously, there are strong disagreements for conferences and other venues excluded by the selection process. They conclude the following:

(1) The traditional CS emphasis on publishing in conference proceedings could hurt CS researchers in evaluations based on on the ISI Web of Knowledge.
(2) Leading conference proceedings compare favorably to mid-ranking journals, surpassing journals in the bottom half of the traditional ISI Web of Knowledge impact ranking.
(3) The commonly held view that conference rejection rates are a good proxy for conference quality did not hold up to scrutiny in this study.

10 Concluding remarks: Audience should assume responsibility

IMHO, competition works remarkably well (eventually). The community will converge on better estimates of better metrics. And better estimates of better metrics will be used for promotion decisions, at some point in the future. In the short term, though, junior faculty should probably hedge their bets, and publish in top venues as well as venues that are picked up by most indexing services.

Venues will have to compete with one another, and if journals find their reviewing processes make them uncompetitive with conferences and even the archives, market forces will force them to make appropriate adjustments to remain competitive. Authors will mostly do the right thing, partly to advance their careers, but more to pass the red-faced test (peer pressure), and for their own satisfaction.

At the end of the day, the ultimate satisfaction comes from meeting (and exceeding) audience expectations. That is unlikely to happen if we have no idea what we are doing, and why we do what we do. The audience must demand more than merely good numbers. It is the responsibility of the audience to expect good work (both numbers as well as insights), and vote early and vote often with citations.

References

Church, K. 2005. Reviewing the reviewers. *Computational Linguistics* **31**(4): 575–578.

Church, K. 2014. TALIP perspectives, guest editorial commentary: what counts (and what ought to count)? *ACM Transactions on Asian Language Information Processing (TALIP)* **13**(1): 5:1–5:5.

https://doi.org/10.1017/S1351324917000067 Published online by Cambridge University Press

Freyne, J., Coyle, L., Smyth, B., and Cunningham, P. 2010. Relative status of journal and conference publications in computer science. *CACM* **53**(11): 124–132.

LeCun, Y., Denker, J. S., Solla, S. A., Howard, R. E., and Jackel, L. D. 1989. Optimal brain damage. *NIPS* **2**: 598–605.

O'Neil, C. 2016. *Weapons of Math Destruction*. Crown, New York: Crown Publishing Group.

Page, L., Brin, S., Motwani, R., and Winograd, T. 1999. The PageRank citation ranking: bringing order to the web. Technical Report, Stanford InfoLab.

https://doi.org/10.1017/S1351324917000067 Published online by Cambridge University Press