Natural Language Engineering (2022), **28**, pp. 249–269
doi:10.1017/S1351324922000043

CAMBRIDGE UNIVERSITY PRESS

EMERGING TRENDS

Emerging Trends: SOTA-Chasing

Kenneth Ward Church[1,*] and Valia Kordoni[2]

[1]Baidu, Sunnyvale, CA, USA and [2]Humboldt-Universitaet zu Berlin, Germany
*Corresponding author. Email: Kenneth.Ward.Church@gmail.com

(10 January 2022; revised 10 January 2022)

Abstract

Many papers are chasing state-of-the-art (SOTA) numbers, and more will do so in the future. SOTA-chasing comes with many costs. SOTA-chasing squeezes out more promising opportunities such as coopetition and interdisciplinary collaboration. In addition, there is a risk that too much SOTA-chasing could lead to claims of superhuman performance, unrealistic expectations, and the next AI winter. Two root causes for SOTA-chasing will be discussed: (1) lack of leadership and (2) iffy reviewing processes. SOTA-chasing may be similar to the replication crisis in the scientific literature. The replication crisis is yet another example, like evaluation, of over-confidence in accepted practices and the scientific method, even when such practices lead to absurd consequences.

Keywords: State-of-the-art; Evaluation; Benchmarks; Leaderboards; Root causes; Leadership; Reviewing; Replication crisis

1. Common ground: SOTA-chasing considered harmful

Given its unpopularity, why are so many papers chasing state-of-the-art (SOTA) numbers? The point of this paper is not to argue against SOTA-chasing but to identify some of the root causes behind SOTA-chasing and to offer some constructive suggestions for the future.

This paper will define SOTA-chasing to refer to papers that report SOTA numbers, but contribute little of lasting value to the literature. The point is the pointlessness.

Rogers posted an excellent blog on SOTA-chasing.[a] Her blog addresses two questions that will also be discussed in this paper:

1. How did we get here?
2. And what can we do about it?

There are plenty of additional criticisms of SOTA-chasing in the literature (Bender *et al.* 2021; Raji *et al.* 2021). We have added our own criticism of SOTA-chasing:

> *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.* (Church 2017)

The next two sections will discuss costs of SOTA-chasing and root causes.

[a]https://hackingsemantics.xyz/2020/reviewing-models/

2. SOTA-chasing: Costs

The next three subsections will discuss three types of costs:

1. Leaderboards emphasize competition, distracting attention from more important opportunities to advance the field,
2. SOTA-chasing is sucking the oxygen out of the room, discouraging interdisciplinary collaboration with colleagues in linguistics, lexicography, psychology, etc., and
3. Claims of superhuman performance (on tasks that appear to be more realistic than they are) create unrealistic expectations that could lead to yet another AI winter.[b]

2.1 Incentives and coopetition: Leaderboards considered harmful

It is a cliche that *whatever you measure, you get*. Leaderboards emphasize winners and losers. If you do a Google search for "meme: mine is bigger than yours," you will find a bunch of rude, childish, and even dangerous images of nuclear brinkmanship. International relations and science should be better than school boys playing king-of-the-hill. Hopefully, there is more to the literature than boys being boys.

TREC[c] (Text REtrieval Conference) emphasizes coopetition.[d] (Voorhees 2021) as opposed competition. In her keynote at SIGIR-2020,[e] as well as her invited talk at the ACL-2021 Workshop on Benchmarking,[f] Voorhees observed that:

- *competing may give you a bigger piece of the pie ...*
- *... while cooperation makes the whole pie bigger*[g]

TREC participants are asked to sign a form that forbids explicit advertising of TREC results. This prohibition was mentioned a number of times in the videos of the 25th anniversary of TREC.[h] While participants appreciate the principle, the temptation to boast is difficult to resist.

Voorhees is making an important point. Consider the overview paper to the TREC Deep Learning track.[i] (Craswell *et al.* 2020), for example, where methods are split into three types: nnlm (neural net language models such as BERT), nn (other types of neural nets), and trad (traditional methods). Their Figure 1 shows performance is best for nnlm and worst for trad. That is, nnlm > nn > trad. In this way, coopetition produces important insights that advance the field in meaningful ways, in contrast with leaderboards that emphasize competition and schoolyard nonsense such as "mine is bigger than yours."

As a second example of coopetition and competition, consider MRQA (Machine Reading for Question Answering) (Fisch *et al.* 2019). The call for papers[j] highlights a number of admirable objectives such as domain transfer, interpretability, robustness, and error analysis, but unfortunately, the shared task[k] leads with a leaderboard and congratulates the winners, with no mention of the more admirable objectives.

[b] https://en.wikipedia.org/wiki/AI_winter
[c] https://trec.nist.gov/
[d] https://en.wikipedia.org/wiki/Coopetition
[e] https://dl.acm.org/doi/10.1145/3397271.3402427
[f] https://github.com/kwchurch/Benchmarking_past_present_future/blob/master/README.md#Voorhees
[g] slides 6-7 of https://github.com/kwchurch/Benchmarking_past_present_future/blob/master/slides/session3/Ellen_benchmarking.pdf
[h] https://trec.nist.gov/celebration/25thcelebration.html
[i] https://microsoft.github.io/msmarco/TREC-Deep-Learning-2020
[j] https://mrqa.github.io/2019/cfp
[k] https://mrqa.github.io/2019/shared

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

With a slightly different design, the shared task could have provided some interesting insights into domain transfer. Table 1 of (Fisch *et al.* 2019) lists 18 QA benchmarks, split up into three groups of six benchmarks. The three groups are used for train, validation, and test, respectively. Suppose instead of using this train/validation/test split, we used a number of different splits. Could we learn that transfer is more successful for some splits than others?

MRQA identifies some interesting similarities and differences among the 18 benchmarks:

- source of documents: Wikipedia/Web snippets/misc
  - Wikipedia (7 benchmarks): DROP, HotpotQA, QAMR, RelationExtraction, SQuAD, TREC, Natural Questions
  - Web snippets (3 benchmarks): TriviaQA, SearchQA, ComplexWebQ
  - misc (8 benchmarks): MCTest, RACE, DuoRC, NewsQA, BioASQ, QAST, BioProcess, TextbookQA
- source of questions: Crowdsourced/Domain Experts/misc
  - Crowdsourced (9 benchmarks): ComplexWebQ, DROP, DuoRC, HotpotQA, MCTest, NewsQA, QAMR, SQuAD, TREC
  - Domain Experts (5 benchmarks): BioASQ, BioProcess, QAST, RACE, TextbookQA
  - misc (4 benchmarks): SearchQA, Questions Natural, RelationExtraction, TriviaQA
- source of answers: based on documents/not based on documents
  - based on documents (9 benchmarks): SQuAD HotpotQA, DROP, RACE, TextbookQA, BioProcess, MCTest, QAMR, QAST
  - not based on documents (9 benchmarks): NewsQA TriviaQA, SearchQA, Natural Questions, BioASQ, DuoRC, RelationExtraction, ComplexWebQ, TREC

It would be very interesting to know if these patterns are important for transfer or not. There is pretty clear evidence, for example, that constructed (crowdsourced) questions are easier than questions from query logs. TREC QA,[l] for example, started with "constructed" questions in 1999, but quickly moved to "real" questions from query logs for subsequent TREC QA tracks (2000–2007) because constructed questions are too easy for systems and unrealistic (Voorhees 2001). Based on these observations, transfer might also be more effective between benchmarks that are similar to one another in terms of the source of questions, documents, and/or answers. In this way, coopetition could produce important insights that advance the field in more meaningful ways than leaderboards and competition.

It also helps to advance the field when benchmarks are realistic. Most of the benchmarks in MRQA are based on benchmarks from academia, except for Natural Questions. To construct more realistic benchmarks, it is advisable to work with industry and make sure the benchmark is representative of a real problem that they care about. A number of companies have been involved in a number of benchmark efforts:

- Microsoft Bing: TREC Web track[m] (1999–2014), TREC Deep Learning track[n] (Craswell *et al.* 2020)
- Baidu: DuReader[o] (He *et al.* 2018)
- Google: Natural Questions[p] (Kwiatkowski *et al.* 2019)

[l] https://trec.nist.gov/data/qamain.html
[m] https://trec.nist.gov/data/webmain.html
[n] https://microsoft.github.io/msmarco/TREC-Deep-Learning-2020
[o] https://github.com/baidu/DuReader
[p] https://ai.google.com/research/NaturalQuestions/download

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

There are also connections between the TREC QA track[q] (1999–2007) and IBM Watson *Jeopardy!* In this case, IBM started in 2006 with a system designed for the TREC QA track and discovered that that system did not work well enough for Jeopardy questions,[r] as discussed at a celebration for the 25th anniversary of TREC.[s] After 5 years of hard work, the IBM system beat the two best human Jeopardy players in 2011, but their 2011 system was probably very different from their 2006 system because, among other things, the TREC QA tasks are not very representative of the Jeopardy task. The Jeopardy task is a problem that matters to IBM marketing, though problems such as web search are probably more real than Jeopardy.[t]

Unfortunately, while we all know that IBM won, much less is remembered about how that was accomplished (Ferrucci *et al.* 2010; Ferrucci 2012), and how that achievement could have advanced the field toward more admirable goals. We should follow Voorhees's advice and replace competition with coopetition. The point is not who wins, but insights that advance the field.

2.2 Sucking the oxygen out of the room

What is not happening as a result of too much SOTA-chasing? It is becoming harder and harder to publish computational linguistics in a conference on computational linguistics. Students preparing for their first ACL paper may find textbooks on machine learning: (Bishop 2016; Goodfellow *et al.* 2016) to be more helpful than textbooks on computational linguistics: (Manning and Schutze 1999; Jurafsky 2000; Eisenstein 2019) and handbooks (Dale *et al.* 2000; Mitkov 2003; Clark *et al.* 2013).

ACL conferences used to be more inclusive. We used to see more people at our meetings from more fields such as linguistics, philosophy, lexicography, psychology, etc. ACL venues used to reach out to HLT (human language technology), a combination of computational linguistics, speech and information retrieval/web search. Lots of people used to publish in more combinations of fields/venues: computational linguistics (ACL, EMNLP, NAACL, EACL, Coling), Machine Learning (NeurIPS), Speech (ICASSP,[u] Interspeech[v]), Information Retrieval (SIGIR,[w] TREC), Web Search (WWW,[x] WSDM[y]), Datamining (KDD[z]), Language Resources (LREC[aa]), etc.

Why do we no longer see these people at ACL? It became clear to us that many of them no longer feel welcome when we attended an ACL-2014 workshop in honor of Chuck Fillmore.[ab] The workshop was bitter sweet. They were grateful that Chuck won a Lifetime Achievement Award,[ac] but they were also mourning his passing, and there were concerns about the relevancy of their work to where ACL was going. Fillmore's "Case for Case" (Fillmore 1968) has more than 11k citations in Google Scholar, but ACL is no longer interested in this approach, or in linguistic resources such as FrameNet (and much of what is discussed at LREC).

Reviewers, these days, sometimes suggest that resources such as FrameNet and WordNet are no longer relevant now that BERT works as well as it does. Such remarks discourage diversity. People who have invested in resources may find such remarks offensive (and unethical).

[q] https://trec.nist.gov/data/qamain.html
[r] https://j-archive.com/
[s] See minute 25 of part 3 of https://www.nist.gov/news-events/events/2016/11/webcast-text-retrieval-conference.
[t] https://futurism.com/neoscope/ibm-watson-ai-selling
[u] https://2022.ieeeicassp.org/
[v] https://www.isca-speech.org/iscaweb/index.php
[w] https://sigir.org/
[x] https://dl.acm.org/conference/www
[y] https://www.wsdm-conference.org/
[z] https://www.kdd.org/
[aa] http://www.lrec-conf.org/
[ab] https://aclanthology.org/volumes/W14-30/
[ac] https://aclweb.org/aclwiki/ACL_Lifetime_Achievement_Award_Recipients

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

Even people in Machine Learning have reservations about SOTA-chasing. Rahimi gave a Test of Time Award talk at NIPS-2017 titled "Machine Learning has become Alchemy."[ad] NIPS (now called NeurIPS) used to be more receptive to theory and what Rahimi referred to as the *rigor police*. Apparently, SOTA-chasing is squeezing out many important topics including theory and computational linguistics.

There is a different kind of rigor in other fields such as Lexicography, Library Science, and Information Retrieval, where proper attribution is taken very seriously. People in these fields care deeply about sampling (balance), what came from where, and what is representative of what. They will feel unwelcome when SOTA-chasing moves too quickly with less rigor. Consider the reference to TREC in HuggingFace[ae] as well as Table 1 of MRQA (Fisch *et al.* 2019), as discussed in Section 2.1. There have been 30 text retrieval conferences (TREC)[af] thus far. For each of those 30 conferences, there are many tracks and many datasets with many contributions from many people. We asked someone familiar with TREC for help disambiguating the references to TREC in HuggingFace and MRQA. The response was uncharacteristically sharp:

> *A reference to simply the "TREC collection" is underspecified to the point of being worthless, as you suspected*

It is important, especially in certain fields, to give credit where credit is due. Citing work with proper attributions will make our field more inclusive and more attractive to people in other fields with different priorities and diverse views of rigor. Proper citations will also facilitate replication.

2.2.1 What counts as computational linguistics (CL)?

As empiricists, we like to start with data. Here are seven ways to characterize CL. The first three have been discussed above, and the last four will be discussed below.

1. Interdisciplinary collaboration: Formal Linguistics, Philosophy,[ag] Psychology, Machine Learning, Statistics, Computer Science, Electrical Engineering, Phonology, Phonetics
2. Textbooks:[ah] (Manning and Schutze 1999; Jurafsky 2000; Eisenstein 2019) and handbooks: (Dale *et al.* 2000; Mitkov 2003; Clark *et al.* 2013)
3. Conference Venues: ACL, NeurIPS, AAAI, ICASSP, INTERSPEECH, TREC, LREC, WWW, WSDM, KDD
4. Organization of ACL Program Committees (Table 1)
5. Concepts suggested by search engine (Figure 1)
6. Studies of ACL Anthology (Anderson *et al.* 2012; Vogel and Jurafsky 2012)
7. Studies of Papers with Code (PWC) (Koch *et al.* 2021)

2.2.2 Organization of ACL program committees

Until recently, areas played an important role in setting the agenda. Table 1 shows the organization of the ACL-2021 Program Committee by area.[ai] Some areas have more SACs (senior area chairs) and ACs (area chairs) than others because some areas receive more submissions than others.

The call for papers[aj] encourages authors to submit their paper to one of these areas. Before the new ARR process (see Section 3.2 on reviewing processes), authors could expect their paper to be

[ad] https://www.youtube.com/watch?v=x7psGHgatGM
[ae] https://huggingface.co/datasets/trec
[af] https://trec.nist.gov/
[ag] https://plato.stanford.edu/entries/linguistics/
[ah] https://web.stanford.edu/~jurafsky/slp3/
[ai] https://web.archive.org/web/20210416171402/https://2021.aclweb.org/organization/program/
[aj] https://web.archive.org/web/20210416171402/https://2021.aclweb.org/calls/papers/

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

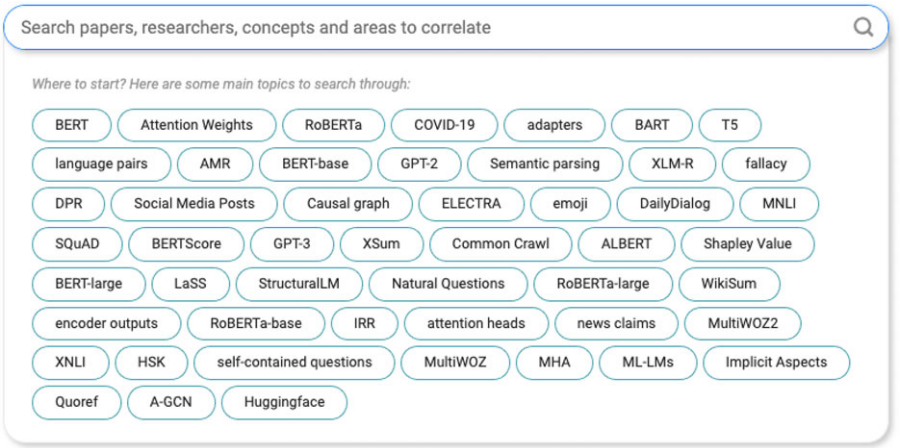

Figure 1. Default search concepts for ACL-2021, according to semantic paths (see footnote al).

handled by experts in the area. Handling includes both reviewing as well as assignments of papers to reviewers. Assignments used to be made by domain experts that know who's who and what's what. When assignments were made by domain experts, reviewers were more qualified and more sympathetic to the area than they are these days. As will be discussed in Section 3.2, the new ARR process no longer uses areas and domain experts to assign papers to reviewers, and consequently, assignments are probably more random.

Sessions were also typically organized by area. This way, the audience knew what to expect in a session on a particular topic.

These changes to the reviewing process have huge consequences on the field. There will be less diversity. Areas that are not well covered by action editors (AEs) will suffer. The rich will get richer. Areas near the top of Table 1 will benefit, and areas near the bottom of Table 1 will suffer, though there will be a few exceptions for a few micro-topics that happen to be favored by the (iffy) software for routing papers to reviewers.

2.2.3 Concepts suggested by search engine

As mentioned above, conference sessions used to be organized by areas. But that is changing as areas become deemphasized. The Program for ACL-2021,[ak] for example, leads with a pointer to a search engine.[al] This search engine offers a set of default concepts to search on, as shown in Figure 1. These concepts in Figure 1 are very different from areas in Table 1. The concepts emphasize datasets (COVID-19, Common Crawl, MNLI, SQuAD) and models (BERT, RoBERTa, BART).

The areas in Table 1 are closer to textbooks/handbooks on computational linguistics: (Manning and Schutze 1999; Dale *et al.* 2000; Jurafsky 2000; Mitkov 2003; Clark *et al.* 2013; Eisenstein 2019), and the concepts in Figure 1 are closer to textbooks on machine learning: (Bishop 2016; Goodfellow *et al.* 2016). Replacing the areas in Table 1 with the concepts in Figure 1 will have a dramatic impact on the field of computational linguistics. There will be less diversity and less room for papers that do not mention popular datasets and popular models.

[ak] https://web.archive.org/web/20210913160131/https://2021.aclweb.org/program/overview/
[al] https://acl2021.semanticpaths.org/

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

Table 1. ACL-2021 topics, sorted by the number of senior area chairs (SACs) and area chairs (ACs), based on footnote ai

SACs	ACs	Topic
4	32	Machine learning for NLP
3	29	Dialogue and interactive systems
3	25	Information extraction
3	23	Question answering
3	22	Machine translation and multilinguality
3	22	Generation
3	18	NLP applications
3	17	Semantics: Sentence-level semantics, textual inference, and other areas
3	14	Sentiment analysis, stylistic analysis, and argument mining
3	12	Interpretability and analysis of models for NLP
3	12	Computational social science and cultural analytics
2	12	Semantics: Lexical
2	12	Resources and evaluation
2	11	Summarization
2	11	Language grounding to vision, robotics, and beyond
2	11	Information retrieval and text mining
2	8	Syntax: Tagging, chunking, and parsing
2	7	Theme
2	6	Speech and multimodality
2	5	Linguistic theories, cognitive modeling, and psycholinguistics
2	5	Ethics in NLP
2	5	Discourse and pragmatics
2	3	Phonology, morphology, and word segmentation
2	3	Multidisciplinary and area chair COI

2.2.4 Studies of ACL anthology

The ACL Anthology provides a very different perspective from Figure 1. At ACL-2012, there was a Special Workshop on Rediscovering 50 Years of Discoveries[am] (Banchs 2012). A number of papers at that workshop provide tables of topics such as Table 1 in (Anderson *et al.* 2012) and Figures 7 and 8[an] in (Vogel and Jurafsky 2012). These topics look similar to topics covered in textbooks such as (Jurafsky 2000). Perhaps that should not be a surprise since Jurafsky is an author of all of those references. We are concerned, though, about the relatively small overlap between these topics and what we see in more recent ACL conferences. Recent ACL meetings appear to be moving toward Figure 1 and away from Table 1.

2.2.5 Studies of papers with code (PWC)

Analyses of PWC may be more useful than ACL Anthology for appreciating the move toward SOTA-chasing. Figure 3 of (Koch *et al.* 2021) reports usage of datasets in PWC. One might expect usage to be heavily skewed, following a Zipf-like law. They find that usage is becoming more and more concentrated over time in a small number of places. From this perspective, it is not surprising that 50% of the usage can be attributed to a short list of institutions:

[am] https://aclanthology.org/volumes/W12-32/
[an] https://nlp.stanford.edu/projects/gender.shtml

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

- Non-profit: Stanford, Princeton, Max Planck, CUHK, TTIC, NYU, Georgia Tech, Berkeley
- Corporate: Microsoft, Google, AT&T, Facebook

These results are perhaps not that surprising, but we found it remarkable what is missing from this list: for example, National Library of Medicine (PubMed),[ao] Linguistic Data Consortium.[ap] PWC probably has better coverage of machine learning than other fields such as speech, medicine, law, information retrieval, web search, etc. Venues such as TREC, LREC, Kaggle[aq] and many others are not well covered in PWC. PWC probably has more coverage of universities than industry and government (NIST, DARPA). PWC is probably missing much of what is happening in Asia, considering how much participation in ACL meetings is coming from Asia.

The PWC view of CL, as well as Figure 1, is probably more representative of where CL is going than other views discussed above such as Table 1, textbooks/handbooks, ACL Anthology. That said, we hope CL does not go down this rabbit hole. It is not a promising direction. We are particularly concerned that the PWC view and SOTA-chasing will reduce diversity and squeeze out many alternative perspectives.

2.3 Unrealistic expectations: Superhuman performance, seriously???

We now turn to the third of the three costs of SOTA-chasing mentioned at the beginning of Section 2. Claims of superhuman performance (on tasks that appear to be more realistic than they are) create unrealistic expectations that could lead to an AI winter.

It is not hard to construct CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart)[ar] as well as the reverse, which we call *reverse CAPTCHAs*. For standard CAPTCHAs, you can bet on people to succeed and machines to fail, whereas for reverse CAPTCHAs, you can bet on machines to succeed and people to fail.

Technology is often amazing, though sometimes exhausting, embarrassing, unethical, and/or dangerous. It is easy to find examples in the news and in social media of amusing/scary "computer errors." Alexa recently told a 10-year-old girl to do something dangerous with a penny and electricity.[as,at] Gmail autocorrect recently sent an embarrassing email where an interest in *speaking* with a business associate somehow came out as an interest in *sleeping* with the business associate.

Computers are being used for all sorts of use cases, raising some serious ethical questions (O'Neil 2016). In one case, a judge ruled that Google translate is not good enough to count as consent for a police search.[au] Society will need to address many more ethical questions like this.

If machines were actually better than people at transcribing speech and machine translation, then why are there so many "computer errors" in captions for services such as YouTube and Zoom? There is always more work to do. There are a few tasks, like playing chess, where computers are much better than people. But there are many tasks that are important for commercial applications, like captions, where there are opportunities for improvement.

[ao] https://pubmed.ncbi.nlm.nih.gov/
[ap] https://www.ldc.upenn.edu/
[aq] https://www.kaggle.com/
[ar] https://en.wikipedia.org/wiki/CAPTCHA
[as] https://www.bbc.com/news/technology-59810383
[at] https://twitter.com/klivdahl/status/1475220450598924297
[au] https://techcrunch.com/2018/06/15/judge-says-literal-but-nonsensical-google-translation-isnt-consent-for-police-search/

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

There have been claims at WMT (Workshop on Machine Translation) (Barrault *et al*. 2019) and elsewhere suggesting machines have achieved more than they have (parity/superhuman performance). The community tends to remember this simple take-away message, despite reassessments (Toral 2020), and cautionary caveats such as this:

> *This result has to be regarded with a great caution and considering the technical details of the... evaluation method as well as... Importantly, the language pairs where the parity was reached last year were not confirmed by the evaluation this year and a similar situation can repeat.* (Barrault *et al*. 2019)

Replication is a major problem for many fields, as will be discussed in Section 4 on the *replication crisis*. One of the root causes behind the replication crisis is over-confidence in the scientific method.

Evaluations can also be misleading because of over-confidence in the methodology and metrics such as BLEU. The community is more likely to remember the superhuman result than cautionary caveats/reassessments such as *your mileage may vary* (YMMV).[av]

Consider technology for translating meetings, for example. This technology is amazingly good, but far from human parity. The first author has considerable experience with this technology because he works for a Chinese company and does not speak Chinese. When he has access to a human interpreter, he is much more engaged in the meeting (and not nearly as exhausted).

When this technology was first introduced, everyone was impressed with how well it worked. The live stream was displayed on stage so everyone in the room could read whatever was said over the house speakers in real time in both English and Chinese. The chair of one high-profile session made a point to call out the technology.[aw]

Now that the technology has been around for a few years, the technology is no longer displayed on stage (perhaps because of a few inevitable embarrassing mistakes). The latest version runs on phones, so only those of us who need the technology can see (and hear) everything (warts and all) in both English and Chinese. The technology is even better than it used to be, especially with respect to latency, but even so, it is far from human parity.

Others who do not use the technology as much as we do may be misled by evaluations that report superhuman BLEU scores and latency. While the machine may be better than people in those terms, BLEU and latency are not the terms that matter. Professional interpreters translate what needs to be said when it needs to be said, and human interpreters do not make "computer errors."

We can correct some computer errors. For example, when the machine says "ecology" in a Baidu meeting, the speaker is probably talking about "eco-systems." Other computer errors are more challenging. One time, we guessed the machine made an error when it used a politically sensitive word. We asked for clarification after the meeting. In this case, it turned out that the translation was actually correct, but the context was "lost in translation." When we have access to a human interpreter, there are more opportunities to ask for clarifications and less need to do so.

Given realities such as these, when evaluations produce numbers that are too good to be true (using inappropriate metrics such as BLEU and average latency), why do we take these numbers so

[av] https://www.urbandictionary.com/define.php?term=ymmv

[aw] *Recently, Robin Li, the CEO of Baidu, gave a talk to about 200 of us in Chinese. The technology worked great for his talk, but they turned off the translation technology at the end of his talk, making it all too clear to me just how dependent I was on the translation technology. After they turned off the translation technology, I am less clear about what happened, but I am guessing that the chair of the session asked the audience for questions. When no one seemed eager to ask the first question, the chair handed the microphone to me and "volunteered" me to ask the first question. It felt a bit like that line from Dirty Harry: "Do you feel lucky, punk?" There was some nervous laughter at that point; it was a rather high-stakes public evaluation of their translation technology (using a novel metric they were not prepared for). Did their technology work well enough that I could ask a sensible question?* (Church 2018*b*)

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

seriously? We have so much confidence in our evaluation methodology that we believe the results (and gloss over caveats/reassessments), even when we know the results cannot be right:

> *The first principle is that you must not fool yourself and you are the easiest person to fool*[ax]
> —Feynman

A number of evaluations are reporting that machines are better than people on a number of tasks[ay,az,ba,bb,bc,bd] (Nangia and Bowman 2019; Nguyen *et al.* 2021). We all know these superhuman numbers are too good to be true and unlikely to transfer beyond academic benchmarks to tasks that matter for commercial practice. No one will remember the caveats/reassessments, but they will remember the unrealistic expectations, and that will not be good for the long-term health of the field.

Viewed in this way, the successes of deep nets on so many benchmarks could be interpreted as a criticism of these benchmarks. Benchmarks tend to focus too much on tasks that are ideal for technologies we already have. But benchmarks should place greater emphasis on opportunities for improvement. Benchmarks should be different from PR hype. The point of benchmarks is not to make our technology look good (or better than it is), but to help set the agenda for future work. Evaluations provide credible measurements of progress, as well as realistic expectations for the future.

We are not objecting to evaluation, and measuring real progress. But we are objecting to "gains" that are more noise/hope/hype than progress. The difference between the top two places on a leaderboard may not be significant or replicable or interesting.

3. Root causes for SOTA-chasing

Sections 3.1 and 3.2 will discuss two possible root causes for SOTA-chasing:

1. Lack of leadership and long-term strategic planning: historically, the agenda was determined top-down by a relatively small number of influential leaders in academia, industry, and government, but these days, the agenda is evolving more bottom-up via social media and websites such as papers with code (PWC)[be] and HuggingFace's lists of frequently downloaded models and datasets.[bf] As a result of these changes, the emphasis has become more short-term and more transactional.
2. Poor reviewing as a result of poor assignments of papers to reviewers by a combination of iffy programs and ineffective processes for correcting the mistakes of these programs.

3.1 SOTA-chasing: A consequence of a lack of leadership

SOTA-chasing may have evolved out of the evaluation tradition, which has a long history. (Raji *et al.* 2021) start by summarizing some of this history in (Lewis and Crews 1985; Liberman 2010;

[ax] https://protect-eu.mimecast.com/s/tzLMCzKPPtNVvgghoR1s2?domain=sites.cs.ucsb.edu
[ay] https://www.wired.com/1999/10/superhuman-speech-machine/
[az] https://cacm.acm.org/news/220485-microsoft-claims-new-speech-recognition-record-achieving-a-superhuman-51-error-rate/fulltext
[ba] https://www.businessinsider.com/ibm-speech-recognition-almost-super-human-2017-3
[bb] https://www.kit.edu/kit/english/pi_2020_095_ai-outperforms-humans-in-speech-recognition.php
[bc] https://ai.facebook.com/blog/facebook-leads-wmt-translation-competition/
[bd] https://medium.com/@yixing.cai/whats-aoa-a-model-that-beats-human-performance-in-squad-2-0-15422559bda2
[be] https://paperswithcode.com/
[bf] https://huggingface.co/

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

Church 2018*a*). Historically, there was a point to the emphasis on evaluation; evaluation used to be more than pointless SOTA-chasing.

Many first-hand accounts of this history were presented at the ACL-2021 Workshop on Benchmarking: Past, Present, and Future (BPPF) (Church *et al.* 2021*a*). Videos and slides are posted on github.[bg]

Much of this history involves influential leaders such as John Mashey, Fred Jelinek, and Charles Wayne, as will be discussed in Section 3.1.1. Before Mashey, Jelinek and Wayne, the agenda was largely set by many other influential leaders: Pierce, Skinner, Shannon, Licklider, Minsky, Chomsky, and others (Church 2011). These days, one might try to argue that the agenda is coming top-down from Turing Award Winners such as Hinton, Bengio, LeCun, Pearl, and others. Bengio, for example, is working on some long-standing hard problems in Artificial Intelligence such as causality[bh] (Bengio *et al.* 2019; Schölkopf *et al.* 2021) and compositionality.[bi] Despite such top-down efforts, though, we view SOTA-chasing as evidence that the agenda is, in fact, emerging more bottom-up from community-driven sources such as papers with code (PWC) and HuggingFace.

This paper will suggest that SOTA-chasing is a consequence of a lack of top-down leadership. Students need help finding projects to work on. Success is measured transactionally. What does it take to get a paper accepted in the next conference? Publish or perish. Unless we offer a more promising alternative, students are likely to turn to PWC to find a project that is likely to "succeed" in the next round of conference reviews. Long-term success is more of a concern for more established researchers with more experience and more responsibility for the long-term health of the field.

Established researchers, such as authors of textbooks, used to play more of a role in setting the agenda. The connection between textbooks (Manning and Schutze 1999; Jurafsky 2000; Eisenstein 2019) and ACL meetings used to be stronger than it is today, as discussed in Section 2.2.

These days, the agenda is determined more bottom-up by mouse clicks. Everyone has an equal vote. The author of a textbook has no more vote than a student just starting out. Consequently, short-term concerns tend to dominate long-term concerns since the voting block of students starting out is much larger than the relatively small number of established researchers. The agenda is no longer determined by authors of textbooks and influencers such as John Mashey, Fred Jelinek and Charles Wayne.

3.1.1 Mashey, Jelinek and Wayne

John Mashey was one of the founders of SPEC,[bj,bk] an important benchmark for measuring CPU performance since 1988. SPEC has probably had more influence over commercial practice than all of the benchmarks in PWC combined.

Fred Jelinek was a manager in charge of much of IBM's work in Speech (Jelinek 1976 1997) and Machine Translation (Brown *et al.* 1990 1993) in the 1970s and 1980s, before moving to Johns Hopkins University and creating CLSP (Center for Language and Speech Processing).[bl] Bob Mercer worked closely with Fred Jelinek when they were both at IBM. Both Jelinek and Mercer received ACL Lifetime Achievement Awards in 2009 and 2014, respectively (see footnote ac).

[bg] https://github.com/kwchurch/Benchmarking_past_present_future
[bh] https://www.youtube.com/watch?v=rKZJ0TJWvTk
[bi] https://slideslive.com/38922794/towards-compositional-understanding-of-the-world-by-agentbased-deep-learning
[bj] https://www.spec.org/
[bk] youtube.com/watch?v=koSuxS3QFDk
[bl] https://www.clsp.jhu.edu/

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

Mercer was an early advocate of end-to-end methods (Church and Mercer 1993). Jelinek preferred the older term: *self-organizing systems* (Farley and Clark 1954; Von Foerster 1960; Jelinek 1990).

Charles Wayne played an important role in US government funding agencies including DARPA[bm] and NSA.[bn] In the US government, projects are typically designed to run for 5 years or so but somehow our field enjoyed nearly continuous funding for three decades starting in the mid-1980s (Church 2018*a*; Liberman and Wayne 2020).

Liberman[bo,bp] attributes the funding success to Wayne's emphasis on evaluation. Before Wayne, there had been an "AI Winter," largely as a result of Pierce's criticism of speech recognition in "Whither Speech Recognition?" (Pierce 1969) and Pierce's criticism of Machine Translation in the ALPAC report[bq] (Pierce and Carroll 1966):

> *It is clear that glamour and any deceit in the field of speech recognition blind the takers of funds as much as they blind the givers of funds. Thus, we may pity workers whom we cannot respect.* (Pierce 1969), p. 1049

Wayne's emphasis on evaluation was more glamour-and-deceit-proof than previous approaches to Artificial Intelligence. This approach enabled funding to start after a long "AI Winter" and to continue for many decades because funders could measure progress over time. Crucially, though, unlike many of the benchmarks that we work on today, the benchmarks under Wayne's leadership were very much driven by top-down strategic planning, with clear long-term goals.

Wayne encouraged interdisciplinary collaboration. He created a series of HLT (Human Language Technology) conferences by reaching out to natural language processing (NLP), information retrieval (IR), and speech. Wayne also played an important role in the creation of TREC (Text REtrieval Conference).[br] TREC is closely associated with NIST (National Institute of Standards and Technology), part of the U.S. Department of Commerce.

3.1.2 Strategic planning in government

There is a long tradition of top-down strategic planning in government agencies such as NIST and DARPA. NIST's mission is:[bs]

> *To promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.*

Their core competences are measurement science, rigorous traceability, and development and use of standards.

In the early 1960s, Licklider created IPTO (Information Processing Techniques Office) under DARPA with the mission to:[bt]

[bm] https://www.darpa.mil/

[bn] https://www.nsa.gov/

[bo] https://www.simonsfoundation.org/event/reproducible-research-and-the-common-task-method

[bp] https://github.com/kwchurch/Benchmarking_past_present_future#Liberman

[bq] https://www.nap.edu/html/alpac_lm/ARC000005.pdf

[br] The history of TREC https://trec.nist.gov/ was discussed at the 25th anniversary. Videos are available online. Wayne's contributions are mentioned at minute 15 of part 1 of https://www.nist.gov/news-events/events/2016/11/webcast-text-retrieval-conference.

[bs] https://www.nist.gov/about-nist/our-organization/mission-vision-values

[bt] https://en.wikipedia.org/wiki/Information_Processing_Techniques_Office

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

> *create a new generation of computational and information systems that possess capabilities far beyond those of current systems. These cognitive systems—systems that know what they're doing:*
>
> 1. *will be able to reason, using substantial amounts of appropriately represented knowledge;*
> 2. *will learn from their experiences and improve their performance over time;*
> 3. *will be capable of explaining themselves and taking naturally expressed direction from humans;*
> 4. *will be aware of themselves and able to reflect on their own behavior;*
> 5. *will be able to respond robustly to surprises, in a very general way.*

Our field has made considerable progress on some of these goals, though much work remains to be done.

While it is tempting to blame many of the leaders mentioned in this history for the current SOTA-chasing craze, that would be unfair. SOTA-chasing, as defined above, involves pointless numbers with little long-term strategic value, whereas the leaders in this history made important long-term contributions to the field largely because they placed such a high value on long-term strategic planning.

3.2 SOTA-chasing: A consequence of poor reviewing processes

In addition to a lack of leadership, another root cause of SOTA-chasing is poor reviewing processes. Rogers's blog attributes SOTA-chasing to lazy/poor reviewing, information overload (drowning in papers), and glorification of benchmarks, though there is more discussion of poor reviewing in her blog than glorification of benchmarks. Her blog leads with the following tweet:[bu]

> *Another @emnlp2019 reviewer's 3-line review concludes: "The main weakness of the paper is the results do not beat the state of the art models." This is a tired take and a lazy, backwards way to think about research.*

It is a shame that EMNLP reviewing is as bad as it is. EMNLP's poor reviewing is particularly ironic given that we created EMNLP largely as a reaction to ACL's poor reviewing, as discussed in Section 1.2 of (Church 2020). EMNLP reviewing is used to be quicker than ACL by construction.[bv] These days, EMNLP reviewing is no quicker (and no better) since all ACL* conferences use the same (broken) processes. Putting all our eggs in one basket is not a solution, especially if the basket is known to be defective.

There have been many criticisms of reviewing recently. Rogers's blog offers a number of constructive suggestions to reviewers. While we agree that reviewing is bad, and even worse than it used to be, blaming the reviewers is unlikely to lead to improvements. Reviewers do what reviewers do. Creating more tutorials,[bw] rules and process are unlikely to help.

It is widely agreed that ACL reviewing is an opportunity for improvement. The ACL has recently rolled out a new rolling review process (ARR)[bx] based on open review. Perhaps ARR will improve matters, though we have serious doubts.

Best practices tend to start by identifying root cause(s). Introducing change for change sake is unlikely to lead to improvements. Roll out new processes gradually; do not make too many changes at the same time.

[bu] https://twitter.com/_jessethomason_/status/1147587570634645504

[bv] EMNLP's submission date was immediately after ACL's notification date, but the conferences were held at the same time.

[bw] https://aclrollingreview.org/reviewertutorial

[bx] https://aclrollingreview.org/cfp

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

Table 2. ACL-2022 (after ARR) is no faster than ACL-2021 (before ARR)

Venue	Submission Date	Start Date	Days
ACL-2021	Feb 1, 2021	Aug 2, 2021	182
ACL-2022	Nov 15, 2021	May 22, 2022	188

One thing is certain: as shown in Table 2, ACL-2022 is no faster than ACL-2021. It is claimed that ARR is faster:

> *The original goal of ARR was to have all reviews and meta-reviews completed within 35 days. The process requires that each paper has 3 reviews, and once those are complete, a meta-review. This is a pretty tight turnaround. By comparison, in ACL 2021, ... the time from submission to notification amounted to 92 days, nearly 3 times what ARR is aiming for.*[by]

but ACL-2022 used ARR and ACL-2021 did not. If ARR really was 3x faster, why doesn't that speed up show up in the schedules in Table 2?

Speed is important, but quality is even more important. Why is reviewing so bad? Reviewers are tired and underpaid, as Rogers points out in her blog. But that is also true of researchers. Most of us do what we do because we care deeply about what we do. Researchers are also tired and underpaid. That is not the root cause for bad reviewing.

A more likely root cause is the assignment of papers to reviewers. ARR has a number of serious design flaws that make it very likely that reviewers will be less qualified/sympathetic than they used to be. It used to be rare for students to be invited to review. Reviewers were typically authors of cited papers, increasing the odds that reviewers would be familiar with relevant background material, and positively inclined toward the general approach. Reviewers used to have more expertise in the topic than the target audience for the paper. Unfortunately, that is no longer the case.

It appears that ARR automates the assignment of papers too much. In Section 2.7.1 of (Church 2020), we discussed a number of common methods for assigning papers to reviewers:

1. Delegate to authors (keywords)
2. Delegate to reviewers (bidding)
3. Delegate to middle management (manual assignments)
4. Delegate to software using automated routing (Yarowsky and Florian 1999)
5. Semi-automatic routing (automatic routing with manual post-editing).

ARR uses automatic routing with too little post-editing. Action editors are asked to over-ride initial assignments,[bz] but this process is unworkable because of fluctuations in the workload. There were just barely enough AEs for the October-2021 round of ARR, but not nearly enough for November 2021 because of a submission deadline for ACL-2022.

To make matters worse, ARR eliminated areas, a huge mistake in our opinion, because action editors cannot be expected to know who's who and what's what in all areas.

We prefer a recursive structure where papers are assigned to subcommittees going down the tree, until the number of papers is small enough that a subcommittee can handle the load. With

[by] https://web.archive.org/web/20211013141438/https://aclrollingreview.org/status-report/

[bz] See https://web.archive.org/web/20211102183707/https://aclrollingreview.org/aes for instructions to action editors (AEs): *It is important to select reviewers with expertise appropriate to the paper. Some suggestions are listed based on an automatic score, but this score may not always be reliable, so feel free to make modifications if they are not reliable. Click on a reviewer's name to see their profile and publications; if they are a good match and do not have too many assignments already, you can unassign one of the current* [sic] *reviewers and add a new reviewer.*

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

this structure, the subcommittee chair can be expected to know who's who and what's what in their subarea. Chairs should be encouraged to send papers back up the tree if they receive papers that go beyond their expertise.

There are some obvious weaknesses with automatic routing and the ARR process. We described our experience with (Yarowsky and Florian 1999) in (Church 2020). Automatic routing loves conflicts of interest. If there is a way to send a paper to reviewers with conflicts of interest, automatic routing programs will do just that.

Since ACL-2022 makes it possible to see the names of the other reviewers, we can check for conflicts of interest. ACL-2022 assigned one paper to two reviewers that work closely together. When we complained to the meta-reviewer, after apologizing for taking a long time to reply (because the system was sending too many unimportant emails), the meta-reviewer made it clear that there are too many such conflicts to fix.

Perhaps ARR chose to use automatic routing to cope with the scale. This is not a good reason though. The routing process is extremely important. If a paper is sent to an unqualified/unsympathetic reviewer, then the reviewer is likely to essentially "abstain" and kill the paper with an average/low score.

Automating the routing process is disrespectful to authors and reviewers and the community. Reviewers do not like to review papers outside their area. Authors work hard on their submissions. We owe them more than "abstentions." The community deserves to know that published papers have been credibly reviewed by qualified reviewers with considerably more expertise in the area than the target audience. It is a violation of ethics to route papers the way we do without a reasonable number of qualified experts involved in the routing process.

Why does ARR believe that it is necessary to automate the process? Scale was mentioned above. ACL does receive quite a few submissions. There were about 3000 submissions in November 2021, about an order of magnitude more than we had for EMNLP in 1999. For EMNLP-1999, the first author assigned each paper to a subcommittee in two days, at a rate of 2 minutes per paper. Those subcommittees would then assign papers to reviewers.

This process is embarrassingly parallel. Thus, with about ten people doing what the first author did for EMNLP-1999, we could route the 3000 papers to subcommittees in two days. The process also scales recursively. If a subcommittee receives too many papers, we can split the subcommittee recursively into subcommittees, as needed. The entire routing process should require no more than a week.

In principle, the law of large numbers should make it easier to find good matches between papers and reviewers. Scale is not an excuse for disrespecting authors and reviewers.

To make matters even worse, ARR uses the wrong population to sample reviewers. Now that all authors are required to review, including students[ca] and authors from other areas, it is very likely that reviewers will be unqualified and unmotivated. The process should sample reviewers from the population of published/cited papers, not submitted papers.

Chairs are always looking for more reviewers. Selecting reviewers from cited papers should help since the set of cited papers is much larger than the set of submitted papers.[cb]

There had been some talk when ARR was first proposed about minimum requirements for reviewers in terms of h-index and/or publications, but since ARR makes it easy to see the names of other reviewers, it is easy to verify that many/most of reviewers are not as qualified as they were a few decades ago.

SOTA-chasing is a natural consequence of these new (but not improved) processes. Since authors cannot assume that reviewers are qualified or sympathetic to the area, authors need to

[ca] There are manual filters that attempt to weed out authors without PhDs, but these filters do not always succeed.

[cb] Authors of cited papers should be given the choice to review or not. But authors of submitted papers are expected to review (if asked). Apparently, ACL is having a problem with freeloading. Some authors are too "busy" to review, though they have time to submit 10+ papers. It may be necessary to introduce penalties for freeloading if freeloading is causing the system to collapse.

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

come up with a simple argument that will work with unmotivated reviewers. Empirically, authors have discovered that SOTA-chasing is effective with random reviewers.

We cannot blame authors for doing what they are doing. Nor can we blame reviewers for doing what they are doing. We have seen the problem, and it is us.

3.2.1 Recommendations

What can we do about bad reviewing?

1. Governance: Leadership and organizational structure
2. Dashboards: Goals, Milestones, Metrics
3. Incentives: Align incentives across organization

ARR looks like it was designed by academics that have never worked in a large organization. Leadership and organizational structure are important now that conferences have become as large as they are. We need to run the review process more like a large company or military organization, with clear roles and responsibilities.

The current ARR process has too much turnover. Chairs of conferences are only involved in the process for a few months. It takes more experience than that to run a large organization.

Executives in large companies have often worked for few decades in a number of different positions within the organization for about 18 months per rotation. After "punching their ticket" in this way, they have a broad understanding of the organization from many perspectives, as well as a valuable personal network so they can call in favors as necessary. The top of the reviewing organization should be an executive with this kind of experience and personal connections.

The rest of the organization needs to scale appropriately so people are not running around like headless chickens because they have too much work to do. As mentioned above, scale is not an excuse for doing a bad job. Large organizations are all about logistics and process. The organization is not running smoothly if people are working too hard. Accomplishments count more than activities.

It should be clear who is responsible for what. If systems are sending out too many emails that no one is reading, it should be clear who is responsible for fixing that. And no one should have responsibility without authority or vice versa.

In addition to governance, we also need goals, milestones, and metrics. It is important that these are aligned. Many organizations end up optimizing the wrong metric (Goldratt and Cox 2016). Factories in the failed Soviet Union, for example, produced too many widgets that no one wanted. It is important to match supply and demand.

So too, a reviewing process should not be optimizing throughput, but the quality of the results. Metrics such as citations can be used to measure progress toward these goals. Milestones are often defined in terms of metrics and dates. For example, a sales organization might set a milestone for sales at the end of every month. Similarly, a reviewing organization could set milestones to make sure the reviewing process is on schedule.

Less is more. If the dashboard is too complicated, no one will look at it. The dashboard should be actionable. It should lead with an executive summary, a single page that an executive can understand. Other people should be able drill down to see more detailed views that are more relevant to their roles and responsibilities.

Visibility is essential. If people cannot see the dashboard, they will not use it.

Avoid micro-managing. If everyone understands their roles and responsibilities, as well as their metrics and milestones, then they will figure out how to get the job done. There are too many long documents on the ACL wiki that no one is reading. Do not create documents that are not read. Better to emphasize metrics. If you create a document, measure usage, and reward people that create resources that are found to be useful by the metrics.

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

With appropriate incentives, governance, and dashboards, everyone in the organization will figure out what needs to be done. This type of structure scales more effectively than centralized planning where everyone waits for instructions from above.

4. Replication crisis

The media has coined the term, *replication crisis*,[cc,cd] following some influential papers such as "Why Most Published Research Findings Are False" (Ioannidis 2005). The replication crisis is yet another example, like evaluation, of over-confidence in accepted practices and the scientific method, even when such practices lead to absurd consequences. Several surveys suggest that most/much of the literature is wrong:

> *Amgen researchers declared that they had been unable to reproduce the findings in 47 of 53 "landmark" cancer papers* (Begley and Ellis 2012; Baker 2016*b*)

> *More than 70% of researchers have tried and failed to reproduce another scientist's experiments*[ce] (Baker 2016*a*).

Even the famous "marshmellow" experiment may be wrong:

> *A new replication study of the well-known "marshmallow test"—a famous psychological experiment designed to measure children's self-control—suggests that being able to delay gratification at a young age may not be as predictive of later life outcomes as was previously thought*[cf,cg] (Watts *et al.* 2018)

Ironically, there are suggestions these criticisms of replicability are themselves difficult to replicate. Some suggest that "only" 14% of the literature is wrong (Jager and Leek 2014*a*; Ioannidis 2014; Jager and Leek 2014*b*), though even 14% is considerably more than chance based on *p*-values. Standard assumptions based on *p*-values may not hold because of an unethical (but not uncommon) practice known as "*p*-hacking" (Bruns and Ioannidis 2016).

Following the cliche, *to err is human; to really foul things up requires a computer*, machine learning could easily make the replication crisis even worse. Suppose *p*-hacking is a reverse CAPTCHA, an optimization that machines can do better than people can. To make matters even worse, given how hard it is to figure out what deep nets are doing, we might not even know if our nets are "doing real science" or adding automation to the replication crisis.

What are the root causes for the replication crisis? Much of the discussion points to incentive structures, and especially: *publish or perish*. Surprising results are more likely to be accepted, especially in top venues with low acceptance rates. As a result, the literature is full of experiments reporting "significant results" on surprising hypotheses. By construction, such results are unexpected, publishable, and probably wrong. This process encourages lots of junior researchers to try lots of long-short experiments, and publish the few that reach "significance." Normally, in a casino, the house almost always wins by construction, but in this case, the literature is designed to fail. A few lucky researchers will win the lottery and land a good job, but the literature will end up full of false positives.

[cc] https://www.youtube.com/watch?v=n4S1rJsStbA
[cd] https://en.wikipedia.org/wiki/Replication_crisis
[ce] https://www.bbc.com/news/science-environment-39054778
[cf] https://www.sciencedaily.com/releases/2018/05/180525095226.htm
[cg] https://roadofneurosurgery.com/wp-content/uploads/2020/11/dba80ddc6f3dcf485c1a9b91b6e7899e.pdf

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

The incentive structures in our field may also be suboptimal. If we encourage too much SOTA-chasing, then the literature will be full of papers that will not stand up to the test of time, by construction.

What can be done about the replication crisis? Many suggestions have been discussed in the references above and elsewhere:

1. More robust experimental design
2. Better statistics
3. Better mentorship
4. More process/requirements/paperwork (checklists)
5. Pre-registration of experiments

Surveys report large majorities in support of many of these proposals.

There is general agreement that the literature is not self-correcting (Ioannidis 2012), and the replication crisis is unlikely to fix itself without intervention. There is also agreement on the need to take action soon:

> *given that these ideas are being widely discussed, even in mainstream media, tackling the initiative now may be crucial. "If we don't act on this, then the moment will pass, and people will get tired of being told that they need to do something"* (Baker 2016*a*)

Much of this literature focuses on the dangers of misusing *p*-values, but misuse of *p*-values is just one of many ways for the literature to fool the public (and itself). As mentioned above, the community tends to remember superhuman performance on a benchmark, but not the cautionary caveats/reassessments.

There has been quite a bit of discussion of replication in our field, as well.[ch] Thanks to websites such as github and HuggingFace, replication in our field is easier than it used to be.

Even so, perhaps due to the use of crowdsourcing on these sites, as well as the lack of peer review, it can be difficult to figure out what came from where and what is representative of what. Consider, for example, two HuggingFace datasets: (1) *ptb_text_only*[ci] and (2) *trec* (see footnote ae). As mentioned in Section 2.2, there have been 30 TREC conferences so far, and each conference runs many tracks. We need to find a way to encourage better citations that give credit where credit is due. Different people are responsible for the data for different tracks.

As discussed in Section 11 of (Church *et al.* 2021*b*), the documentation suggests that *ptb_text_only* is from (Marcus *et al.* 1993). That said, credit should go to the people that collected the corpus since the parse trees, the main contribution of (Marcus *et al.* 1993), are no longer in the "text-only" version: *ptb_text_only*. In addition, most of the content words have been replaced with $< unk >$, making it much easier (and less useful) to predict the next "word," though these predictions have subsequently become the standard PTB task in PWC.[cj] Since this task has little to do with P (the data was not collected at Penn) and little to do with TB (there are no treebanks) and little to do with words, perhaps we should refer to this task as "$< unk >$ prediction in the non-P non-TB corpus." $< unk >$ prediction, is, of course, a silly task. No one should care very much about how well computers can predict $< unk >$, or whether they can do that better than people. <unk> prediction might appear to be related to (Shannon 1951), but predicting entropy of English is a real problem with important applications, and <unk> prediction is not.

Similar comments may apply to many other popular benchmarks such as SQuAD, where superhuman performance is claimed (see footnote bd) on a task with constructed questions. As

[ch] https://www.slideshare.net/aclanthology/joakim-nivre-2017-presidential-address-acl-2017-challenges-for-acl

[ci] https://huggingface.co/datasets/ptb_text_only

[cj] https://paperswithcode.com/sota/language-modelling-on-penn-treebank-word

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

discussed in Section 2.1, TREC QA started with constructed questions, but quickly moved to real questions from query logs for subsequent TREC QA tracks (2000–2007) because constructed queries are too easy for systems, and unrealistic (Voorhees 2001).

To summarize, many fields including our own are undergoing a replication crisis. It is helpful to shed light on the crisis, though we wish we had more constructive suggestions to offer. Many fields are taking steps to address these concerns. Github and HuggingFace have been helpful in our field though we need to be more careful documenting what came from where, and how these datasets/benchmarks fit into a larger strategic roadmap for advancing the field.

That said, we fear that process improvements, such as many of the suggestions above (that are supported by large majorities), are unlikely to be effective, and may even distract the community from addressing root causes. The system would be more self-correcting if incentives were aligned. Researchers need to believe that the best way to advance their career is to do what is in the best long-term interest of the field. As long as researchers are thinking short-term and transactional, then their incentives will not be aligned with the long-term interests of the field.

5. Conclusions/Recommendations

Many papers are SOTA-chasing, and more will do so in the future. SOTA-chasing comes with many costs. We discussed three costs:

1. Too much competition (and not enough coopetition)
2. Sucking the oxygen out of the room (discouraging diversity)
3. Claims of superhuman performance set unrealistic expectations

The first two are opportunity costs. SOTA-chasing distracts attention away from more promising opportunities. The last cost is a risk. Everyone will remember the claims, and no one will remember cautionary caveats/reassessments. We may not have claimed to have solved all the worlds' problems, but that is what they will remember, and they will be disappointed when we fail to deliver.

After discussing those three costs, we discussed two root causes: (1) lack of leadership and (2) iffy reviewing processes. Historically, the agenda was determined top-down by influencers in academia, industry, and government, but these days, the agenda is determined more bottom-up by social media (e.g., papers with code, HuggingFace).

SOTA-chasing can also be viewed as a vote of no confidence in the reviewing process. The reviewing process is so bad that authors have discovered that SOTA-chasing is more likely to get past unqualified/unsympathetic reviewers than alternatives that require reviewers with more domain expertise. It is tempting to blame the reviewers for their lack of expertise, but we believe the problem is more with the matching process than the reviewers. It should be possible to find qualified reviewers, but not with the current processes for assigning papers to reviewers.

After discussing root causes, we then discussed the replication crisis. The replication crisis is yet another example, like evaluation, of over-confidence in accepted practices and the scientific method, even when such practices lead to absurd consequences.

What do we do about SOTA-chasing? It is tempting to skip steps and discuss diagnosis and therapy, but the first step is to get past denial. If we can agree on priorities such as

1. need for leadership, and
2. need for better processes for matching papers and reviewers

then we can come up with a list of next steps to make progress on those priorities.

These days, it is hard for the ACL exec to set the agenda because the exec is too large, and too many positions rotate too quickly. When Don Walker was in charge, the ACL exec was more like the executive branch of government, and less like a committee/legislative branch of government.

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

Before Don Walker passed away, the exec was smaller, and he controlled most of the votes from 1976 to 1993.[ck] It is difficult for a large organization such as the ACL to run effectively without more leadership. Committees are effective for some tasks such as reaching consensus, but it is hard for large committees to lead.

References

Anderson A., Jurafsky D. and McFarland D.A. (2012). Towards a computational history of the ACL: 1980–2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, Jeju Island, Korea: Association for Computational Linguistics, pp. 13–21.

Baker M. (2016a). 1,500 scientists lift the lid on reproducibility. *Nature News* **533**(7604), 452.

Baker M. (2016b). Biotech giant posts negative results. *Nature* **530**(7589), 141.

Banchs R.E. (ed) (2012). *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, Jeju Island, Korea: Association for Computational Linguistics.

Barrault L., Bojar O., Costa-jussà M.R., Federmann C., Fishel M., Graham Y., Haddow B., Huck M., Koehn P., Malmasi S., Monz C., Müller M., Pal S., Post M. and Zampieri M. (2019). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy: Association for Computational Linguistics, pp. 1–61.

Begley C.G. and Ellis L.M. (2012). Raise standards for preclinical cancer research. *Nature* **483**(7391), 531–533.

Bender E.M., Gebru T., McMillan-Major A. and Shmitchell S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.

Bengio Y., Deleu T., Rahaman N., Ke R., Lachapelle S., Bilaniuk O., Goyal A. and Pal C. (2019). A meta-transfer objective for learning to disentangle causal mechanisms.

Bishop C. (2016). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer New York.

Brown P.F., Cocke J., Della Pietra S.A., Della Pietra V.J., Jelinek F., Lafferty J.D., Mercer R.L. and Roossin P.S. (1990). A statistical approach to machine translation. *Computational Linguistics* **16**(2), 79–85.

Brown P.F., Della Pietra S.A., Della Pietra V.J. and Mercer R.L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* **19**(2), 263–311.

Bruns S.B. and Ioannidis J.P. (2016). P-curve and p-hacking in observational research. *PloS One* **11**(2), e0149144.

Church K. (2011). A pendulum swung too far. *Linguistic Issues in Language Technology* **6**(5), 1–27.

Church K. (2017). Emerging trends: I did it, I did it, I did it, but... *Natural Language Engineering* **23**(3), 473–480.

Church K. (2018a). Emerging trends: A tribute to Charles Wayne. *Natural Language Engineering* **24**(1), 155–160.

Church K. (2018b). Emerging trends: APIs for speech and machine translation and more. *Natural Language Engineering* **24**(6), 951–960.

Church K. (2020). Emerging trends: Reviewing the reviewers (again). *Natural Language Engineering* **26**(2), 245–257.

Church K., Liberman M. and Kordoni V. (2021a). Benchmarking: Past, present and future. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pp. 1–7.

Church K. and Mercer R. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics* **19**(1), 1–24.

Church K., Yuan X., Guo S., Wu Z., Yang Y. and Chen Z. (2021b). Emerging trends: Deep nets for poets. *Natural Language Engineering* **27**(5), 631–645.

Clark A., Fox C. and Lappin S. (eds) (2013). *The Handbook of Computational Linguistics and Natural Language Processing*. John Wiley & Sons.

Craswell N., Mitra B., Yilmaz E., Campos D. and Voorhees E.M. (2020). Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*.

Dale R., Moisl H. and Somers H. (eds) (2000). *Handbook of Natural Language Processing*. CRC Press.

Eisenstein J. (2019). *Introduction to Natural Language Processing*. MIT Press.

Farley B. and Clark W. (1954). Simulation of self-organizing systems by digital computer. *Transactions of the IRE Professional Group on Information Theory* **4**(4), 76–84.

Ferrucci, D., Brown, E. Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N. and Welty, C. (2010). Building Watson: An overview of the DeepQA project. *AI magazine* **31**(3), 59–79.

Ferrucci, D.A. (2012). Introduction to "This is Watson". *IBM J. Res. Dev.* **56**(3), 235–249.

Fillmore C.J. (1968). In Emmon Bach & R. Harms (eds.), Universals in Linguistic Theory. Holt, Rinehart, and Winston.

Fisch A., Talmor A., Jia R., Seo M., Choi E. and Chen D. (2019). MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of 2nd Machine Reading for Reading Comprehension (MRQA) Workshop at EMNLP*.

Goldratt E.M. and Cox J. (2016). *The Goal: A Process of Ongoing Improvement*. Routledge.

Goodfellow I., Bengio Y. and Courville A. (2016). *Deep Learning*. MIT Press.

[ck] https://www.aclweb.org/archive/misc/History.html

https://doi.org/10.1017/S1351324922000043 Published online by Cambridge University Press

He W., Liu K., Liu J., Lyu Y., Zhao S., Xiao X., Liu Y., Wang Y., Wu H., She Q., Liu X., Wu T. and Wang H. (2018). DuReader: A Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, Melbourne, Australia: Association for Computational Linguistics, pp. 37–46.

Ioannidis J.P. (2005). Why most published research findings are false. *PLoS Medicine* **2**(8), e124.

Ioannidis J.P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science* **7**(6), 645–654.

Ioannidis J.P. (2014). Discussion: Why "an estimate of the science-wise false discovery rate and application to the top medical literature" is false. *Biostatistics* **15**(1), 28–36.

Jager L.R. and Leek J.T. (2014a). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* **15**(1), 1–12.

Jager L.R. and Leek J.T. (2014b). Rejoinder: An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* **15**(1), 39–45.

Jelinek F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE* **64**(4), 532–556.

Jelinek F. (1990). Self-organized language modeling for speech recognition. *Readings in Speech Recognition*, pp. 450–506.

Jelinek F. (1997). *Statistical Methods for Speech Recognition*. MIT Press.

Jurafsky D. (2000). *Speech & Language Processing*. Upper Saddle River, NJ, USA: Pearson Education.

Koch B., Denton E., Hanna A. and Foster J.G. (2021). Reduced, reused and recycled: The life of a dataset in machine learning research. *NeurIPS*.

Kwiatkowski T., Palomaki J., Redfield O., Collins M., Parikh A., Alberti C., Epstein D., Polosukhin I., Devlin J., Lee K., Toutanova K., Jones L., Kelcey M., Chang M.-W., Dai A.M., Uszkoreit J., Le Q. and Petrov S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics* **7**, 453–466.

Lewis B.C. and Crews A.E. (1985). The evolution of benchmarking as a computer performance evaluation technique. *MIS Quarterly*, 7–16.

Liberman M. (2010). Fred jelinek. *Computational Linguistics* **36**(4), 595–599.

Liberman M. and Wayne C. (2020). Human language technology. *AI Magazine* **41**(2), 22–35.

Manning C. and Schutze H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Marcus M.P., Santorini B. and Marcinkiewicz M.A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* **19**(2), 313–330.

Mitkov R. (ed) (2003). *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Nangia N. and Bowman S.R. (2019). Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, pp. 4566–4575.

Nguyen T.-S., Stueker S. and Waibel A. (2021). Super-human performance in online low-latency recognition of conversational speech.

O'Neil C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Pierce J.R. 1969. Whither speech recognition? *The Journal of the Acoustical Society of America* **46**(4B), 1049–1051.

Pierce J.R. and Carroll J.B. (1966). Language and machines: Computers in translation and linguistics.

Raji I.D., Bender E.M., Paullada A., Denton E. and Hanna A. (2021). Ai and the everything in the whole wide world benchmark. *NeurIPS*.

Schölkopf B., Locatello F., Bauer S., Ke N.R., Kalchbrenner N., Goyal A. and Bengio Y. (2021). Toward causal representation learning. *Proceedings of the IEEE* **109**(5), 612–634.

Shannon C.E. (1951). Prediction and entropy of printed english. *Bell System Technical Journal* **30**(1), 50–64.

Toral A. (2020). Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal: European Association for Machine Translation, pp. 185–194.

Vogel A. and Jurafsky D. (2012). He said, she said: Gender in the ACL Anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, Jeju Island, Korea: Association for Computational Linguistics, pp. 33–41.

Von Foerster H. (1960). On self-organizing systems and their environments. In *Idem, Understanding Understanding: Essays of Cybernetics and Cognition, New York: Springer*, pp. 1–20.

Voorhees E. (2001). The TREC question answering track. *Natural Language Engineering* **7**(4), 361–378.

Voorhees E. (2021). Coopetition in IR research. In *ACM SIGIR Forum*, vol. 54, pp. 1–3. New York, NY, USA: ACM.

Watts T.W., Duncan G.J. and Quan H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science* **29**(7), 1159–1177.

Yarowsky D. and Florian R. (1999). Taking the load off the conference chairs-towards a digital paper-routing assistant. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Cite this article: Church KW and Kordoni V (2022). Emerging Trends: SOTA-Chasing. *Natural Language Engineering* **28**, 249–269. https://doi.org.10.1017/S1351324922000043