Natural Language Engineering (2019), **25**, pp. 753–767
doi:10.1017/S1351324919000275

CAMBRIDGE UNIVERSITY PRESS

EMERGING TRENDS

A survey of 25 years of evaluation

Kenneth Ward Church[1,*] and Joel Hestness[2]

[1]Kenneth Ward Church, Baidu, Sunnyvale, CA 94089, USA, [2]Cerebras Systems, Los Altos, CA 94022, USA
*Corresponding author. Email: kenneth.ward.church@gmail.com

(Received 27 April 2019; accepted 6 June 2019; first published online 19 July 2019)

Abstract

Evaluation was not a thing when the first author was a graduate student in the late 1970s. There was an Artificial Intelligence (AI) boom then, but that boom was quickly followed by a bust and a long AI Winter. Charles Wayne restarted funding in the mid-1980s by emphasizing evaluation. No other sort of program could have been funded at the time, at least in America. His program was so successful that these days, shared tasks and leaderboards have become common place in speech and language (and Vision and Machine Learning). It is hard to remember that evaluation was a tough sell 25 years ago. That said, we may be a bit too satisfied with current state of the art. This paper will survey considerations from other fields such as reliability and validity from psychology and generalization from systems. There has been a trend for publications to report better and better numbers, but what do these numbers mean? Sometimes the numbers are too good to be true, and sometimes the truth is better than the numbers. It is one thing for an evaluation to fail to find a difference between man and machine, and quite another thing to pass the Turing Test. As Feynman said, "the first principle is that you must not fool yourself – and you are the easiest person to fool."

Keywords: evaluation; survey; systems research

1. Introduction

This paper will survey 25 years of evaluation. There was very little evaluation 25 years ago, and now there is much more than there was. There is a clear appreciation that the field is better off than it was, though there has always been pushback. We will mention some of the pushback, and then provide some of our own. We all agree that evaluation is better than content-free debates (dominated by Highest Paid Person's Opinion (HiPPOs)), but there is a risk that evaluation can devolve into mindless metrics. We expressed a concern in Church (2017) that 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, it is no longer even considered desirable.

There are a number of metrics of metrics that might help make the metrics more meaningful. The community has made considerable progress toward reproducibility by sharing code and data (Wieling, Rawee, and van Noord 2018), making metrics more comparable. The psychological notions of reliability and validity—driving toward insight—are perhaps even more meaningful than reproducibility.

In systems research, there is an emphasis on general purpose computing. Many of our papers report performance on a single task with a single corpus under various specific conditions. Systems research places considerable value on general solutions that address a wide variety of use cases.

Historically, general purpose computing has tended to beat out specific solutions, because of Moore's Law and economies of scale; The rich get richer. Whoever has a larger market share has more money for R&D, and therefore, their hardware gets better faster than the competition. This emphasis on general purpose solutions will have more and more impact on our field going forward with the creation of MLPerf, a suite of Machine Learning benchmarks. We are already seeing a move toward more general purpose models of language such as ELMo (Peters *et al.* 2018),[a] GPT-2 (Radford *et al.* 2019),[b] and BERT (Devlin *et al.* 2018),[c] where the cost of training a single model can be amortized over a variety of use cases (perhaps with a little bit of fine-tuning).

2. Resources

Resources for linguistics and Machine Learning used to be hard to come by, but we now seem to have an embarrassment of riches. Tables 1 and 2 list some popular metrics[d] and shared tasks.[e] There are a number of conferences with shared tasks such as TREC,[f] CONNL,[g] WMT,[h] and KDD Cup.[i] Kaggle runs an unbelievable number of competitions.[j] They provide convenient access to a huge number of data sets.[k] The community has made considerable progress toward reproducibility by sharing code and data (Wieling *et al.* 2018). Conference papers these days are often based on standard benchmarks distributed by Linguistic Data Consortium (LDC)[l] and others. Numerous python packages such as natural language took kit (NLTK)[m] and statistics packages such as R,[n] so on, provide handy access to tools and data. All these methods make corpus-based work easier than it used to be; a rising tide lifts all boats.

2.1 Empiricism: 25 years ago

We did not always have such great resources, especially when we first revived empirical methods about 25 years ago. Even though we did not have as much as we have now, those were exciting times, as described in Church (2011):

> "The revival of empiricism in the 1990s was an exciting time. We never imagined that that effort would be as successful as it turned out to be. At the time, all we wanted was a seat at the table. In addition to everything else that was going on at the time, we wanted to make room

[a] https://github.com/allenai/allennlp/blob/master/tutorials/how_to/elmo.md.
[b] https://openai.com/blog/better-language-models/.
[c] https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html.
[d] Some useful URLs: https://en.wikipedia.org/wiki/Precision_and_recall (precision and recall), https://en.wikipedia.org/wiki/Mean_reciprocal_rank (MRR), https://en.wikipedia.org/wiki/Discounted_cumulative_gain (NDCG), https://github.com/vrama91/cider (CIDEr), https://en.wikipedia.org/wiki/Word_error_rate (WER), https://en.wikipedia.org/wiki/Mean_opinion_score (MOS), http://www.xavieranguera.com/phdthesis/node108.html (DER), https://en.wikipedia.org/wiki/BLEU (BLEU), https://en.wikipedia.org/wiki/METEOR (METEOR), https://github.com/grammarly/gr-parseval (PARSEVAL), https://github.com/grammarly/gr-parseval (ROUGE), and https://en.wikipedia.org/wiki/Cohen%27s_kappa (Cohen's Kappa).
[e] Some useful URLs: https://catalog.ldc.upenn.edu/LDC99T42 (Penn Treebank) http://yann.lecun.com/exdb/mnist/index.html (MNIST), http://www.image-net.org/ (ImageNet), http://cocodataset.org/ (COCO), https://visualgenome.org/ (Visual Genome), https://catalog.ldc.upenn.edu/LDC97S62 (Switchboard), https://catalog.ldc.upenn.edu/LDC97S42 (CALLHOME), https://coml.lscp.ens.fr/dihard/index.html (DIHARD), https://en.wikipedia.org/wiki/Standard_Performance_Evaluation_Corporation (SPEC), and https://www.mlperf.org/ (MLPerf).
[f] https://trec.nist.gov/.
[g] http://www.signll.org/conll.
[h] http://www.statmt.org/.
[i] https://www.kdd.org/kdd-cup.
[j] https://www.kaggle.com/.
[k] https://www.kaggle.com/datasets.
[l] https://www.ldc.upenn.edu/.
[m] https://www.nltk.org/book/ch02.html.
[n] https://vincentarelbundock.github.io/Rdatasets/datasets.html.

https://doi.org/10.1017/S1351324919000275 Published online by Cambridge University Press

Table 1. Some popular metrics

Metric	Literature
Accuracy	Part of Speech Tagging, Vision, etc.
L_p-norms	
Mean square error (MSE)	
Precision/recall	Information Retrieval
Mean reciprocal rank (MRR)	Information Retrieval
NDCG	Web Search
Entropy/perplexity	Information Theory/Language Modeling
CIDEr (Vedantam, Zitnick, and Parikh 2015)	Vision
Word error rate (WER)	Speech Recognition
Mean opinion score (MOS)	Speech Synthesis
Diarization error rate (DER)	Speech Diarization
BLEU (Papineni *et al.* 2002)	Machine Translation
METEOR (Banerjee and Lavie 2005)	Machine Translation
PARSEVAL	Parsing
ROUGE (Lin 2004)	Text Summarization, Machine Translation, NLP
Cohen's Kappa	Inter-Annotator Agreement (IAA)

Table 2. Some popular tasks/benchmarks

Task	Literature
Penn Treebank (Marcus, Marcinkiewicz, and Santorini 1993)	Parsing
MNIST (LeCun *et al.* 1998)	Vision
ImageNet (Russakovsky *et al.* 2015)	Vision
COCO (Lin *et al.* 2014)	Vision
Visual Genome (Krishna *et al.* 2016)	Vision
Switchboard (Godfrey, Holliman, and McDaniel 1992)	Speech recognition
CallHome (Canavan, Graff, and Zipperlen 1997)	Speech recognition
DIHARD	Diarization (speech)
SPEC	Hardware systems
MLPerf	Hardware systems

for a little work of a different kind. We founded SIGDAT to provide a forum for this kind of work. SIGDAT started as a relatively small Workshop on Very Large Corpora in 1993 and later evolved into the larger EMNLP Conferences. At first, the SIGDAT meetings were very different from the main ACL conference in many ways (size, topic, geography), but over the years, the differences have largely disappeared. It is nice to see the field come together as it has, but we may have been too successful. Not only have we succeeded in making room for what we were interested in, but now there is no longer much room for anything else."

https://doi.org/10.1017/S1351324919000275 Published online by Cambridge University Press

In Church (2011), we point out the clear shift from rationalism to empiricism. In the early 1990s, it was unusual to see a paper with an evaluation section, and a decade later, it was unusual to see a paper without an evaluation section.

The switch to empiricism came later in other fields. Machine Learning has only recently become empirical. It was not that many years ago when NIPS (now known as NeurIPS)[o] was mostly theoretical.

2.2 Shared tasks and information retrieval

Other fields were always empirical. Information Retrieval (SIGIR)[p] has been promoting empiricism and shared tasks since the 1960s.[q] Historically, Information Retrieval grew out of a combination of library schools and computer science. Shared tasks in Information Retrieval typically start with a library of documents and relevance judgments (pairs of test queries and relevant documents). Ranking systems take a query as input and sort the documents in the library by relevance. Ranking systems are scored by precision (minimize errors of commission) and recall (minimize errors of omission). A single point in precision–recall space is computed by placing a threshold on the ranked output from a system and computing precision and recall for the documents above the threshold. We compute a precision–recall curve by sweeping the threshold over all possible thresholds.

Precision and recall are related to ROC (Receiver operating characteristic) curves[r] which played an important role in the development of radar during World War II. It is well known that one can trade off Type I and Type II errors (errors of omission and errors of commission) in uninteresting ways. When proposing a new method, we would like to claim that the proposed method dominates the state of the art (SOTA)[s] over all trade-offs of Type I and Type II errors. That is, the precision/recall curve for the proposed method should be above the curve for the SOTA method.

One could imagine that some methods might be better for high precision, and other methods might be better for high recall. In principle, it is possible for such curves to cross one another, although that does not happen much in practice. It is often convenient to reduce the two quantities (precision and recall) down to a single figure of merit such as F-measure,[t] equal error rate (EER),[u] MRR,[v] mean average precision (MAP), area under the curve (AOC), so on.[w] Different communities prefer different figures of merit: EER is popular in speaker identification, and F is popular in Information Retrieval.

With a single figure of merit, it is easy to create a leaderboard, reporting a list of candidate systems sorted by the figure of merit. Technically, in order to justify sorting systems in this way, the metric ought to be a proper distance metric[x] supporting the triangle inequality. It is common practice, nevertheless, to sort systems by all sorts of metrics. Leaderboards ought to make it clear which differences are significant and which are not, though leaderboards rarely report error bars.

2.3 Post hoc judging: Is there a single correct answer?

It is common in Information Retrieval to consider multiple answers as correct (relevant). But in many other applications, it is often assumed that each question has a single correct answer.

[o] https://nips.cc/.
[p] https://sigir.org/.
[q] http://sigir.org/resources/museum/.
[r] https://en.wikipedia.org/wiki/Receiver_operating_characteristic.
[s] https://en.wikipedia.org/wiki/State_of_the_art.
[t] https://en.wikipedia.org/wiki/F1_score.
[u] https://www.webopedia.com/TERM/E/equal_error_rate.html.
[v] https://en.wikipedia.org/wiki/Mean_reciprocal_rank.
[w] https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval).
[x] https://en.wikipedia.org/wiki/Metric_(mathematics).

https://doi.org/10.1017/S1351324919000275 Published online by Cambridge University Press

Figure 1. There may not be a single unique correct label. Candidate labels: baseball cap, cap, green hat, hat, and head. Can you guess which one is in the gold standard?

The Visual Genome website,[y] for example, offers more than 100k images with millions of labels (Krishna *et al.* 2016). This data has been used to test and train a number of systems such as Zhang *et al.* (2019). P@1 (precision in top position) is about 50%, but we believe the system is actually performing better than this number suggests because the grading system is overly harsh. This metric counts candidate hypotheses as correct if they match the (single) answer in the gold standard and incorrect otherwise. It is common practice to assume missing judgments are incorrect, but this assumption is problematic, as illustrated in Figure 1. The metric will count just one of the candidate labels (baseball cap, cap, green hat, hat, and head) as correct, since the gold standard provides just one judgment per bounding box.

There are many tasks and metrics that have come up with solutions to this problem. BLEU, for example, makes use of multiple reference translations. It does not make sense, in the case of Machine Translation, to assume that each source sentence has a single reference translation. There are obviously many ways to translate a sentence. Inter-annotator agreement would be awful if we insisted on a single reference translation.

Another popular solution is post hoc judging. That is, instead of judging first and then running systems, we do it the other way around. This way, the judges do not have to judge all possible candidate hypotheses but only candidates proposed by one of the systems. Both TREC[z] and Bing[aa] use post-hoc judging. It is fairly straightforward to use post hoc judging to estimate P@1 (precision in top position) and P@10 (precision in top 10 positions), but recall is harder to estimate, unless one assumes a value for missing judgments.

Missing judgments are a challenge for training as well as evaluation. Learning to rank[ab] was used to train Bing. It is a standard practice to judge as many candidates as the budget will permit,

[y] https://visualgenome.org/.
[z] https://trec.nist.gov/.
[aa] https://www.bing.com/.
[ab] https://en.wikipedia.org/wiki/Learning_to_rank.

https://doi.org/10.1017/S1351324919000275 Published online by Cambridge University Press

and then assume low ratings for missing judgments. There has been considerable work on active learning,[ac] another approach for training with missing labels.

These issues come up for even relatively simple problems like Part of Speech Tagging. One might have thought that each word has one and only one Part of Speech Tag, but there is much more room for differences of opinion than one might have thought. Estimates of inter-annotator agreement rates show that judges agree with one another about as often as systems agree with judges. Does this mean that systems are performing as well as people? Can these systems pass the Turing Test?

The problem is that there is a big difference between a difference of opinion and an error. When two judges disagree, it is a difference of opinion, but when the computer differs from a judge, it is a computer error. It is hard to describe the difference, but computers make mistakes that no person would ever make.

"I know it when I see it."[ad]

Progress in Part of Speech Tagging has been held back for decades, because we do not know how an automated evaluation could differentiate differences of opinion from computer errors (Church 1992). Post hoc judging (also known as human-in-the-loop)[ae] could help the field make progress on Part of Speech Tagging (and many other tasks such as visual genome). That is, when the computer disagrees with the gold standard, one could send the two candidates to human judges (mechanical turkers). Unfortunately, the business case for Part of Speech Tagging is not as compelling as other tasks such as web search, and therefore, it has been difficult to raise funding to improve the gold standard for Part of Speech Tagging.

Similar problems come up in evaluations of speech recognition. A number of companies are reporting WERs on the Switchboard corpus (Godfrey *et al.* 1992) that are within the range of human disagreement rates.[af] That said, one probably would not want to conclude that these machines are as good as people. This blog[ag] concluded quite sensibly that *the real test is using it.*

There has been a trend for publications to report better and better numbers, but are we really making that much progress, or are we fooling ourselves? Sometimes the numbers are too good to be true, and sometimes the truth is better than the numbers. Sometimes the problem is not with the numbers but with the interpretation. It is one thing for an evaluation to fail to find a difference between man and machine, and quite another thing to pass the Turing Test.

3. Meta-considerations and metrics of metrics

3.1 Content-free debates versus mindless metrics

Before evaluations were taken seriously, there was little agreement about what mattered. There were many (unpleasant) debates in Theoretical Linguistics and Computational Linguistics. Facts did not seem to matter as much as personalities. Chomsky would turn his back on his former students. Feelings were hurt.[ah]

Each school had its champions. ACL reviewers gave high grades to papers from their school and low grades to papers from other schools. Information Retrieval was different. The numbers were taken more seriously than personalities. Admittedly the field was small and ingrown. There

[ac] https://en.wikipedia.org/wiki/Active_learning.
[ad] https://en.wikipedia.org/wiki/I_know_it_when_I_see_it.
[ae] https://en.wikipedia.org/wiki/Human-in-the-loop.
[af] https://www.theverge.com/2016/10/18/13326434/microsoft-speech-recognition-human-parity.
[ag] https://thenewstack.io/speech-recognition-getting-smarterstate-art-speech-recognition/.
[ah] https://en.wikipedia.org/wiki/Linguistics_wars.

https://doi.org/10.1017/S1351324919000275 Published online by Cambridge University Press

was a time when almost everyone at SIGIR was a student or a grand-student of Gerald Salton[ai] or Karen Spärck-Jones.[aj] But even so, there was much more agreement in Information Retrieval than Computational Linguistics about what was important and what was not. If the evaluation section reported good numbers, the reviewers would give the paper high grades, no matter which school the paper came from.

Computational Linguistics is now more like Information Retrieval used to be. Reviewers tend to grade papers more on numbers (and English) than insight. While most people would agree that our field is better off with evaluation than without, there is a danger that content-free debates could be replaced with mindless metrics. There are a number of ways to address the mindless concern: reproducibility, reliability, validity, and insight.

3.2 Reproducibility, reliability, validity, generalization, and insight

While there is much to be said in favor of evaluation, much of the reason for the success of Minsky and Chomsky's rebellion against 1950s-style empiricism was frustration with overly burdensome methodology and a lack of insight. The experimental methods of the 1950s were too inflexible for the let-it-all-hang-out 1970s.[ak] Currently popular evaluation methods may suffer from the same concerns that led to the demise of 1950s-style empiricism: burdensome methodology and lack of insight. We discussed the lack of insight in Church (2017):

> "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."

On the other hand, what passes for evaluation these days may not stand up well to the rigorous demands of the 1950s.[al] While we have made some progress toward reproducibility by sharing code and data (Wieling *et al.* 2018), our current metrics may not stand up well to the kind of scrutiny practiced in experimental psychology. Reproducibility is far from reliability and validity. See Pittenger (1993) for an example of a criticism of popular (though probably flawed) personality tests on the basis on reliability and validity.

Many of our methods are probably exposed to similar criticisms. Section 4.2 will discuss Godfrey's Gap. Jack Godfrey (personal communication) observed a large gap between performance of systems on standard academic bake-offs and performance on real tasks of interest to our sponsors (typically in government and industry). Funding agencies have attempted to address this gap by encouraging work on domain adaptation, surprise languages,[am] low resources (and even zero resources).[an]

What our sponsors really care about is how well our solutions are likely to generalize to their problems. Too many of our evaluations are too specific to a specific task and a specific corpus. We can warn the sponsors that *their mileage may vary*,[ao] but that's a pretty lame excuse. Our evaluations ought to address the sponsors' concerns, which are not unreasonable.

[ai] https://en.wikipedia.org/wiki/Gerard_Salton.
[aj] https://en.wikipedia.org/wiki/Karen_Sparck_Jones.
[ak] https://www.youtube.com/watch?v=XWN65nAkk20.
[al] https://en.wikipedia.org/wiki/Experimental_psychology.
[am] http://universaldependencies.org/conll17/surprise.html.
[an] https://zerospeech.com/2017/.
[ao] https://en.wiktionary.org/wiki/your_mileage_may_vary.

https://doi.org/10.1017/S1351324919000275 Published online by Cambridge University Press

3.3 Evaluations in systems

Similar concerns can be found in many other literatures. Consider hardware and software systems. During the let-it-all-hang-out 1970s, the field was dominated by charismatic personalities like Steve Jobs,[ap] Bill Gates,[aq] Gordon Moore,[ar] Seymour Cray,[as] and many others. These people were extremely successful and produced many useful (rationalist) insights such as Moore's Law.[at] Nevertheless, about the same time that empirical evaluations started to take off in Computational Linguistics, empirical evaluations also took off in systems. These days, data are hot and personalities are not. Amazon has a particular (not nice) acronym for personalities: HiPPO.[au] The point is that data ought to trump opinions.

SPEC[av] was founded in 1988. The second author had the distinct pleasure of sitting in on an interview of John Mashey,[aw] a founder of SPEC, while helping to prepare the new Machine Learning benchmark suite, MLPerf.[ax]

In the late 1980s, there were lots of hardware options: Intel, Commodore, VAX, DEC Alpha, MIPS, HP, IBM, SGI, Sun. It was not easy for potential buyers to make sensible choices. First, they needed to decide which applications they cared about, because some machines are more appropriate for some applications and other machines are more appropriate for other applications. Even if the buyers know which applications they care about, vendors might not provide numbers for those applications. Some vendors provide some numbers for some applications, but not all vendors provide numbers for all applications. Even when numbers are available, it is not clear how well those numbers will generalize to other applications. It was likely that vendors invested significant effort to obtain impressive results that may or may not generalize to what the buyer really cares about, and the buyer may be less motivated than the vendor (and less capable).

Mashey explained that too many numbers could make a buyer's head spin. It would help if vendors could be persuaded to report comparable numbers on a particular application. Then we could produce a simple rank ordering (like modern leaderboards in Machine Learning and Computational Linguistics).

But we need more than this. Different buyers care about different applications. Hardware vendors need to produce general purpose solutions because of Moore's Law and economies of scale. Special purpose solutions rarely survive the test of time because general purpose solutions can sell more units to more markets, and consequently, they can afford to invest more on future improvements. In short, it is important to cover a spectrum of different applications. Mashey and colleagues designed the SPEC benchmark suite as a broad collection of the important applications, spanning integer applications (compilers/interpreters, compression, graph traversals) as well as floating point (physical simulations, numerical methods, image processing) in the 1990s and 2000s. Later in the 2010s, applications from databases and parallel computing were added.

The key to SPEC's evaluation was to aggregate the results in a simple manner: performance was measured as speed-ups relative to a baseline processor (e.g., a 1997 Sun Ultrasparc server for SPEC2006), and vendors reported a single number, the geometric mean of these speed-ups. Speed-ups are ratios, so the geometric mean worked best; an arithmetic mean of run times would favor long-running applications. This form simplified the buyer's effort to decide whether they should look more closely at a particular system's performance. Further, the vendors could report

[ap] https://en.wikipedia.org/wiki/Steve_Jobs.
[aq] https://en.wikipedia.org/wiki/Bill_Gates.
[ar] https://en.wikipedia.org/wiki/Gordon_Moore.
[as] https://en.wikipedia.org/wiki/Seymour_Cray.
[at] https://en.wikipedia.org/wiki/Moore%27s_law.
[au] http://www.enricdurany.com/agile-startup-entrepreneur/hippos-highest-paid-person-opinion-data-driven-decision-making/.
[av] https://en.wikipedia.org/wiki/Standard_Performance_Evaluation_Corporation.
[aw] https://en.wikipedia.org/wiki/John_Mashey.
[ax] https://mlperf.org/.

https://doi.org/10.1017/S1351324919000275 Published online by Cambridge University Press

results on each individual benchmark if the buyer wanted a deeper look. The common agreement on benchmarks meant the buyer would be able to compare directly between systems.

A level deeper, we can interpret the geometric mean speed-up as a measure of the general-purposeness of a system—its ability to "adapt" and perform well on many different kinds of applications. The systems and hardware community has broadly adopted these evaluation criteria over the last 25 years, and maybe not surprisingly, MLPerf's system run time characterization is in the process of adopting this same evaluation scheme.

This way of thinking is about to make its way into the Machine Learning community with the creation of MLPerf. Machine Learning is beginning to experience some of the same kinds of economies of scale as mentioned above. Just as the hardware community appreciated the value of general purpose solutions, so too, there may be advantages to designing general purpose networks that do not need as much training for each specific task. We are already beginning to see models such as ELMo (Peters *et al.* 2018), BERT (Devlin *et al.* 2018), and GPT (Radford *et al.* 2019) where a single general purpose model is trained for a range of different tasks. Some fine-tuning may be useful for particular tasks, but even so, there seems to be a promising possibility that one can train a single general purpose model extremely well, and the cost of doing so can be amortized over a range of applications.

Historically, chip design was so expensive (in terms of time and money) that the business case required amortizing that cost over a range of applications. So too, the cost of training big Machine Learning models over very large corpora might become easier to justify if we move toward general purpose designs.

There might be an intriguing analog in the recent growth and development of multi-task learning and fine-tuning in language domains. In particular, as techniques like ELMo (Peters *et al.* 2018), BERT (Devlin *et al.* 2018), and GPT (Radford *et al.* 2019) have ramped up the use of pre-trained contextualized embeddings and attentional transformers, these models have been fine-tuned and applied to numerous language-understanding tasks. Papers report tables of results with a column for each individual task. Like hardware buyers in the 1990s, one's head might start spinning wondering what each result means. We probably care most about the general-purposeness of these language-understanding techniques.

3.4 Declaring success is not a formula for success

Evaluation was always important in Information Retrieval, but Computational Linguistics was different. When the first author was in graduate school, there was an Artificial Intelligence (AI) boom. Funding was easy to come by, no matter what we did. We believed that our systems worked better than they did.

Before writing one of the early papers on Part of Speech Tagging (Church 1988), we were told that not only had Part of Speech Tagging been solved, but all the problems in syntax had been declared solved, as well. We were told to go work on difficult problems in pragmatics, because no one had declared success on those problems (yet).

The field had painted itself into a corner. Over the years, the field had gone to the funding agencies and proposed to do more than what was promised in the last proposal. At the end of each round of funding, all the problems in the last proposal would be declared solved, and the field would attack something even more ambitious. This pyramid system worked for a while (during boom times), but eventually led to a bust.

It was scary to get up at an ACL meeting in 1988 and talk about Part of Speech Tagging. The other papers at the conference were addressing more difficult tasks. How can we talk about Part of Speech Tagging when parsing had been declared solved?

As it turned out, the reaction was surprisingly positive. The field appreciated a solution that actually worked and could stand up to credible evaluation. Everyone knew what everyone knew

https://doi.org/10.1017/S1351324919000275 Published online by Cambridge University Press

(but was not saying); the pyramid system was not good for the field. Declaring success is not a formula for success.

> "the first principle is that you must not fool yourself – and you are the easiest person to fool."[ay]

When Eugene Charniak won the ACL Lifetime Achievement Award, he started his talk with a timeline of his career. In the middle of this timeline, he wrote an "S." His career split neatly into about two equal intervals, the period before statistics and the period after. His talk would focus on the second part, because the first part was just BS. Unfortunately, this part of his talk did not make it into the written version (Charniak 2011).

4. Extrapolating from the past to the future

4.1 A pendulum swung too far

When we revived empiricism in the 1990s, we were well aware of the empiricism from the 1950s. In fact, the first author grew up with empiricism all around him. His father studied at Harvard in the 1950s when Skinner was there, and taught Behaviorism until he retired just a few months ago.[az] Some people ask if the first author has been in a Skinner box; the answer is, "yes."

We suggested in Church (2011) that the shift from Rationalism to Empiricism mentioned above was part of a larger cycle where the field oscillates back and forth every 20 years between Empiricism and Rationalism, based on a simple mechanism.

> "The reason grandparents and grandchildren get along so well is that they have a common enemy."[ba]

Just as Chomsky and Minsky rebelled against the previous generation, our generation returned the favor by rebelling against them in the 1990s.

(1) 1950s: Empiricism (Shannon, Skinner, Firth, Harris).
(2) 1970s: Rationalism (Chomsky, Minsky).
(3) 1990s: Empiricism (IBM Speech Group, AT&T Bell Labs).
(4) 2010s: Rationalism?

The first author studied at MIT in the 1970s and 1980s and took classes from both Minsky and Chomsky. Chomsky was a student of Zellig Harris. Harris's distributional hypothesis[bb] is nicely described by Firth's famous quote, "You shall know a word by the company it keeps."[bc] Chomsky rejected empiricism and the use of statistics and distributional evidence and much more as mere performance. He preferred to focus his attention on linguistic competence.[bd]

Our (incorrect) prediction that the pendulum would swing back to Rationalism by now was not exactly a prediction, but more of a plea for inclusiveness. Computational Linguistics used to be an interdisciplinary combination of Humanities and Engineering, with more Humanities in Europe and more Engineering in Asia. As the field took a hard turn toward Empiricism in the 1990s,

[ay] https://en.wikiquote.org/wiki/Richard_Feynman.
[az] https://scholar.google.com/citations?user=8ZhR8sYAAAAJ&hl=en.
[ba] https://www.brainyquote.com/quotes/sam_levenson_100238.
[bb] https://aclweb.org/aclwiki/Distributional_Hypothesis.
[bc] https://en.wikipedia.org/wiki/John_Rupert_Firth.
[bd] https://en.wikipedia.org/wiki/Linguistic_competence.

https://doi.org/10.1017/S1351324919000275 Published online by Cambridge University Press

we have gained new interdisciplinary connections to Machine Learning,[be] but the connections to Linguistics and Humanities are no longer as strong as they used to be. We would be better off if we could find ways to work together. There has been way too much talk about firing linguists.[bf]

The revival of empiricism in the 1990s was not merely a way for our generation to do onto our teachers as they had done onto their teachers. In Church (2011), we suggested a more practical (pragmatic) motivation:

> "What motivated the revival of empiricism in the 1990s? What were we rebelling against? The revival was driven by pragmatic considerations. The field had been banging its head on big hard challenges like AI-complete problems and long-distance dependencies. We advocated a pragmatic pivot toward simpler more solvable tasks like Part of Speech Tagging. Data was becoming available like never before. What can we do with all this data? We argued that it is better to do something simple (than nothing at all). Let's go pick some low hanging fruit. Let's do what we can with short-distance dependencies. That won't solve the whole problem, but let's focus on what we can do as opposed to what we can't do. The glass is half full (as opposed to half empty)."

This plea for inclusiveness ended by arguing that we ought to teach the next generation both Empiricism and Rationalism because it is likely that the next generation will have to take Chomsky's objections more seriously than we have. Our generation has been fortunate to have plenty of low-hanging fruit to pick (the facts that can be captured with short *n*-grams), but the next generation will be less fortunate since most of those facts will have been pretty well picked over before they retire, and therefore, it is likely that they will have to address facts that go beyond the simplest *n*-gram approximations.

As it turned out, with the rise of deep nets (and end-to-end systems), the next generation is learning more new things (e.g., neural nets) and fewer old things (e.g., generative capacity). Chomsky's concerns about long distance dependencies were closely related to his work on generative capacity and the Chomsky Hierarchy.[bg] It is often suggested that modern nets such as LSTMs[bh] can capture the kinds of long distance dependencies that Chomsky was interested in, though it is not clear if this is correct in theory, or in practice (Daniluk *et al.* 2017).

4.2 Godfrey's gap and fooling ourselves

There is, perhaps, too much satisfaction with the latest swing of the pendulum. There are many reasons for Godfrey's gap, but a big problem is a common (but unrealistic) assumption that the test set and the training set are drawn from the same population. In practice, we train the system as best we can. But at training time, it is hard to know what the users will do with the system (in the future). We have to train with data that are available at train time, but it is likely that users will use the system on data that was not available at training time (perhaps because of privacy considerations, or perhaps because it did not exist at training time, among other things). Language is a moving target. Topics change quickly over time. Tomorrow's news will not be the same as yesterday's news. Tomorrow's kids will invent new ways to use social media (that their parents could never have anticipated).

The assumption that the training data are representative of the test data is somewhat similar to a request that professors often hear from students. *What is going to be on the exam?* Students want

[be] https://www.earningmyturns.org/2017/06/a-computational-linguistic-farce-in.html.
[bf] http://www.lrec-conf.org/lrec2004/doc/jelinek.pdf.
[bg] https://en.wikipedia.org/wiki/Chomsky_hierarchy.
[bh] https://en.wikipedia.org/wiki/Long_short-term_memory.

https://doi.org/10.1017/S1351324919000275 Published online by Cambridge University Press

to know how they can do well on the exam (without mastering the material). That is only natural, but it is not fair. So too, we should expect our systems to do well on the test without knowing what is going to be on the exam. The assumption that the training data are representative of the test is only natural, but it is not fair. If we cheat on the test, we are only fooling ourselves into believing that our systems are doing better than they are.

4.3 Evaluation: a tough sell

It is hard to remember that evaluation was a tough sell 25 years ago. Charles Wayne restarted funding in the mid-1980s by emphasizing evaluation. No other sort of program could have been funded at the time, at least in America. Wayne's emphasis on evaluation helped pull the field out of an AI Winter.[bi]

Wayne's idea was not an easy sell, especially at first. As discussed in a previous emerging trends article (Church 2018), Liberman pointed out that "not everyone liked it."

The research community tends to think of funding agencies as the source of funding, but in fact, funding agencies are often middle men, somewhat like real estate agents. Real estate agents do not own the house (either before or after the transaction). They are merely market makers.

Funding agencies like DARPA bring together sellers of technology (researchers) with buyers (department of defense). When Wayne first advocated evaluation, there were objections from both sides of the transaction.

- Buyers: *You can not turn water into gasoline, no matter what you measure.*
- Sellers: *It is like being in first grade again—you are told exactly what to do, and then you are tested over and over.*

But according to Liberman, Wayne's idea eventually succeeded because "it worked." Why did it work? Liberman starts out with the obvious. It enabled funding to start because the project was glamour-and-deceit-proof, and to continue because funders could measure progress over time. Wayne's idea makes it easy to produce plots such as Figure 2[bj] which help sell the research program to potential sponsors.

A less obvious benefit of Wayne's idea is that it enabled hill climbing. Researchers who had initially objected to being tested twice a year began to evaluate themselves every hour. An even less obvious benefit, according to Liberman, was the culture. Participation in the culture became so valuable that many groups joined without funding. As obvious as Wayne's idea may appear to us today, Liberman reminds us that back in 1986, "This obvious way of working was a new idea to many!"

Sometimes the high climbing could be automated. Och (2003a) was disturbing when it was first proposed. When BLEU was first proposed (Papineni *et al.* 2002), it was remarkable that the Machine Translation community could agree on a metric, and that the metric had as much validity as it did,[bk] but no one expected the metric to stand up to automated hill climbing. One would have thought that hill climbing could find a way to game the metric. That is, although the validity experiments in Section 5 of Papineni *et al.* (2002) showed large correlations between human ratings and BLEU scores, at least for a small set of candidate translations that they considered, one might be concerned that hill climbing on a large set of candidate translations would likely find bad translations that happen to score well under BLEU. Fortunately, hill climbing worked remarkably well.

[bi] https://en.wikipedia.org/wiki/AI_winter.

[bj] https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation.

[bk] BLEU has had a huge positive impact on the field. We surveyed Machine Translation metrics before BLEU in Church and Hovy (1993). It is remarkable just how much better BLEU is than what came before it.

https://doi.org/10.1017/S1351324919000275 Published online by Cambridge University Press

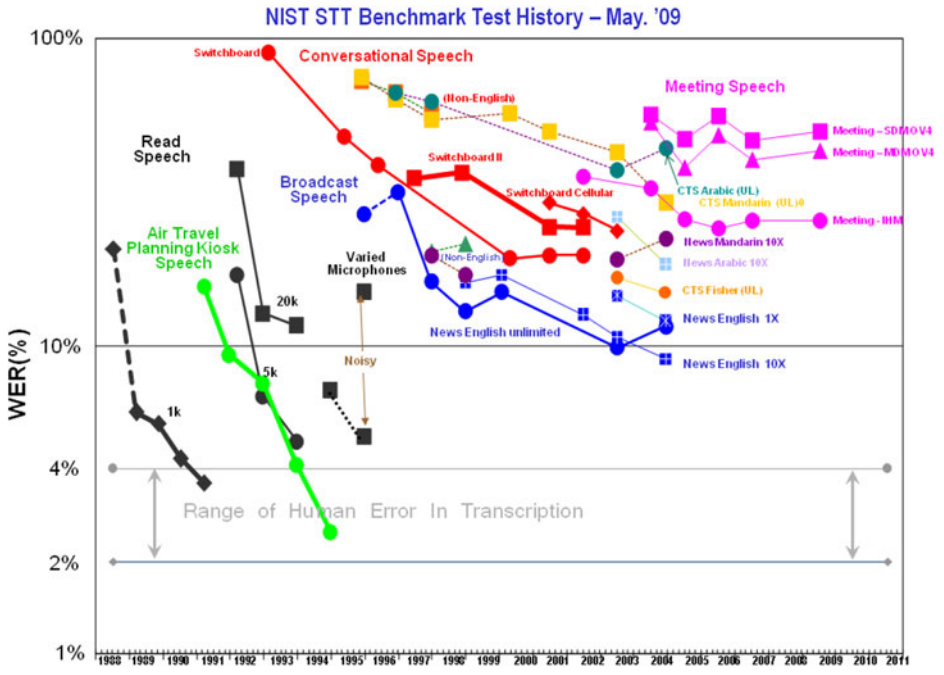

Figure 2. Thirty years of progress in speech recognition.

5. Conclusions

In this survey of 25 years of evaluation, we hoped to not only cover many of the high points, but also call out risks and opportunities for the future. Much of the discussion focused on Computational Linguistics, but also mentioned related work in related fields such as Vision, Psychology, and Systems. Most people would agree that there is more evaluation these days than there was then. Evaluation has been very good for the field. The 1990s revival of empirical methods was an exciting time, but there is even more excitement today. Conferences are considerably bigger today. More and more people are publishing more and more results than ever before. It is hard to remember that evaluation was a tough sell 25 years ago. At the time, there were too many unproductive content-free debates with HiPPOs. We are better-off these days, now that data trumps opinions.

On the other hand, there is always a risk that our evaluations could become mindless metrics. We discussed a number of meta-considerations (metrics of metrics): reproducibility, reliability, validity, and generalization. Our field has made considerable progress on reproducibility. Many papers these days refer to standard corpora and share code and pre-trained models on GitHub. That said, there are always opportunities to borrow insights and best practices from other fields. Consider validity, for example. It is much more common to see a discussion of validity in Psychology than in our field, though a nice exception is the seminal paper on BLEU (Papineni *et al.* 2002).

A second example is the emphasis on general purpose solutions in Systems. MLPerf will encourage more and more general purpose solutions in Machine Learning. BERT and other models are showing how training costs can be amortized over a range of applications. Too much of the work in our field is tied to too many specifics. We ought to do a better job of addressing Godfrey's Gap. Our evaluations ought to be more insightful than they are in helping sponsors understand how well proposed methods will generalize to real applications that matter.

https://doi.org/10.1017/S1351324919000275 Published online by Cambridge University Press

References

Banerjee S. and Lavie A. (2005). METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72.

Black A. and Tokuda K. (2005). The Blizzard Challenge-2005: evaluating corpus-based speech synthesis on common datasets. In *INTERSPEECH*, pp. 77–80.

Canavan A., Graff D. and Zipperlen G. (1997). CALLHOME American English Speech LDC97S42, Web Download. Philadelphia, PA, USA: Linguistic Data Consortium, University of Pennsylvania.

Charniak E. (2011). ACL Lifetime Achievement Award: the brain as a statistical inference engine—and you can too. *Computational Linguistics* **37**(4), 643–655. MIT Press, Cambridge, MA, USA.

Church K. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing*, Austin, Texas, ACL, pp. 136–143.

Church K. (1992). Current practice in part of speech tagging and suggestions for the future, In *Sbornik praci: In Honor of Henry Kucera*, Michigan Slavic Studies. Ann Arbor, MI, USA: University of Michigan, pp. 13–48.

Church K. and Hovy E. (1993). Good applications for crummy machine translation. *Machine Translation* **8**(4), 239–258. Springer.

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. Cambridge University Press.

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

Daniluk M., Rocktäschel T., Welbl J. and Riedel S. (2017). Frustratingly short attention spans in neural language modeling. In *ICLR*, April 24-26, Toulon, France

Devlin J., Chang M.-W., Lee K. and Toutanova K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805.

Godfrey J., Holliman E. and McDaniel J. (1992), SWITCHBOARD: telephone speech corpus for research and development. In *ICASSP*. Washington, DC, USA: IEEE Computer Society, pp. 517–520.

Goodman J. (1996). Parsing algorithms and metrics. In *ACL*, pp. 177–183.

Krishna R., Zhu Y., Groth O., Johnson J., Hata K., Kravitz J., Chen S., Kalantidis Y., Jia-Li L., Shamma D.A., Bernstein M.S. and Li F.-F. (2016) Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations, https://arxiv.org/abs/1602.07332.

LeCun Y., Bottou L., Bengio Y. and Haffner P. (1998). Gradient-based learning applied to document recognition. *Proceedings of IEEE* **86**(11), 2278–2324.

Lin, C.-Y. (2004). Rouge: a package for automatic evaluation of summaries. In *Text Summarization Branches Out, ACL Workshop*.

Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P. and Zitnick L. (2014). Microsoft coco: common objects in context. In *European Conference on Computer Vision*. Springer, pp. 740–755.

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

Och, F.J. (2003). Minimum error rate training in statistical machine translation. In *ACL*, pp. 160–167.

Och, F.J. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics* **29**(1), 19–51. MIT Press, Cambridge, MA, USA.

Papineni K., Roukos S., Ward T. and Zhu W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318.

Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K. and Zettlemoyer L. (2018). Deep contextualized word representations. In *NAACL*.

Pittenger D.J. (1993). The utility of the Myers-Briggs type indicator. *Review of Educational Research* **63**(4), 467–488.

Radford A., Wu J., Child R., Luan D., Amodei D. and Sutskever I. (2019). Language Models are Unsupervised Multitask Learners, OpenAI Blog. https://openai.com/blog/better-language-models/.

Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A.C. and Fei-Fei L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* **115**(3), 211–252. Springer.

Snover M., Dorr B., Schwartz R., Micciulla L. and Makhoul J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pp. 223–231.

Vedantam R., Zitnick L. and Parikh D. (2015). CIDEr: consensus-based image description evaluation. In *CVPR*.

https://doi.org/10.1017/S1351324919000275 Published online by Cambridge University Press

Wieling M., Rawee J. and van Noord G. (2018). Reproducibility in computational linguistics: are we willing to share? *Computational Linguistics* **44**(4), 641–649. MIT Press, Cambridge, MA, USA.

Wang Y., Wang L., Li Y., He D., Liu T.-Y. and Chen W. (2013). A theoretical analysis of NDCG type ranking measures. In *COLT*.

Zhang J., Kalantidis Y., Rohrbach M., Paluri M., Elgammal A. and Elhoseiny M. (2019) Large-scale visual relationship understanding. In *AAAI*.

Cite this article: Church KW and Hestness J (2019). Emerging trends: A survey of 25 years of evaluation. *Natural Language Engineering* **25**, 753–767. https://doi.org/10.1017/S1351324919000275