Benchmarking: Past, Present and Future

Kenneth Church
Baidu, CA, USA

Mark Liberman
University of Pennsylvania, PA, USA

Valia Kordoni
Humboldt-Universitaet zu Berlin, Germany

Abstract

Where have we been, and where are we going? It is easier to talk about the past than the future. These days, benchmarks evolve more bottom up (such as papers with code).[1] There used to be more top-down leadership from government (and industry, in the case of systems, with benchmarks such as SPEC).[2] Going forward, there may be more top-down leadership from organizations like MLPerf[3] and/or influencers like David Ferrucci[4]. Tasks such as reading comprehension become even more interesting as we move beyond English. Multilinguality introduces many challenges, and even more opportunities.

1 Abstracts for Invited Talks

We have an amazing collection of invited speakers that can share with us first hand knowledge of how benchmarking became important in Information Retrieval, and then in speech (starting around 1975), and then in language (in 1988). Much of this history is described in this video[6] and two 2016 Interspeech keynotes: Makhoul describes how benchmarking overcame resistance in speech in this keynote,[7] and Jurafsky describes how this approach moved from speech to language in this keynote.[8]

Web site for workshop is here[9]

1.1 What Will it Take to Fix Benchmarking in Natural Language Understanding?

Sam Bowman
New York University
https://cims.nyu.edu/~sbowman/
https://twitter.com/sleepinyourhat

Evaluation for many natural language understanding (NLU) tasks is broken: Unreliable and biased systems score so highly on standard benchmarks that there is little room for researchers who develop better systems to demonstrate their improvements. The recent trend to abandon IID benchmarks in favor of adversarially-constructed, out-of-distribution test sets ensures that current models will perform poorly, but ultimately only obscures the abilities that we want our benchmarks to measure. In this position paper, we lay out four criteria that we argue NLU benchmarks should meet. We argue most current benchmarks fail at these criteria, and that adversarial data collection does not meaningfully address the causes of these failures. Instead, restoring a healthy evaluation ecosystem will require significant progress in the design of benchmark datasets, the reliability with which they are annotated, their size, and the ways they handle social bias.

1.1.1 Bio

Sam Bowman has been on the faculty at NYU since 2016, when he completed PhD with Chris Manning and Chris Potts at Stanford. At NYU, he is a member of the Center for Data Science, the Department of Linguistics, and Courant Institute's Department of Computer Science. His research focuses on data, evaluation techniques, and modeling techniques for sentence and paragraph

[1] https://paperswithcode.com/

[2] https://www.spec.org/benchmarks.html

[3] https://mlperf.org/

[4] https://en.wikipedia.org/wiki/David_Ferrucci, who was responsible for IBM's success with Jeopardy,[5] and has recently written a paper suggesting how the community should think about benchmarking for machine comprehension (Dunietz et al., 2020)

[6] https://www.simonsfoundation.org/search/liberman/

[7] https://www.superlectures.com/interspeech2016/isca-medalist-for-leadership-and-extensive-contributions-to-speech-and-language-processing

[8] https://www.superlectures.com/interspeech2016/ketchup-interdisciplinarity-and-the-spread-of-innovation-in-speech-and-language-processing

[9] https://github.com/kwchurch/Benchmarking_past_present_future

Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, pages 1–7
August 5–6, 2021.

understanding in natural language processing, and on applications of machine learning to scientific questions in linguistic syntax and semantics. He is the senior organizer behind the GLUE and SuperGLUE benchmark competitions; he organized a twenty-three-person research team at JSALT 2018; and he received a 2015 EMNLP Best Resource Paper Award, a 2019 *SEM Best Paper Award, and a 2017 Google Faculty Research Award.

1.2 Context for Interpreting Benchmark Performances

Eunsol Choi

Interpreting benchmark results requires a more nuanced study than simply comparing a single number (e.g., accuracy). For example, higher performance on benchmark focusing on multi-hop reasoning does not translate to model architecture focusing on multi-hop reasoning but often a bigger pretrained model. In the first half of the talk, I will discuss the nuances of interpreting benchmark results, and our previous efforts in integrating highly relevant axis, computational resources, into evaluation. In the second half of the talk, I will talk about the issues with the static benchmarks in the evolving world. Unlike traditional benchmarks which mostly targeted linguistic knowledge, modern benchmark embraces common sense, social context, and encyclopedic world knowledge into the task definition. All these components change over time, urging NLP benchmarks to be refreshed.

1.2.1 Bio

Eunsol Choi is an assistant professor in the computer science department at the University of Texas at Austin. Her research focuses on natural language processing, various ways to recover semantics from unstructured text. Prior to UT, she was a visiting faculty researcher at Google AI. She received a Ph.D. from the University of Washington (with Luke Zettlemoyer and Yejin Choi) and an undergraduate degree in mathematics and computer science from Cornell University. She is a recipient Facebook Research Fellowship, Google Research Award and has co-organized many workshops related to question answering at NLP and ML venues.

1.3 Moving out of the comfort zones: desired shifts in NLP benchmarking

Ido Dagan
Bar-Ilan University
https://u.cs.biu.ac.il/~dagan/

As the deep-learning era has transformed the NLP field, benchmarking practices haven't changed that much, often addressing earlier language analysis tasks and applications. While performance on many benchmarks rocketed, mostly in deep learning comfort zones, profound language technology is still a long way ahead. In this talk, I will argue for three desired interrelated shifts in NLP benchmarking, which motivate and support each other, that should direct further research.

First, much more emphasis should be given to typical realistic settings, in which large training data for the target task is not available, like few-shot and transfer learning. Moreover, benchmarks design should fit realistic data compositions, rather than synthetic ones within the comfort zone, as I will illustrate by a recent few-shot relation classification dataset. Second, recognizing the limits of foreseeable fully-automated methods in addressing the hard NLP challenges, I suggest developing principled evaluation methodologies for various interactive NLP settings. Interaction may lead to better results, with the help of a human in the loop, and moreover allow personalized and explorative behavior, as I will demonstrate with a recent framework for evaluating interactive summarization. Lastly, while many current models operate in an end-to-end manner over implicit language structures, I argue that it is pertinent to pursue also explicit representations for textual information structure, to facilitate refined and better-controlled modeling. Unlike traditional semantic formalisms, I propose pursuing semi-structured representations, consisting of natural language expressions over which current powerful text-embeddings can be applied. I will illustrate this direction by an approach for decomposing the information in single and multiple texts into sets of question-answer pairs, and draw some analogies from our successful experience in designing the Recognizing Textual Entailment (RTE, later aka NLI) task.

1.3.1 Bio

Ido Dagan is a Professor at the Department of Computer Science at Bar-Ilan University, Israel, the founder of the Natural Language Processing (NLP) Lab at Bar-Ilan, the founder and head of the nationally-funded Bar-Ilan University Data Science Institute and a Fellow of the Association for Computational Linguistics (ACL). His interests are in applied semantic processing, focusing on tex-

tual inference, natural open semantic representations, consolidation and summarization of multi-text information, and interactive text summarization. Dagan and colleagues initiated textual entailment recognition (RTE, later aka NLI) as a generic empirical task. He was the President of the ACL in 2010 and served on its Executive Committee during 2008-2011. In that capacity, he led the establishment of the journal Transactions of the Association for Computational Linguistics, which became one of two premiere journals in NLP. Dagan received his B.A. summa cum laude and his Ph.D. (1992) in Computer Science from the Technion. He was a research fellow at the IBM Haifa Scientific Center (1991) and a Member of Technical Staff at AT&T Bell Laboratories (1992-1994). During 1998-2003 he was co-founder and CTO of FocusEngine and VP of Technology of LingoMotors, and has been regularly consulting in the industry. His academic research has involved extensive industrial collaboration, including funds from IBM, Google, Thomson-Reuters, Bloomberg, Intel and Facebook, as well as collaboration with local companies under funded projects of the Israel Innovation Authority.

1.4 MLPerf

Greg Diamos, Peter Mattson and **David Kanter**
https://www.anandtech.com/show/14754/hot-chips-31-live-blogs-mlperf-benchmark

Two topics: (1) What is MLPerf? (2) Advice for groups wanting to create new sets of benchmarks.

1.4.1 Bio

Greg is helping build Landing AI, a new company focused on bringing AI to every major industry starting with our first manufacturing visual inspection product, LandingLens. Greg co-founded MLPerf and MLCommons. Greg helped found Baidu's Silicon Valley AI Lab, where he contributed to the DeepSpeech, DeepVoice, and Mixed Precision training systems. Greg contributed the independent thread scheduling system to the NVIDIA Volta GPU.

He holds a Ph.D. in electrical engineering from the Georgia Institute of Technology.

1.5 Really Reaching Human Parity? –Addressing NLP Benchmark Issues on Robustness, Constraint, Bias and Evaluation Metrics

Nan Duan (Microsoft Research Asia)
Qi Zhang (Fudan University)
Ming Zhou (Sinovation Ventures)

We use Machine Reading Comprehension as an example to recap the current status of NLP benchmarks and highlight four key issues with the existing benchmarks including (1) lack of robustness testing on the new independent (but similar) dataset or adversarial inputs, (2) strong constraints on experimental conditions, (3) bias brought by data sampling or human annotation, and (4) lack of suitable evaluation metrics. Then we present our thoughts and experiments on the possible solutions to these challenges from various aspects.

1.6 Machine Understanding in Context

Dave Ferrucci
Founder & CEO, Elemental Cognition
https://ec.ai/
davef@ec.ai

The ability for machines to read, understand and reason about natural language would dramatically transform the knowledge economy across all industries. Today's latest Deep Learning marvels do not understand what they read to the extent required for rational problem solving and transparent decision making. And yet we need machines to read, understand and engage with us at a rational level for us to take responsibility for their predictions.

A potential problem slowing the advancement of natural language understanding may be that we are not ambitiously or rigorously defining what it means to comprehend language in the first place. Current metrics and tests may be insufficient to drive the right results. In this talk, I will present a definition of comprehension and early experimental results that strongly suggest existing systems are not up to the task. I will also demonstrate a system architecture and behavior that reflects the sort of language understanding capabilities we envision would do better to advance the field of NLU.

1.6.1 Bio

Dave Ferrucci is an award-winning Artificial Intelligence researcher who started and led the IBM Watson team from its inception through its landmark Jeopardy success in 2011. Dr. Ferrucci's

more than 25 years in AI and his passion to see computers fluently think, learn, and communicate inspired him to found Elemental Cognition in 2015. Elemental Cognition is an AI company focused on deep natural language understanding. It explores methods of learning that result in explicable models of intelligence and cross-industry applications.

Dr. Ferrucci graduated from Rensselaer Polytechnic Institute with a Ph.D. in Computer Science. He has over 100 patents and publications. He is an IBM Fellow and has worked at IBM Research and Bridgewater Associates directing their AI research. He has keynoted at highly distinguished venues around the world. Dr. Ferrucci serves as a member of the Connecticut Academy of Science and Engineering and an Adjunct Professor of Entrepreneurship and Innovation at the Kellogg School of Management at Northwestern University.

1.7 Rethinking Benchmarking in AI

Douwe Kiela
Facebook AI Research
https://douwekiela.github.io/
@douwekiela on Twitter

The current benchmarking paradigm in AI has many issues: benchmarks saturate quickly, are susceptible to overfitting, contain exploitable annotator artifacts, have unclear or imperfect evaluation metrics, and do not necessarily measure what we really care about. I will talk about our work in trying to rethink the way we do benchmarking in AI, specifically in natural language processing, focusing mostly on the Dynabench platform.

1.7.1 Bio

Douwe Kiela is a Research Scientist at Facebook AI Research, working on natural language processing and multimodal reasoning and understanding. His work has mainly been focused on representation learning, grounded language learning and multi-agent communication. Recently, he has become interested in improving the way we evaluate AI systems.

1.8 The Dawn of Benchmarking

John Makhoul

Benchmarking, or common evaluations, can be traced back to a speech recognition workshop in 1987 that pitted a knowledge- or rule-based method against an automatically trainable method on an evaluation task with a defined corpus. The workshop was part of the DARPA Strategic Computing Program. Deciding on an evaluation metric was a contentious issue that was settled soon after into the currently used word error rate. Program managers at DARPA continued to champion the idea of metrics-based common evaluations with defined training and test corpora and, by inviting international research groups to participate in these annual common evaluations, this benchmarking paradigm took hold and spread to other DARPA programs and internationally. DARPA also provided seed funding for the establishment of the Linguistic Data Consortium, which was instrumental in making common corpora available to the world at large.

1.8.1 Bio

John Makhoul is a Chief Scientist at Raytheon BBN Technologies, Cambridge, MA, where he has been working on various aspects of speech and language processing, including speech analysis and synthesis, speech coding, speech recognition, speech enhancement, artificial neural networks, human-machine interaction using voice, optical character recognition, machine translation, and cross-lingual information retrieval. He is a Fellow of the IEEE, the International Speech Communication Association (ISCA), and the Acoustical Society of America. Makhoul is the recipient of the ISCA medal and several IEEE awards, including the Flanagan medal in speech and audio processing.

1.9 Benchmarking as a Method for Long-Term Research Management: The Common Task Method

Mark Liberman
Linguistic Data Consortium, University of Pennsylvania

Over the course of half a century, DARPA's Human Language Technology program created capabilities such as speech recognition, machine translation, and text understanding, turning them from science fiction fantasies to everyday practical fact. This sustained success was based on the development of the Common Task Method, which allowed decades of incremental progress in advance of commercial viability. I'll describe the origin and (sometimes counter-intuitive) progress of this method, distinguish it from other uses of benchmarking, and speculate about its future.

1.9.1 Bio

Mark Liberman is the Christopher H. Browne Professor of Linguistics at the University of Pennsyl-

vania, with positions in the department of computer science and in the psychology graduate group. He is also founder and director of the Linguistic Data Consortium. Before coming to the University of Pennsylvania, he was head of the linguistics research department at AT&T Bell Laboratories.

1.10 Detection of Dementia from Speech Samples

Brian MacWhinney (Language Technologies and Modern Languages, CMU)
Saturnino Luz (University of Edinburgh)
https://www.research.ed.ac.uk/en/persons/saturnino-luz-filho

Diagnosis or early detection of the onset of dementia is important for interventions and planning for life-style changes. Ideally, we would like to achieve accurate diagnosis based on samples of naturalistic language production, as well as samples ellicited using some standard formats, such as narrative, script reading, or picture description. Currently, research in this area relies primarily on the Pitt Corpus in DementiaBank which includes cookie theft narratives from 104 controls, 208 persons with dementia, and 85 persons with unknown diagnosis. These data were used in the ADReSS challenge for INTERSPEECH2020 and will be used in a new challenge for 2021. The previous challenge used hand-created transcripts. The new challenge focuses on a pipeline that can be applied automatically, using ASR and NLP methods. The four major gaps in the current data set are: 1) we need fuller ancillariy data on cognitive and medical status, 2) we need longitudinal data on progression, 3) we need more data across language task and interaction types, and 4) ideally, we would like to have data recorded in the home with voice assistant technology. Currently, challenge participants are committed to open sharing of algorithms, but we need more sharing of primary language data, including data outside of English.

1.10.1 Bios

Brian MacWhinney is Teresa Heinz Professor of Psychology, Computational Linguistics, and Modern Languages at Carnegie Mellon University. His Unified Competition Model analyzes first and second language learning as aspects of a single basic system. He has developed a series of 13 TalkBank open access online databases for the study of language learning, multilingualism, and language disorders. The databases for language disorders include AphasiaBank, ASDBank, DementiaBank, FluencyBank, RHDBank, and TBIBank. These databases provide transcriptions of spoken language linked to audio and video media, along with programs for analysis and linguistic profiling. His other research topics include methods for online learning of second language vocabulary and grammar, neural network modeling of lexical development, fMRI studies of children with focal brain lesions, ERP studies of between-language competition, and the role of embodied perspectival imagery in sentence processing.

Dr. Luz is a reader in medical informatics at the Usher Institute, Edinburgh medical School. His is interested in the use of computational methods in the study of behavioural changes caused by neurodegenerative diseases, with focus on vocalisation and linguistic behaviour. He has also studied interaction in multidisciplinary medical team meetings, doctor-patient consultations, telemedicine and patient safety.

1.11 Lessons from SPEC

John Mashey
https://en.wikipedia.org/wiki/John_Mashey
https://www.spec.org/benchmarks.html
Twitter:@johnmashey
(Mashey, 2004, 2005)
https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story

In the 1980s, amidst fierce competition among new microprocessor architectures, CPU benchmarking was in poor condition. Many commonly-used benchmarks were small synthetic benchmarks like Whetstone and Dhrystone that poorly-matched realistic programs. Companies sometimes outright cheated by special-casing compilers to recognize major benchmarks. Some vendors honestly reported results from realistic benchmarks, but even when running the same programs, often used different inputs, so that potential customers could not easily make direct comparisons. Many customers did not trust performance claims.

The talk reviews the odd way SPEC got started in 1988, initially by MIPS, Apollo, Hewlett-Packard and Sun, later joined by many others, then covers the ground rules that evolved to let fierce competitors work together successfully to produce benchmarks that became industry standards and exemplars of good methodologies for selecting bench-

marks, validating results, reporting them carefully and deciding when they had to be retired as obsolete for one reason or another.

SPEC of course is still active, 30+ years later. The talk reviews lessons learned about high-stakes benchmarking, evolution of benchmark suites over time, competitor social issues, credibility issues when people think the foxes are guarding the henhouse, as we were asked by a member of the press. From the beginning, SPEC reported performance on a set of benchmarks as a set of ratios versus a base system, so that people could find benchmarks they thought relevant to their own and ignore the others. Many arguments had occurred over summary means, but as had been done in some performance reports, SPEC correctly used the Geometric Mean, but without really delving into the underlying statistics, which only happened in 2004.

A set of benchmark ratios can be viewed as a sample (representative if selected by experts) from a large population of programs. In practice, many sets of benchmark ratios are well-fit by the log-normal distribution, whose mean is the Geometric Mean, but also allows computation of a (Multiplicative) Standard Deviation, Confidence Intervals, etc. The talk briefly reviews the relevant, simple statistics and the rationale for them.

1.11.1 Bio

John Mashey is a semi-retired computer scientist/corporate executive at Bell Labs, Convergent Technologies, MIPS Computer Systems and Silicon Graphics, where he is was originator of the phrase "Big Data" (according to NY Times). He later consulted for venture capitalists, advised startups and occasionally consulted for companies like Nvidia. He is a 20-year Trustee at the Computer History Museum. He was one of the 4 cofounders of the SPEC benchmarking group in 1988 and was asked in 2018 to advise the MLperf benchmarking group on relevant statistics.

1.12 Benchmarking for diarization. Lessons from the DIHARD evaluation series

Neville Ryant
Linguistic Data Consortium, University of Pennsylvania

Recently, there has been renewed interest in speaker diarization – that is, the task of determining "who spoke when" in a recording. With this renewed interest has come major improvements in system performance with error rates for the DIHARD challenge falling by 33in the span of 4 years. However, despite these successes, the goal of truly robust diarization which is resilient to the full range of natural variation in recordings (e.g., conversational domain, recording equipment, reverberation, ambient noise) remains elusive. In this talk we will review the evolution of the state-of-the-art on multiple domains from the DIHARD dataset as well as some challenges we have encountered in attempting to construct a representative diarization benchmark.

1.12.1 Bio

Neville Ryant is a researcher at the Linguistic Data Consortium (LDC) at the University of Pennsylvania, where he has worked on many topics in speech recognition including: forced alignment, speech activity detection, large scale corpus linguistics, computational paralinguistics, and automated analysis of tone. Since 2017, he has been the principal organizer of the DIHARD challenge, the most recent iteration of which (DIHARD III) completed in December 2020.

1.13 5 Ways to Make Your Data More Relevant

Anders Søgaard
University of Copenhagen
https://anderssoegaard.github.io/

This talk briefly summarizes works I've been involved in that propose improvements to how we evaluate our models, e.g., presenting sampling strategies that better simulate real-life scenarios. The talk will be a sort of self help talk with simple, practical advice for how to add value to your existing data.

1.14 Benchmarking and TREC

Ellen Voorhees
National Institute of Standards and Technology

urlhttps://www.nist.gov/people/ellen-m-voorhees

Coopetitions are activities in which competitors cooperate for a common good. Community evaluations such as the Text REtrieval Conference (TREC) are prototypical examples of coopetitions in information retrieval (IR) and have now been a part of the field for thirty years. This longevity and the proliferation of shared evaluation tasks suggest that, indeed, the net impact of community evaluations is positive. But what are these benefits, and what are the attendant costs?

This talk will use TREC tracks as case studies to explore the benefits and disadvantages of different evaluation task designs. Coopetitions can improve state-of-the-art effectiveness for a retrieval task by establishing a research cohort and constructing the infrastructure—including problem definition, test collections, scoring metrics, and research methodology—necessary to make progress on the task. They can also facilitate technology transfer and amortize the infrastructure costs. The primary danger of coopetitions is for an entire research community to overfit to some peculiarity of the evaluation task. This risk can be minimized by building multiple test sets and regularly updating the evaluation task.]

1.14.1 Bio

Ellen Voorhees is a Senior Research Scientist at the US National Institute of Standards and Technology (NIST). Her primary responsibility at NIST is to manage the Text REtrieval Conference (TREC) project, a project that develops the infrastructure required for large-scale evaluation of search engines and other information access technology. Voorhees' research focuses on developing and validating appropriate evaluation schemes to measure system effectiveness for diverse user tasks.

Voorhees is a fellow of the ACM and an inaugural member of the ACM SIGIR Academy. She has published numerous articles on information retrieval techniques and evaluation methodologies and serves on the review boards of several journals and conferences.

1.15 Benchmarks: An Industry Perspective

Hua Wu and **Jing Liu**
Baidu
https://wuhuanlp.github.io/
https://www.machinereading.ai/

In recent years, the researchers from academia created large-scale datasets mainly in a crowdsourcing way, that accelerate the development of NLP technology. However, these datasets might present different distributions and different challenges from the ones in real-world applications. In this talk, we will introduce our efforts on building NLP benchmarks from an industry perspective. Specifically, we will describe our released datasets on the tasks including question answering, dialogue and simultaneous translation that were created to tackle with the problems in industrial applications. We will present the challenges of these datasets and show how these datasets drive the advancements of NLP technologies. Additionally, we will talk about LUGE, which is an Open-Source Project of Chinese NLP benchmarks. LUGE aims to evaluate NLP models in terms of robustness and adaptability across multiple tasks and multiple domains, which are very crucial for their success in industrial applications.

1.15.1 Bios

Hua Wu is the chair of Baidu tech committee and tech leader of Baidu NLP. Before that, she worked for Toshiba (China) R&D center and Microsoft Research Asia. She obtained her Ph.D. degree from Institute of Automation, Chinese Academy of Science in 2001. Her research interests span a wide range of topics including machine translation, dialogue systems, knowledge graph, etc. She was the Program Co-Chair of ACL 2014 and AACL 2020 (Asia-Pacific Chapter of ACL).

Jing Liu is a principal architect and a tech leader of deep question answering team at Baidu NLP since 2017. Before that, he was a researcher at Microsoft Research Asia (MSRA). He obtained Ph.D. degree in computer science from Harbin Institute of Technology (HIT) in 2014. He is interested broadly in natural language processing and information retrieval, with a particular focus on building robust end-to-end question answering system. He published over 30 research papers in prestigious conferences including ACL, EMNLP, NAACL, SIGIR, WSDM, CIKM, etc. He served as an Area Chair in ACL 2021.

References

Jesse Dunietz, Gregory Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and David Ferrucci. 2020. To test machine comprehension, start by defining comprehension. *arXiv preprint arXiv:2005.01525*.

John R Mashey. 2004. War of the benchmark means: time for a truce. *ACM SIGARCH Computer Architecture News*, 32(4):1–14.

John R Mashey. 2005. Summarizing performance is no mean feat [computer performance analysis]. In *IEEE International. 2005 Proceedings of the IEEE Workload Characterization Symposium, 2005.*, pages 1–1. IEEE Computer Society.