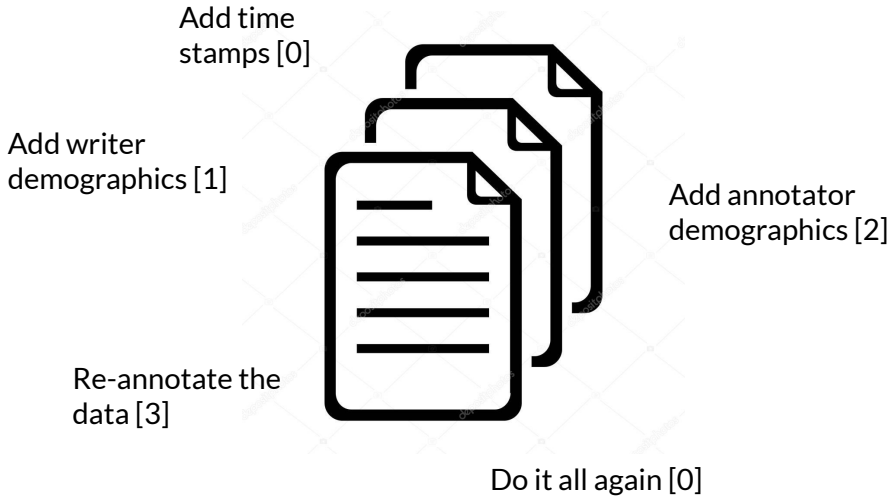
5 Ways to Make Your Data More Relevant

Anders Søgaard

coAStaL

Spoiler



[0] EACL 2021, [1] ACL 2015 / ACL 2021, [2] *to appear*, [3] ACL 2014 / EACL 2014 / NAACL 2015

—

Bowman & Dahl

Me climbing up to stand on the shoulders of giants...

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

Samuel R. Bowman
New York University
bowman@nyu.edu

George E. Dahl
Google Research, Brain Team
gdahl@google.com

Abstract

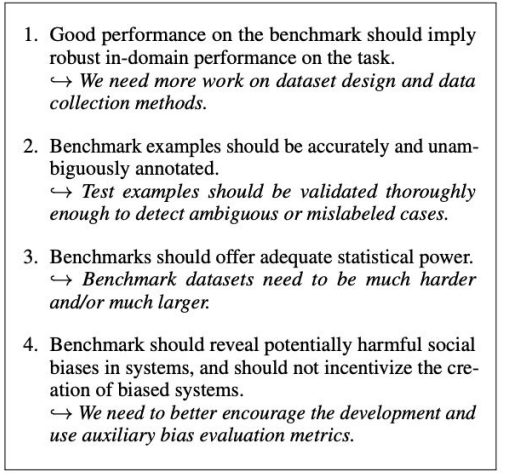
Evaluation for many natural language understanding (NLU) tasks is broken: Unreliable and biased systems score so highly on standard benchmarks that there is little room for researchers who develop better systems to demonstrate their improvements. The recent trend to abandon IID benchmarks in favor of 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. Good performance on the benchmark should imply robust in-domain performance on the task.
   ↪ *We need more work on dataset design and data collection methods.*
2. Benchmark examples should be accurately and unambiguously annotated.
   ↪ *Test examples should be validated thoroughly enough to detect ambiguous or mislabeled cases.*
3. Benchmarks should offer adequate statistical power.
   ↪ *Benchmark datasets need to be much harder and/or much larger.*
4. Benchmark should reveal potentially harmful social biases in systems, and should not incentivize the creation of biased systems.
   ↪ *We need to better encourage the development and use auxiliary bias evaluation metrics.*

Figure 1: A summary of the criteria we propose.

Bowman & Dahl

1. I agree, but domains drift (see Søgaard et al., 2021).
2. Hm, *wait a minute*. Not sure I agree.
3. See Card et al. (2020) and Rodriguez et al. (2021).
4. I agree and think of this in terms of demographics. Also, ❤️

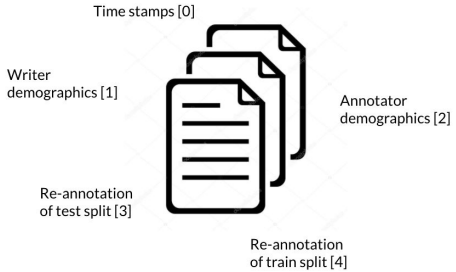


1. Good performance on the benchmark should imply robust in-domain performance on the task.
   ↪ *We need more work on dataset design and data collection methods.*
2. Benchmark examples should be accurately and unambiguously annotated.
   ↪ *Test examples should be validated thoroughly enough to detect ambiguous or mislabeled cases.*
3. Benchmarks should offer adequate statistical power.
   ↪ *Benchmark datasets need to be much harder and/or much larger.*
4. Benchmark should reveal potentially harmful social biases in systems, and should not incentivize the creation of biased systems.
   ↪ *We need to better encourage the development and use auxiliary bias evaluation metrics.*

Figure 1: A summary of the criteria we propose.

Bowman & Dahl

1. I agree, but domains drift (see Søgaard et al., 2021).
2. Hm, *wait a minute*. Not sure I agree.
3. See Card et al. (2020) and Rodriguez et al. (2021).
4. I agree and think of this in terms of demographics. Also, ❤️



1. Good performance on the benchmark should imply robust in-domain performance on the task.
   ↪ *We need more work on dataset design and data collection methods.*
2. Benchmark examples should be accurately and unambiguously annotated.
   ↪ *Test examples should be validated thoroughly enough to detect ambiguous or mislabeled cases.*
3. Benchmarks should offer adequate statistical power.
   ↪ *Benchmark datasets need to be much harder and/or much larger.*
4. Benchmark should reveal potentially harmful social biases in systems, and should not incentivize the creation of biased systems.
   ↪ *We need to better encourage the development and use auxiliary bias evaluation metrics.*

Figure 1: A summary of the criteria we propose.

[0] EACL 2021, [1] ACL 2015 / ACL 2021, [2] *to appear*, [3] ACL 2014, [4] EACL 2014 / NAACL 2015

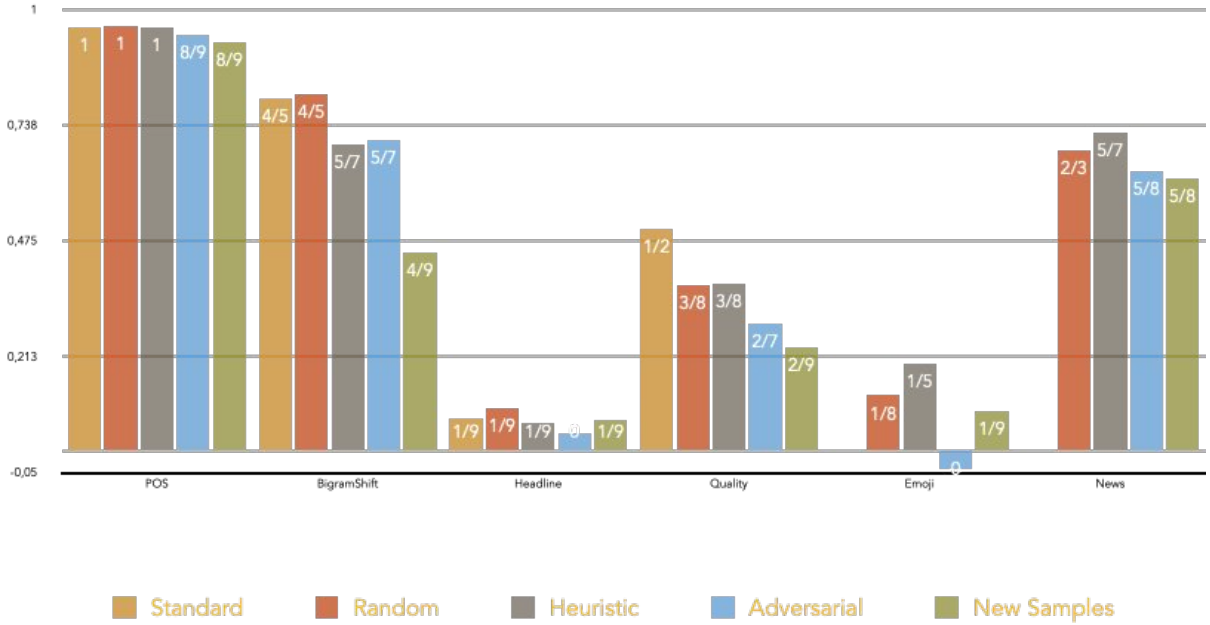
Time stamps

Joint work with Sebastian Ebert, Jasmijn Bastings, and Katja Filippova

1. Levenberg et al. (2010) quantified performance drops over time in SMT
2. Lukes and Søgaard (2018); same for sentiment analysis, predicting lexical polarity shifts

Temporal drift

Søgaard et al. (2021) was mostly a reply to Gorman and Bendrick (2020), arguing against random splits; but the paper also shows that temporal drift is often, even at small scale, more significant than the effects of adversarial splitting/sampling.



Demographics

Joint work with Sheng Zhang and Victor Petrén Hansen

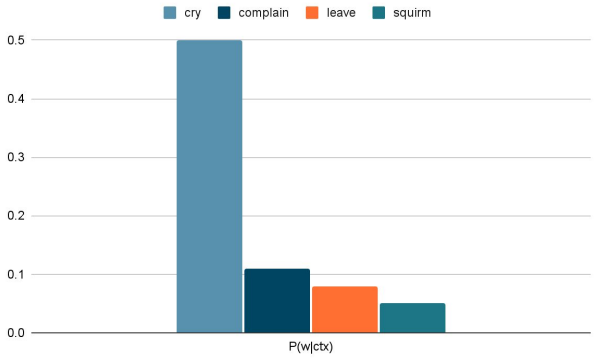
1. Hovy and Søgaard (2015) show POS taggers are sensitive to author age and gender
2. Jørgensen et al. (2016) show POS taggers are sensitive to race

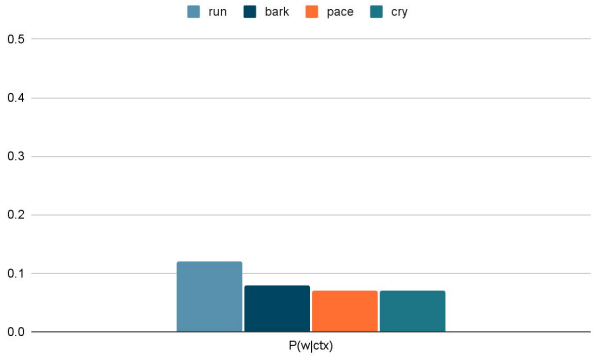
Demographic biases

- Language modeling (newswire)
- Sentiment analysis (product reviews)
- Argument mining (social media)
- Abstractive and extractive summarization (Wikipedia)
- Genre classification (song lyrics)
- Document classification (legal documents and blogs)
- Handwriting recognition (digit/character)

Cloze task

After waiting three hours, Cal whined and started to _____

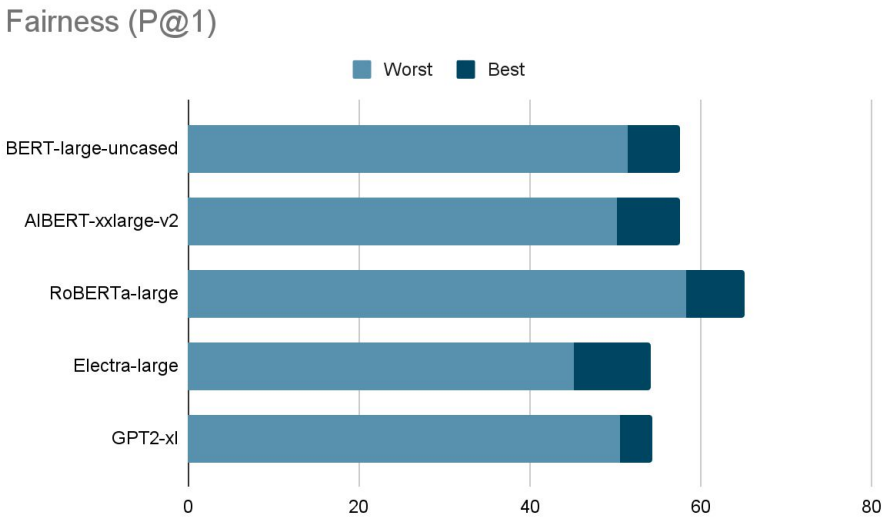




Cloze task

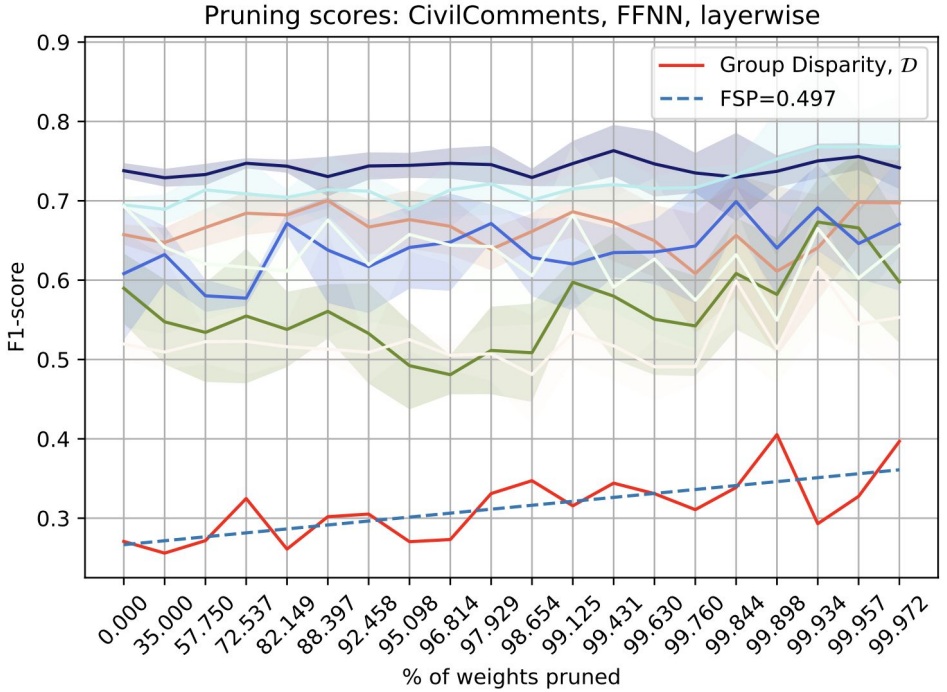
After waiting three hours, Cal whined and started to _____

Groups defined by age, gender, race, and level of education.



Sentiment analysis

Groups defined by age, gender and location.



Multiplicity

Joint work with Terne Thorn Jakobsen and Maria Barrett

1. Plank et al. (2014) show how to use cost-sensitive learning to leverage disagreements
2. Alonso et al. (2015) generalize the result to dependency parsing

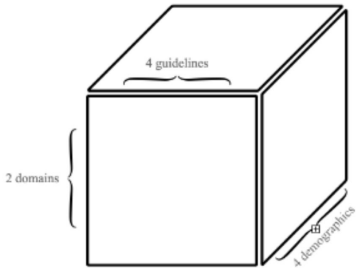
Argument mining

Method: 16 annotations of the same data (group x guideline).

Table: **Ratio** of arguments to non-arguments.

Finding: Men generally more conservative, especially when guidelines require ‘evidence’.

	LIBERAL		CONSERVATIVE		μ
	♀	♂	♀	♂	
G1	0.650	0.517	0.690	0.363	0.555
G2	**0.805**	0.382	0.700	0.342	0.557
G3	0.733	0.487	0.683	0.653	0.639
G4	0.668	0.432	0.383	0.480	0.496
μ	0.714	0.454	0.638	0.460	–



Benchmarks are irrelevant (on their own)

Joint work with Sebastian Ebert, Jasmijn Bastings, and Katja Filippova

Søgaard et al. (2014) argue for evaluation across a dozen benchmarks.

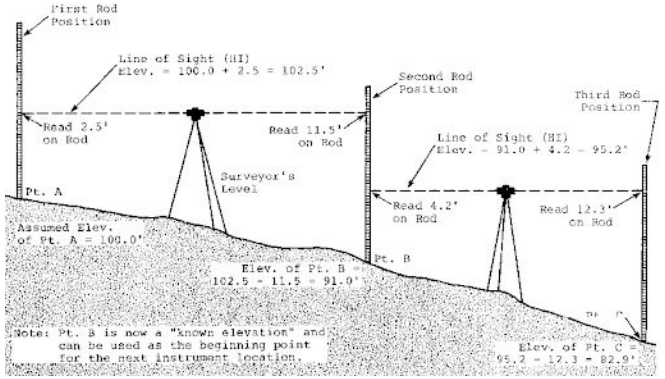
What is a benchmark?

A benchmark is *a mark for accurately repositioning a leveling rod* (‘bench’).

Benchmarks form networks and are used to draw height maps.

Benchmarks

A single benchmark is **not** useful for a surveyor. Height maps rely on multiple benchmarks.



—

Beyond de facto benchmarks

Søgaard et al. (2014) argue we need to compute significance across samples, not across data points (within samples). Ideally, these should be sampled in a shift-pessimistic fashion (Liu et al., 2015).

https://www.aclweb.org › anthology ▾ PDF

Farewell Freebase: Migrating the SimpleQuestions Dataset to ...

by M Azmy · 2018 · Cited by 13 — The SIMPLEQUESTIONS dataset (Bordes et al., 2015) has emerged as the **de facto benchmark** for evaluating these simple questions over ...

https://www.aclweb.org › anthology ▾ PDF

Strong Baselines for Simple Question Answering over ...

by S Mohammed · 2018 · Cited by 55 — has emerged as the **de facto benchmark** for evalu- ating simple QA over knowledge graphs. The original solution of Bordes et al. (2015) fea- tured memory ...

https://www.aclweb.org › anthology ▾ PDF

A Supervised Term Weighting Scheme for Sentiment Analysis ...

by Y Kim · 2014 · Cited by 21 — has become the **de facto benchmark** for sen- timent analysis (Pang and Lee, 2004). • IMDB: 50k full-length movie reviews (25k training, 25k ...

https://www.aclweb.org › anthology ▾ PDF

Replication issues in syntax-based aspect extraction for ...

by E Marrese-Taylor · 2017 · Cited by 9 — 2004b) which became the **de facto benchmark** for evaluation in syntax-based aspect-based opinion mining. This is also a very important part of our environment.

https://www.aclweb.org › anthology PDF

Proceedings of the Student Research Workshop at the 15th ...

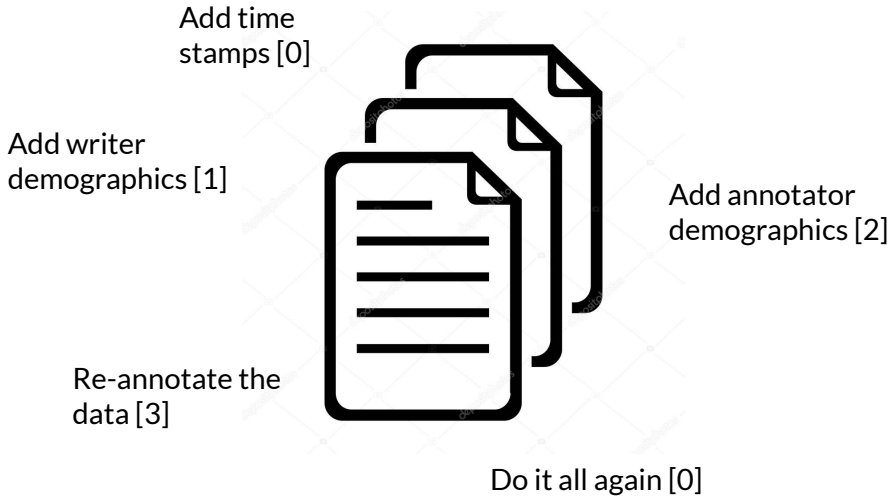
2004b) which became the **de facto benchmark** for evaluation in syntax-based aspect-based opinion mining. This is also a very important part of our environment.

https://www.aclweb.org › anthology ▾ PDF

Proceedings of the 5th Workshop on ... - ACL 2014

27 Jun 2014 — has become the **de facto benchmark** for sen- timent analysis (Pang and Lee, 2004). • IMDB: 50k full-length movie reviews (25k training, 25k ...

Summary



[0] EACL 2021, [1] ACL 2015 / ACL 2021, [2] *to appear*, [3] ACL 2014 / EACL 2014 / NAACL 2015

Sidenote: Guidelines

Joint work with Victor Petrén Hansen and Terne Thorn Jakobsen

Johannsen et al. (2014) argue for guideline-free frame semantic annotation to avoid biases.