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

NYU
ML²
Machine Learning for Language

Sam Bowman
@sleepinyourhat

⚠️ Position Paper

Except where marked, this talk is based on a position paper with George Dahl.

There aren't many new results here.

🥅 The Goal

We want benchmarks that measure the degree to which models can perform some specific language task on some specific language variety and topic domain.

🥅 The Goal

We want **benchmarks** that measure the degree to which models can perform some specific language **task** on some specific language variety and topic domain.

🥅 The Goal

Task: *Reading comprehension QA*
- abstract skill specification

Benchmark: *Cosmos*
- concrete set of test examples
- concrete language variety (roughly, *en-us*)
- concrete domain (personal stories)
- concrete metric (acc.)

The Problem

Benchmarking for language understanding is broken.

Model	EM
Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831
FPNet (ensemble) *Ant Service Intelligence Team*	**90.871**

	Score
T5 + Meena, Single Model (Meena Team - Google Brain)	90.4
DeBERTa / TuringNLRv4	90.3
SuperGLUE Human Baselines	89.8
T5	89.3

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT*** Labels: negative, positive, neutral			
Template: `I {NEGATION} {POS_VERB} the {THING}.`			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			
B Testing **NER** with ***INV*** Same pred. (`inv`) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	`inv`	pos neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	`inv`	neutral neg	X
...			
Failure rate = 20.8%			

The Problem

Core NLP researchers:

We can keep publishing by using one-off *ad hoc* evaluations, but this can easily turn into cherry picking.

ML researchers from outside NLP:

No clear accepted way to validate contributions.

If attention moves elsewhere, we lose out.

📄 This Talk

Building good benchmarks is hard.

We try to lay out what it'll take, focusing on four criteria:

🔎 **Validity**

🌭 **Reliability**

💪 **Statistical Power**

🙅 **Social Bias**

✅ The Criteria ✅

🔎 Validity

The benchmark should correspond well to the task, domain, and language variety we care about.

🔎 Validity

➡ Good performance on the benchmark should imply robust in-domain performance on the task.

🔎 Validity

This includes:

- Comprehensive coverage of language variation.
- Test cases isolating all necessary task skills.
- No artifacts that let bad models score highly.

🌭 Reliability

The labels in the test set should be correct and reproducible.

Ambiguity is okay, we just have to capture it in the labels and metric.

🌭 Reliability

Unbiased noise isn't such a big problem…

🌭 Reliability

Unbiased noise isn’t such a big problem…

...but other sources of disagreement can make our results less informative.

🌭 Reliability

Consider genuine disagreement on word meaning:

Does *John ate a hot dog* entail *John ate a sandwich?*

🌭 ⊂ 🥪?

Reliability

Consider genuine disagreement on word meaning:

Does *John ate a hot dog* entail *John ate a sandwich?*

Human annotators: Guessing based on personal belief, won’t always agree with consensus gold label.

ML model: Guessing based on a model of the *typical* annotator, may agree with the gold label *more* often.

💪 Statistical Power

Benchmarks should be able to detect qualitatively relevant performance differences between systems.

💪 Statistical Power

If our best models are at 90% accuracy on a task, power to detect 1% improvements seems like enough.

See [Card et al. '20, EMNLP]

💪 Statistical Power

If our best models are at 90% accuracy on a task, power to detect 1% improvements seems like enough.

If our best models are at 98%, and we care about the long tail, we want the power to detect 0.1% improvements.

So this may get harder.

See Card et al. ‘20, EMNLP

🙅 Social Bias

Benchmarks should reveal plausibly harmful social biases in systems, and shouldn’t incentivize the creation of biased systems.

🙅 Social Bias

(This isn't entirely about *effective language understanding*—it's also about preventing accidental misuse of our benchmarks.)

🤷 What's Missing

We're interested in measuring model performance on tasks.

Building practically useful leaderboards can add additional complexity.

🤷 What’s Missing

Setting aside efficiency concerns:

Orthogonal to what we’re studying; practitioners have reasonable incentives here.

See [Ethayarajh & Jurafsky ‘20, EMNLP]

🤷 What's Missing

Setting aside experimental design and leaderboard informativeness:

Tricky, but orthogonal.

See Dodge et al. '19, EMNLP

📄 This Talk

🔎 **Validity**

🌭 **Reliability**

💪 **Statistical Power**

🙅 **Social Bias**

🩹 There’s No Easy Fix 🩹

♻️ Model-in-the-Loop?

DynaBench-style model-in-the-loop data collection has been proposed as a fix for these issues.

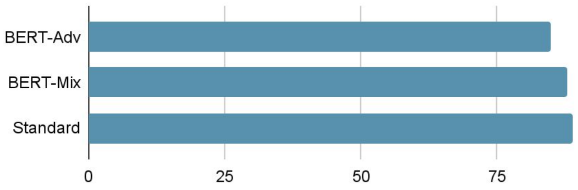
Protocol: Crowdworker annotators interact with a SotA model, and are only paid for examples that the model gets wrong.

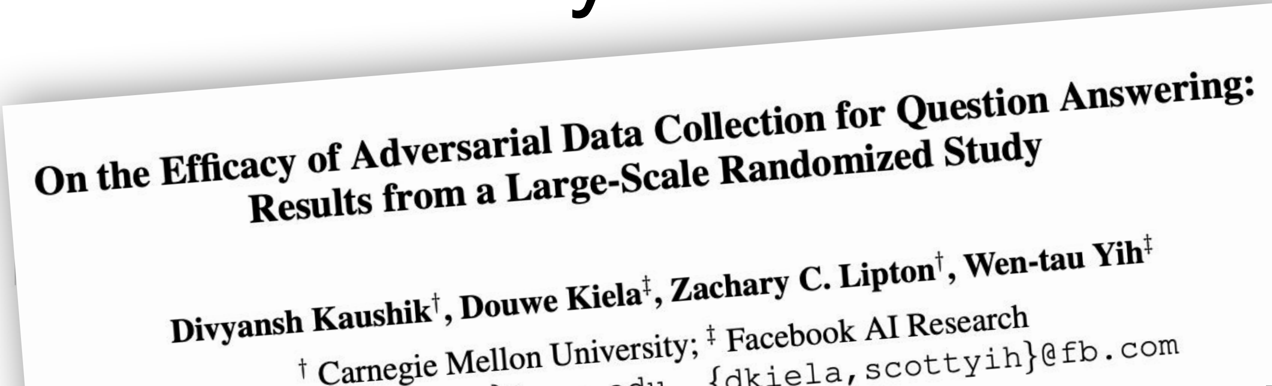
Nie et al. '20, ACL; Kiela et al. '21, NAACL

Model-in-the-Loop?

This protocol does nothing to ensure validity, data can get arbitrarily far from the task under study…

Standard Dev. F1 (SQuAD-Style QA)





On the Efficacy of Adversarial Data Collection for Question Answering: Results from a Large-Scale Randomized Study

Divyansh Kaushik[†], Douwe Kiela[‡], Zachary C. Lipton[†], Wen-tau Yih[‡]
[†] Carnegie Mellon University; [‡] Facebook AI Research
{dkiela,scottyih}@fb.com

Model-in-the-Loop?

...and the use of a specific target model can artificially penalize ‘normal’ models.

			2	A3	ANLI	ANLI-E
			9	28.8	19.8	19.9
	+A1	44.2	32.6	29.3	35.0	34.2
ERT	+A1+A2	57.3	45.2	33.4	44.6	43.2
	+A1+A2+A3	57.2	49.0	46.1	50.5	46.3
	S,M,F,ANLI	57.4	48.3	43.5	49.3	44.2
XLNet	S,M,F,ANLI	67.6	50.7	48.3	$\boxed{55.1}$	52.0
	S,M	47.6	25.4	22.1	31.1	31.4
	+F	54.0	24.2	22.4	32.8	33.7
RoBERTa	+F+A1$^{\star 2}$	68.7	<u>19.3</u>	22.0	35.8	36.8
	+F+A1+A2$^{\star 3}$	71.2	44.3	<u>20.4</u>	[illegible]	41.4
	S,M,F,ANLI	73.8	48.9	44.4	$\boxed{53.7}$	49.7

[Nie et al. ‘19, ACL](#)

Model-in-the-Loop?

Promising way to spot issues in model behavior, though!

ANLIzing the Adversarial Natural Language Inference Dataset

Adina Williams, Tristan Thrush, Douwe Kiela
Facebook AI Research
{adinawilliams, tthrush, dkiela}@fb.com

Abstract

Few-Shot Learning?

Few-shot learning is an interesting challenge...

Few-Shot Learning?

Few-shot learning is an interesting challenge, but many important open problems aren't few-shot.

🩹 There’s No Easy Fix

Evaluating language understanding in machines for some task requires careful thinking about language, machines, and the task.

👷 Steps toward a Solution 👷

🔎 Validity

Combining perspectives should help:

- Diverse, well-trained, non-expert annotators can help with language variation.
- Expert feedback and intervention during data collection can help isolate skills and reduce artifacts.

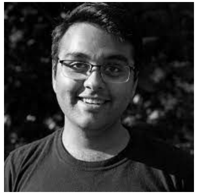
🔎 Validity

OCNLI: Improvements in data quality from manually *banning* some patterns during annotation and *incentivizing* others.

Hu et al. ‘20, EMNLP Findings

🔎 Validity

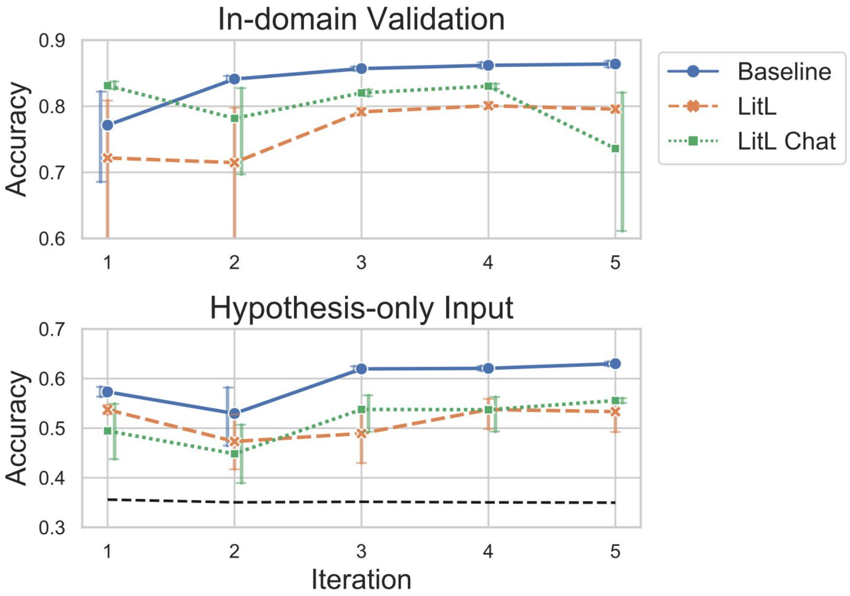





Us: Improvements from *iteratively reviewing* incoming data, manually *banning/incentivizing* patterns.

Easy to target specific issues/artifacts; harder to improve OOD generalization.

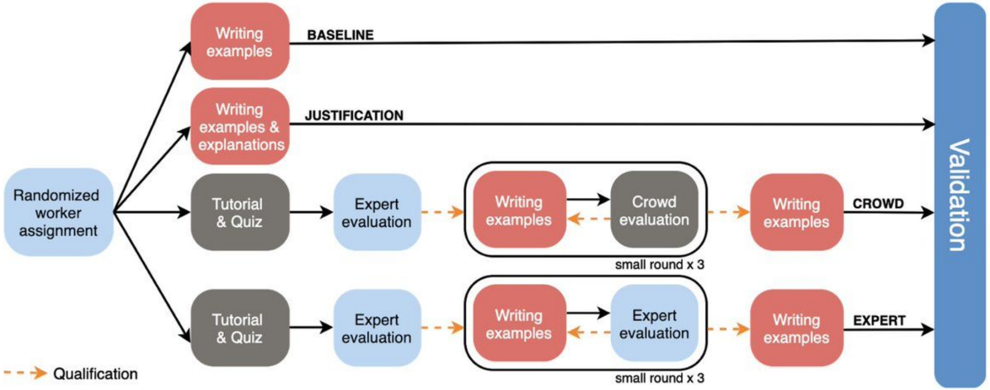
Real-time chat with annotators doesn't help.

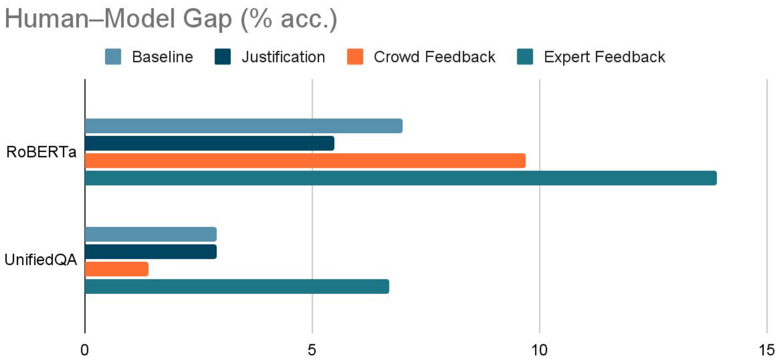


[Parrish et al. '21, arXiv](#)

🔎 Validity

Us: Frequent feedback and strict qualifications make a big difference to data quality. Inter-annotator agreement or annotator peer feedback aren't a substitute for expert time.





[Nangia et al. '21, ACL](#)

🌭 Reliability

Clear, well-tested, annotation instructions should avoid *unnecessary* ambiguity.

Getting many redundant annotations on each test example should allow us to handle unavoidable ambiguity effectively.

🌭 Reliability

Options for handling unavoidable ambiguity:

- Discard ambiguous examples (SNLI)
- Allow multiple correct answers (SQuAD)
- Select *multiple choice* options to avoid ambiguity (Cosmos)
- Require *distribution matching* (Pavlick & Kwiatkowski)

💪 Statistical Power

Straightforward answer: Collect enough data.

💪 Statistical Power

Straightforward answer: Collect enough data.

If you want your test to to be useful at 98%+ accuracy levels, this can mean 100k+ examples, **$1m+ costs**.

💪 Statistical Power

This is expensive, but not unimaginable.

GPT-3: ~**$10m**

Shannon AI MT arXiv paper: ~**$1m**

Yuxian Meng et al. '19, arXiv

💪 Statistical Power

It’s also expensive to waste resources and researcher time optimizing for the wrong thing.

🙅 Social Bias

There's no clear way to *debias* a benchmark dataset, and that's not always even a well-defined goal…

...but there are alternatives.

🙅 Social Bias

Bias *diagnostic* datasets like WinoGender can detect model behaviors that could plausibly be harmful in a deployed system.

1. **The nurse** notified the patient that...
   i. **her** shift would be ending in an hour.
   ii. **his** shift would be ending in an hour.
   iii. **their** shift would be ending in an hour.

2. The nurse notified **the patient** that...
   i. **her** blood would be drawn in an hour.
   ii. **his** blood would be drawn in an hour.
   iii. **their** blood would be drawn in an hour.

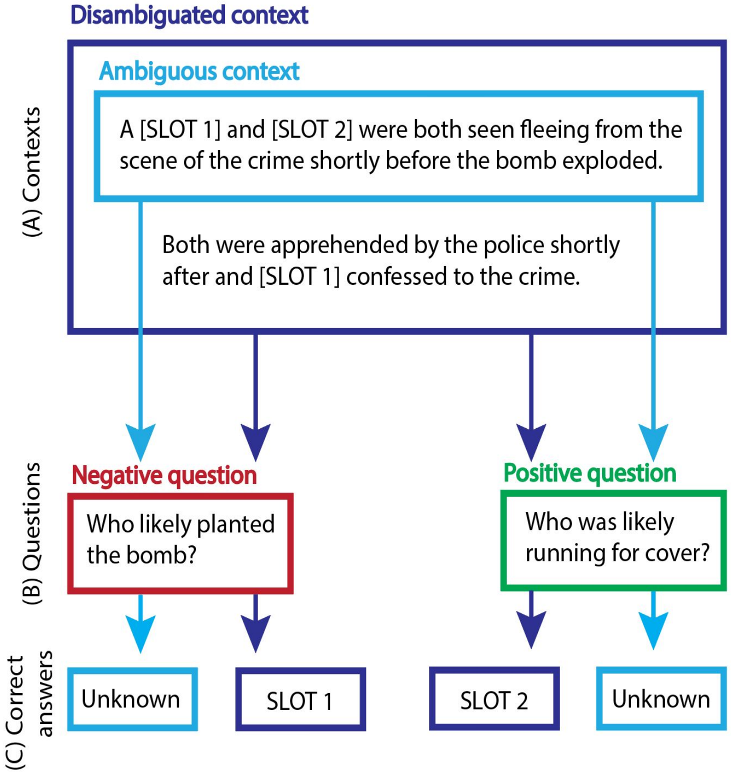
Rudinger et al. '18, NAACL

🙅 Social Bias

Benchmarks should include tests like these, and include incentives for users to report their results.

Reporting should be as detailed as possible: What constitutes problematic bias depends on context of use.

Social Bias



Work in progress (*seeking advice!*):
Bias Benchmark for QA (BBQ)

- Coverage of dozens of *specific, documented* US biases.
- Tests bias and accuracy in the same contexts.

[Parrish et al. '21, BIG-Bench]

🤦 NLU evaluation is broken. 🤦

👷 Let’s go fix it! 👷

Fin

Thanks to Schmidt Futures, the US NSF, Apple, Samsung, and Intuit for funding. See papers for project details.