Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Benchmark in Natural Language Processing (NLP)

Teerapong Aramruang

Introduction

Artificial Intelligence
Index Report 2021

- **Natural Language Processing (NLP) outruns its evaluation metrics:** Rapid progress in NLP has yielded AI systems with significantly improved language capabilities that have started to have a meaningful economic impact on the world. Google and Microsoft have both deployed the BERT language model into their search engines, while other large language models have been developed by companies ranging from Microsoft to OpenAI. Progress in NLP has been so swift that technical advances have started to outpace the benchmarks to test for them. This can be seen in the rapid emergence of systems that obtain human level performance on SuperGLUE, an NLP evaluation suite developed in response to earlier NLP progress overshooting the capabilities being assessed by GLUE.

Does this mean that we have solved natural language processing? Far from it.

https://aiindex.stanford.edu/wp-content/uploads/2021/11/2021-AI-Index-Report_Master.pdf#page=11

Introduction

Dataset	Task	Train	Dev	Test	Evaluation Metrics
BC5-chem	NER	5,203	5,347	5,385	F1 entity-level
BC5-disease	NER	4,182	4,244	4,424	F1 entity-level
NCBI-disease	NER	5,134	787	960	F1 entity-level
BC2GM	NER	15,197	3,061	6,325	F1 entity-level
JNLPBA	NER	46,750	4,551	8,662	F1 entity-level
EBM PICO	PICO	339,167	85,321	16,364	Macro F1 word-level
ChemProt	Relation Extraction	18,035	11,268	15,745	Micro F1
DDI	Relation Extraction	25,296	2,496	5,716	Micro F1
GAD	Relation Extraction	4,261	535	534	Micro F1
BIOSSES	Sentence Similarity	64	16	20	Pearson
HoC	Document Classification	1,295	186	371	Micro F1
PubMedQA	Question Answering	450	50	500	Accuracy
BioASQ	Question Answering	670	75	140	Accuracy

Note: We list the numbers of instances in train, dev, and test (e.g., entity mentions in NER and PICO elements in evidence-based medical information extraction).

- The traditional practices for evaluating performance of NLP models, using a **single metric** such as accuracy or BLEU
- Relying on **static benchmarks** and **abstract task formulations** might have to be **re-consideration.**

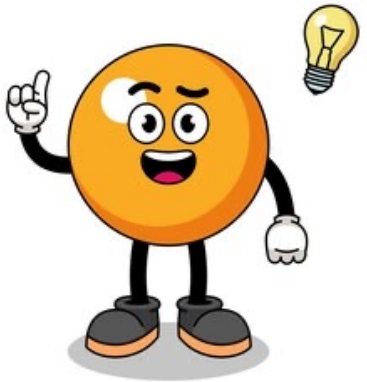
Gu Y, et al. ACM Trans Comput Healthcare. 2021;3(1):Article 2.

Introduction

We thus need to rethink how we design our benchmarks and evaluate our models so that they can still serve as useful indicators of progress going forward.

The GOAL

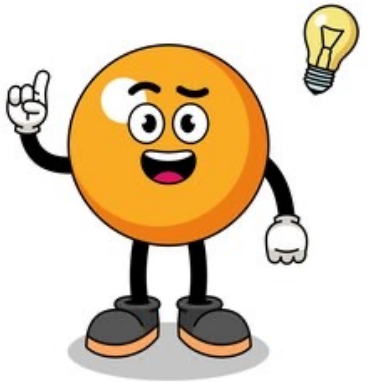
We want benchmarks that measure the degree to which models can perform some specific language task on some specific language variety and topic domain.



Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

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.



Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Benchmark in NLP

"*Datasets are the telescopes of our field*" — Aravind Joshi

For people in NLP field, benchmarks are crucial tools to track progress. — Aravind Joshi said that "*without benchmarks to assess the performance of our models, we are just like astronomers wanting to see the stars but refusing to build telescopes*".

For outsiders in NLP, benchmarks provide an objective lens into a field that enables them to identify useful models and keep track of a field's progress OR as a proxy for overall progress in natural language processing.

Benchmark in NLP

Tasks

A *task* is a **language-related skill** or **competency** that we want a model to demonstrate in the context of a specific input–output format OR **abstract skill** specification.

Task: Multiple-choice reading-comprehension question answering

Benchmarks

A *benchmark* attempts to **evaluate performance on a task** by grounding it to a text domain and instantiating it with a concrete dataset and evaluation metric.

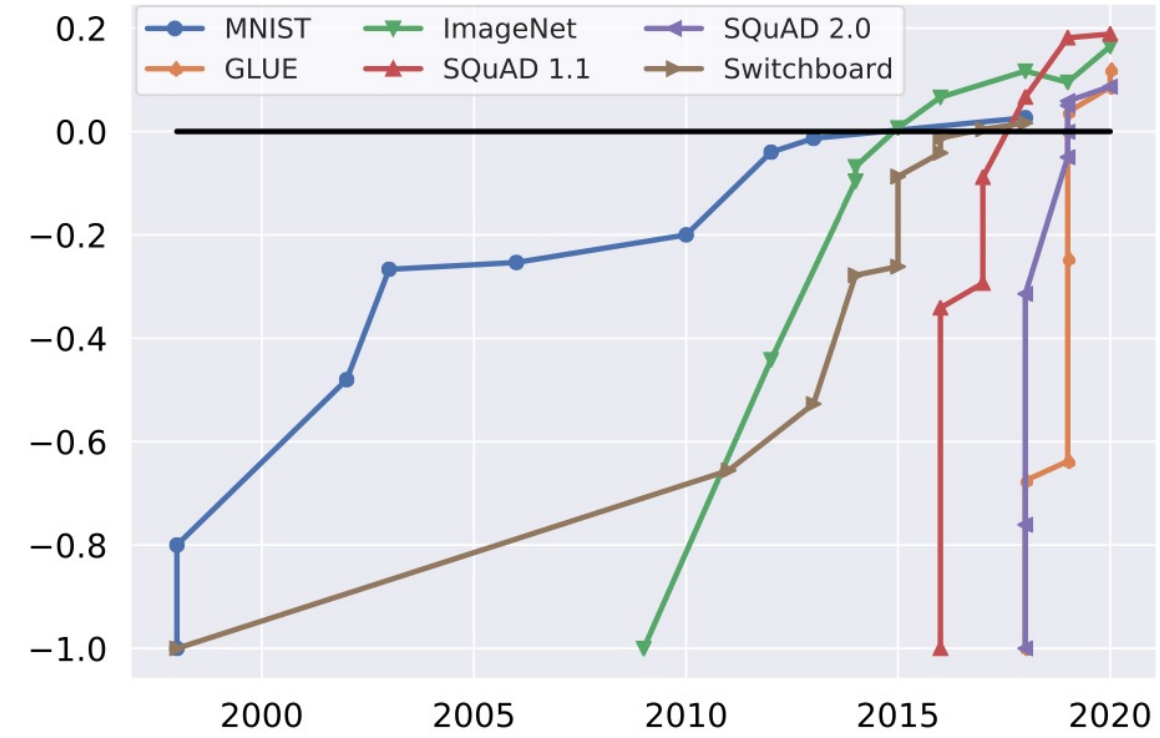
Benchmark: Cosmos benchmark (Huang et al., 2019)

- a specific sample of passages and questions (set of test examples)
- from the English personal narrative domain (language variety and domain)
- test using an accuracy metric (concrete metric)

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Problem in Benchmarking

Benchmark saturation



Reached human-level performance

Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero.

Kiela D, et al. Dynabench: Rethinking Benchmarking in NLP2021.

Problem in Benchmarking

Benchmark saturation

- One factor that has contributed to the saturation of these benchmarks is that **limitations and annotation artefacts** of recent datasets have been identified much more quickly compared to earlier benchmarks.

- In SNLI, annotators have been shown to rely on **heuristics**, which allow models to make the correct prediction in many cases using the hypothesis alone.

Gururangan S, et al. Annotation Artifacts in Natural Language Inference Data. 2018.

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Problem in Benchmarking

Benchmarking is broken

Capability	Min Func Test	INVariance	DIRectional
Vocabulary	Fail. rate=15.0%	16.2%	C 34.6%
NER	0.0%	B 20.8%	N/A
Negation	A 76.4%	N/A	N/A
...			

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%	
C Testing **Vocabulary** with ***DIR***	Sentiment monotonic decreasing (↓)		
@AmericanAir service wasn't great. You are lame.	↓	neg / neutral	x
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg / neutral	x
...			
		Failure rate = 34.6%	

Ample evidence has emerged that the systems that have topped these leaderboards can **fail dramatically on simple test cases.**

Long term issue, **people in NLP filed might be keep publishing** by using one-off ad hoc evaluations, but this can easily turn into cherry picking; on the other hand, **ML researchers from outside NLP, no clear accepted way to validate contributions**.

Ribeiro M, et al. Beyond Accuracy: Behavioral Testing of NLP models with CheckList2020.
Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

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.

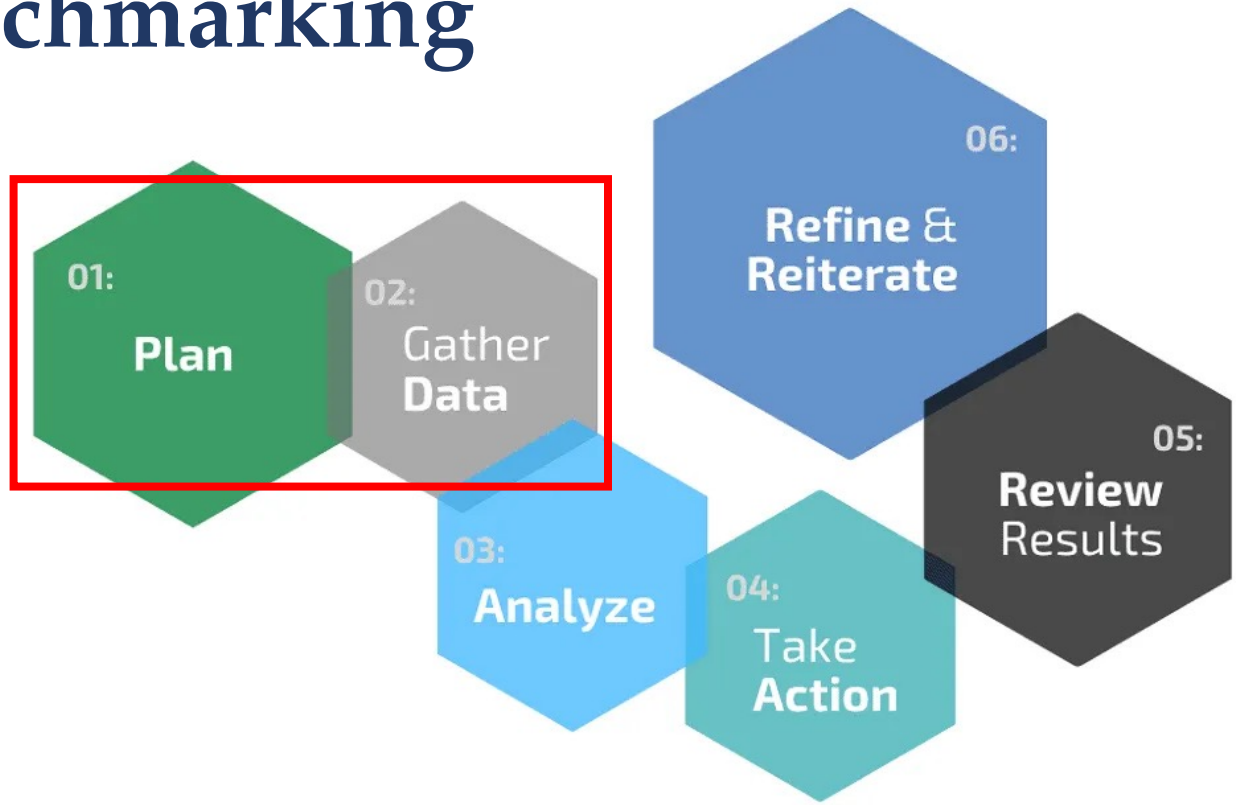
1. Good performance on the benchmark should imply robust in-domain performance on the task.
   ↪ *We need more work on dataset design and data collection methods.*
2. Benchmark examples should be accurately and unambiguously annotated.
   ↪ *Test examples should be validated thoroughly enough to remove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

Figure 1: A summary of the criteria we propose.

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Literature

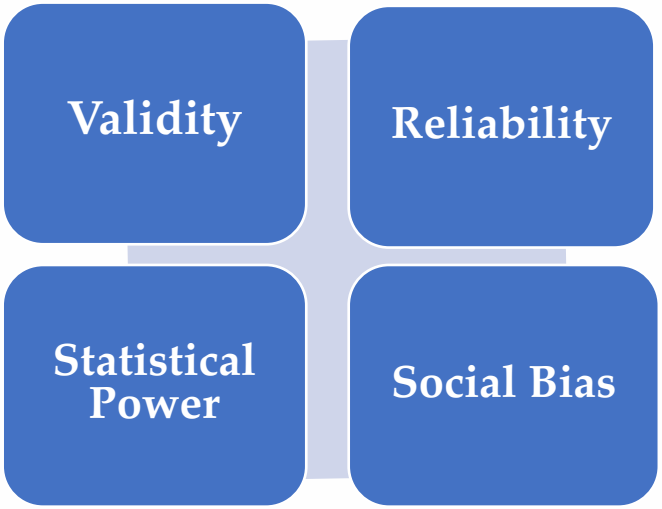
Selecting steps to
Benchmarking



https://chacc.co.uk/small-business-advice/benchmark-your-business/

Literature

- Building good benchmarks is hard.
- They lays out **four criteria** that we would like our benchmarks to satisfy in order to facilitate further progress toward a primarily scientific goal.



- They also attempt to sketch out some **possible directions for each criteria** for improvement along each axis.

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Validity

If a model shows good performance in the benchmark in a particular task and domain, it should result in a good performance in other benchmarks in the same task and domain as well **(transferable across benchmarks).**

- In other words, Good performance on the benchmark should imply robust in-domain performance on the task.
- This criterion is **difficult** to fully formalize.
- Knowing that no simple test that will allow one to determine if a benchmark presents a valid measure of model ability.

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Validity

Minimally, though, it requires the following:

I. Comprehensive coverage of language variation

- Reflect the full range of linguistic variation, including words and higher-level constructions—that is used in the relevant domain, context, and language variety.

II. Test cases isolating all necessary task skills

- Test all of the language-related behaviors that we expect the model to show in the context of the task

III. No artifacts that let bad models score highly

- Acceptable to have artifacts but don't let spurious correlation benefits some model over others.

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

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.

Avoiding three failure cases:

I. Examples that are **carelessly mislabeled**

II. Examples that have **no clear correct label** due to **unclear** or **underspecified** task guidelines

III. Examples that have **no clear correct label** under the relevant metric due to **legitimate disagreements** in interpretation among annotators

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Reliability

III. Examples that have **no clear correct label** under the relevant metric due to **legitimate disagreements** in interpretation among annotators

Unbiased noise isn't such a big problem (random noise)…. but other sources of disagreement can make our results less informative (systematic patterns).

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

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

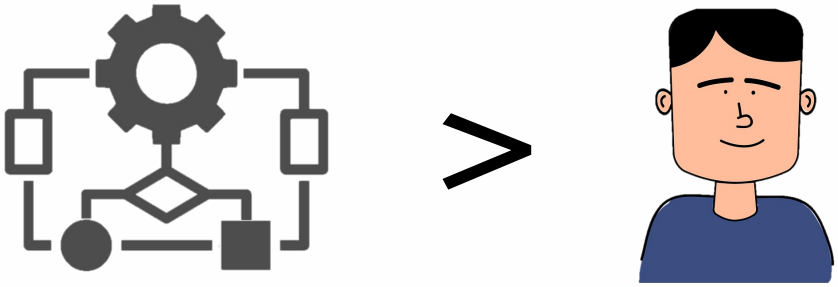
Reliability

Consider genuine disagreement on word meaning:

Does ***John eat a hot dog*** entail ***John eat 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.



Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Statistical Power

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

If our best models are at 90% accuracy on a task, power to detect 1% improvements seems like enough. **Not hard**

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

Long tails phenomena —> later model cannot develop to more powerful than the previous one because the benchmark is not statistical power enough to quantify it.

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Statistical Power

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

This criterion introduces a **trade-off**:

- If we can create benchmark datasets that are both **reliable and highly difficult** for the systems that we want to evaluate, then **moderate dataset sizes** will suffice.
- However, if our benchmark datasets contain many examples that are **easy** for current or near-future systems, then we will need **dramatically larger evaluation sets** to reach adequate power.

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Social Bias

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

- This isn't entirely about *effective NLP*—it's also about preventing accidental misuse of our benchmarks.
- Once model is great in particular socially bias benchmark —> implement (**downstream processing**) may result **unethical** or **illegal.**
- For example, **associations between race or gender and occupation** are generally considered to be undesirable and potentially harmful in most contexts.

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Social Bias

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

- For example, associations between race or gender and occupation are generally considered to be undesirable and potentially harmful in most contexts.
  - If a set of word representations encodes typically **Black female** names like *Keisha* as being **less similar to professional occupation terms** like *lawyer* or *doctor* than typically **White male names** like *Scott* are.
  - Then a model using those representations is likely to **reinforce harmful race** or **gender biases** in any downstream content moderation systems or predictive text systems it gets used in.

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Steps toward a Solution

- Validity
- Reliability
- Statistical Power
- Social Bias

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Validity

Building valid benchmarks will require significant **new research into data collection methods**, at least some of which will be specific to the task under study.

Combining diverse perspective of all different people using language and all different ways represented in your tasks

- 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.

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

OCNLI (Original Chinese Natural Language Inference) benchmark

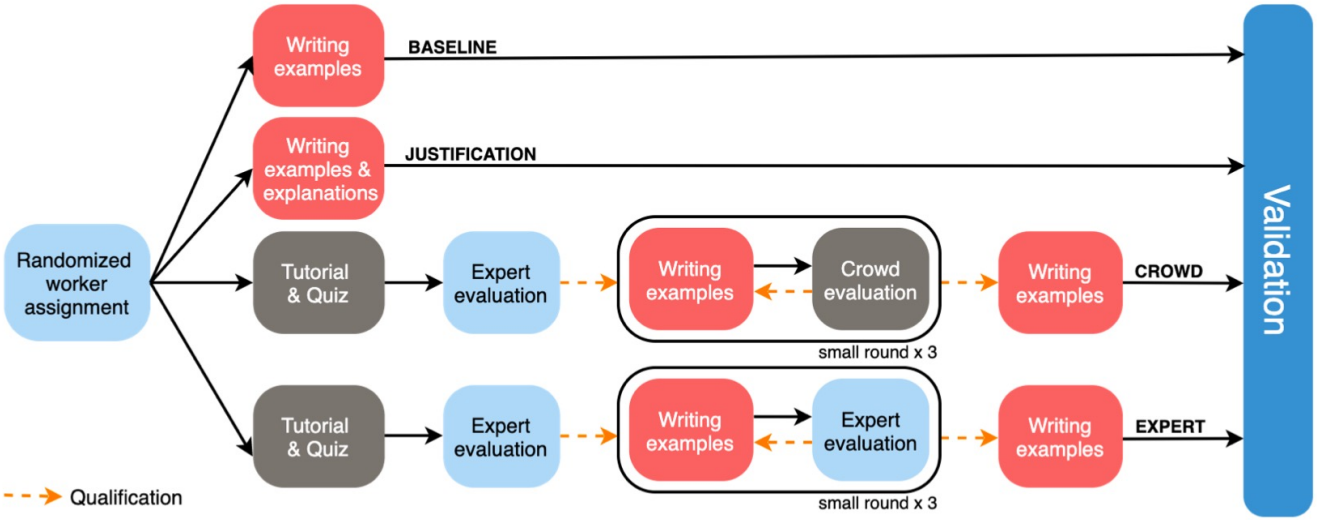
- Improvements in data quality from manually *banning* some patterns during annotation and *incentivizing* others

Hu H, et al. OCNLI: Original Chinese Natural Language Inference. FINDINGS; 2020.

Validity

Another strategy

- 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.



The initial pool of crowd workers are randomly assigned to one of four protocols and the datasets are collected in parallel.

Nangia N, et al. What Ingredients Make for an Effective Crowdsourcing Protocol for Difficult NLU Data Collection Tasks?2021. 1221-35 p.

Reliability

Careful planning and pilot work before data collection can largely resolve the issue of ambiguous annotation guidelines.

- In validation phase, we can systematically identify and discard ambiguously labeled examples.

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.

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Reliability

Options for handling unavoidable ambiguity:

- Discard ambiguous examples (SNLI benchmark)
- Allow multiple correct answers (SQuAD benchmark)
- Select multiple choice options to avoid ambiguity (Cosmos benchmark)
- Require distribution matching (Pavlick E, et al. 2019;7:677-94.)

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.
Pavlick E, et al. Inherent Disagreements in Human Textual Inferences. Transactions of the Association for Computational Linguistics. 2019;7:677-94.

Statistical Power

Straightforward answer: simply **estimate the number of examples** required to reach the desired statistical power for any plausible short-to-medium term system evaluation for the task, and **collect that number** of examples.

For settings in which large datasets are necessary, we see **no clear way to avoid high costs.**

If you want your test to to be useful at **98%+ accuracy levels**, this can mean 100k+ examples, **causing $1m+ costs** (the long tail of benchmark performance).

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Statistical Power

Ultimately, we need to compare the **cost** of making serious investments in **better benchmarks** to the **cost of wasting researcher time and computational resources** due to our inability to measure progress.

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

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**.

A viable alternate approach could involve the expanded use of **auxiliary metrics**:

- Rather than trying to **fully mitigate bias** within a single general dataset and metric for some task
- Benchmark creators can introduce a family of additional expert-constructed **test datasets and metrics** that each isolate and **measure a specific type of bias.**

For example, **WinoGender test,** other working groups (Parrish et al. '21, BIG-Bench)

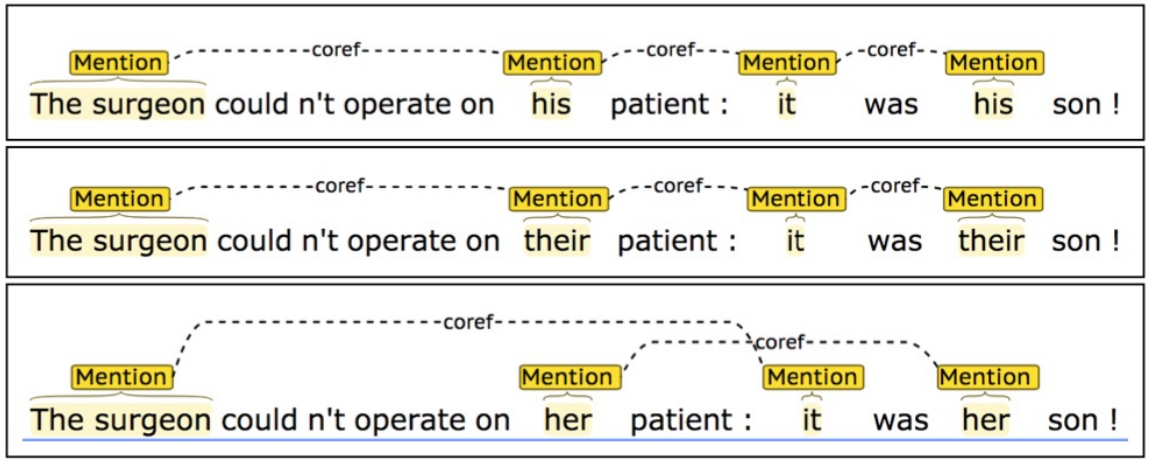
Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p
Zhao J, et al. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods2018. 15-20 p.
https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/bbq_lite

Social Bias

Bias diagnostic datasets like **WinoGender** can detect model behaviors that could plausibly be harmful in a deployed system.
(e.g., **gender-occupation stereotype** social bias)

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.



Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

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.

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

Conclusion

Validity

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.*

Reliability

2. Benchmark examples should be accurately and unambiguously annotated.
   ↪ *Test examples should be validated thoroughly enough to remove erroneous examples and to properly handle ambiguous ones.*

Statistical Power

3. Benchmarks should offer adequate statistical power.
   ↪ *Benchmark datasets need to be much harder and/or much larger.*

Social Bias

4. Benchmarks should reveal plausibly harmful social biases in systems, and should not incentivize the creation of biased systems.
   ↪ *We need to better encourage the development and use of auxiliary bias evaluation metrics.*

Bowman S, et al. What Will it Take to Fix Benchmarking in Natural Language Understanding?2021. 4843-55 p.

…Let's go fix it!