Evaluating Evaluations: A Perspective on Benchmarks

https://github.com/kwchurch/WSDM_2024_tutorial

Validity and Reliability

Krippendorff (2018) *Content analysis: An introduction to its methodology*

- Reliability is about data, and validity is about truth:
  - Reliability:
    - The attribute of Data on which researchers can rely in answering their Research questions – Krippendorff, p. 411
  - Validity:
    - The quality of a claim to be as stated, true, or correct. – Krippendorff, p. 413
- Validity assumes a hypothesis/claim.
  - Hard to test validity without a clearly stated hypothesis
- There are more papers on reliability than validity in our field(s).
  - 18k matches for “inter-annotator agreement” in ACL Anthology
  - https://aclanthology.org/search/?q=inter-annotator+agreement

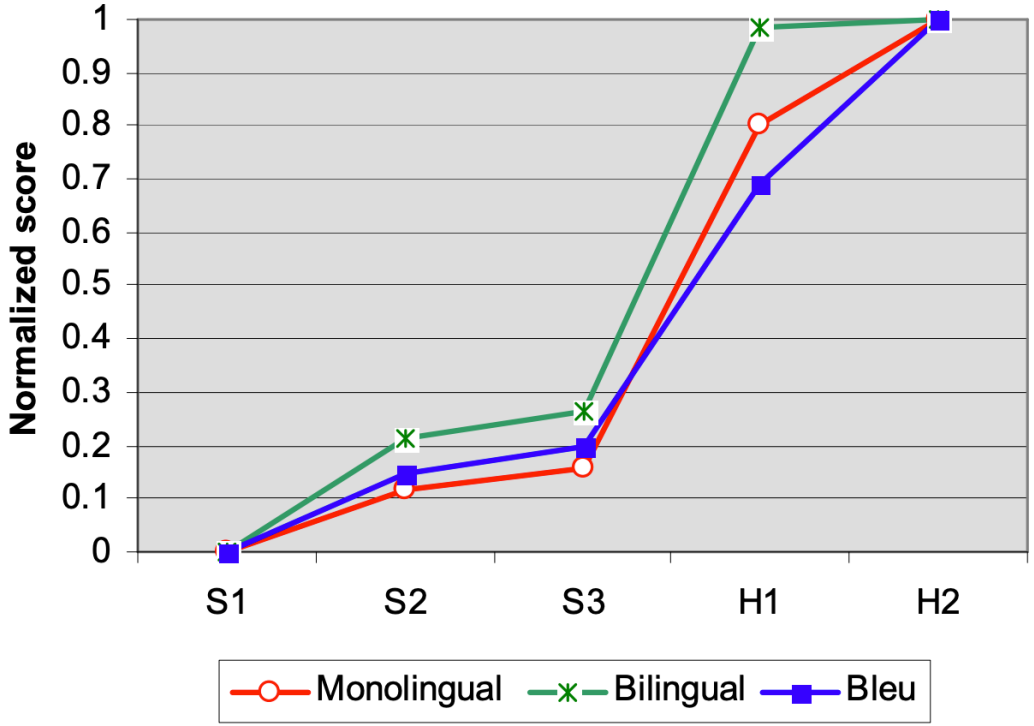
Reliability of BLEU https://aclanthology.org/P02-1040/

- The original paper asks 3 excellent questions about reliability
  - How reliable is the difference in BLEU metric?
  - What is the variance of BLEU score?
  - If we were to pick another random set of 500 sentences,
    - would we still judge S3 to be better than S2?

Reliability of BLEU

- In their Figure 7,
  - they use BLEU to compare three machine translation systems (S1, S2 and S3) and
  - two human (non-professional) translators (H1 and H2).
- They report that humans score better than machines.
- They conclude:
  - "BLEU's strength is that it correlates highly with human judgments
    - by averaging out individual sentence judgment errors over a test corpus
    - rather than attempting to divine the exact human judgment for every sentence:
    - *quantity leads to quality*."

Figure 7: BLEU vs Bilingual and Monolingual Judgments



https://aclanthology.org/P02-1040/

Validity of BLEU

- Less discussion of validity in original paper
- Some papers raise serious questions about validity
  - of BLEU and other metrics
  - at least, for some use cases that go beyond the original BLEU paper

- It is hard to talk about validity without a clearly stated use case
  - https://aclanthology.org/D17-1238/
    - *This paper shows that state-of-the-art automatic evaluation metrics for NLG systems*
      - *do not sufficiently reflect human ratings,*
      - *which stresses the need for human evaluations.*
  - https://aclanthology.org/J18-3002
    - *Overall, the evidence supports using BLEU for diagnostic evaluation of MT systems*
      - *(which is what it was originally proposed for),*
    - *but does not support using BLEU outside of MT, for evaluation of individual texts, or for scientific hypothesis testing.*

Use Cases

- BLEU was originally proposed
  - to compare a small number of systems for DARPA competitions.
- But soon after BLEU was introduced,
  - Och suggested using BLEU for a very different use case.
- It had been standard practice to to use
  - different metrics for testing and training.
- Och found that if one is going to use BLEU to evaluate systems,
  - then his system would do better in the evaluation if he also used BLEU for training.
- Och's suggestion did well in competitions,
  - but raises questions about reliability and validity

Consequence of Evaluation: Proposed Scale

- Minor: e.g., SOTA-chasing, leaderboards
- Moderate: e.g., Multitask learning
  - Generalizing results over workloads and tasks.
- Major: e.g., Och training
  - Significant consequences for system performance
- Mission Critical: e.g., SPEC
  - What should I buy? And what performance should I expect on my workloads?
  - Go/No-go decisions

Averaging: Arithmetic vs. Geometric

- There are many benchmarks in our field:
  - e.g., GLUE, SciRepEval, MS MARCO and Big Bench
- Many designed for SOTA-chasing,
  - but hopefully, results will generalize to more important use cases.
- More likely to generalize to more important use cases
  - if they were designed to do so in the first place

- SPEC was designed to report performance relative to baseline
  - How much better is the candidate CPU relative to VAX 11/780?
  - On the user's (unspecified) workload?
- Mashey argues that
  - geometric means generalize better
  - over workloads than arithmetic means.
- Mashey suggests that results on our benchmarks would
  - generalize beyond less important SOTA-chasing use cases
  - if we replaced arithmetic means with geometric means in GLUE
  - (and many of our other benchmarks).

SPEC: A benchmark for evaluating CPUs

Code	App Area	Lines	Remarks
gcc	Compiler	87,800	CNU C Compiler V1.35, compiles 76 soruces, 10% I/O
Espresso	Logic Design	14,800	PALs generation tool, heuristic minimization, little paging
Li	Interpreter	7700	Lisp interpreter (XLIST 16), solves 8-queens problem using recursive backtracking, many jumps/loops
Eqntott	Logic design	3500	Creates truth tables; >95% of time in qsort
Compress	Data compression	1500	Compress/decompress 1MB file 20 times using adaptive Lempel-Ziv coding
Sc	Spreadsheet	8500	Spreadsheet app based on the Unix "curses"

Table 1: SPEC CINT92 suite (from Table 2 in [8])

Code	App Area	Lines	Remarks
Spice2g6	Circuit Design	18,900	Analog circuit simulation tool, unvectorizable, unparallelizable, uses cache
Doduc	Physics, simulation	5300	Monte Carlo simulation of thermohydraulic neclear reactor, unvectorizable, many jumps/loops
Fpppp	Quantum chemistry	2700	Electron integral, unvectorizable, no jumps (good for pipelining)
Tomcatv	Geometry	200	Mesh generation, 90-98% vectorizable, exercises data cache
Nasa7	Kernels	1300	Some kernels are vectorizable
Mdljdp2	Chemistry	4500	Motion equations for 500-atom model
Wave5	Physics	7600	Particle and Maxwell's equations
Ora	Optics	500	Ray Tracing
Alvinn	Robotics	300	Neural network propagation training
Ear	Medicine	5200	Ear simulation using FFT
Mdljsp2	Chemistry	3900	Single precision of Mdljdp2
Swm256	Simulation	500	Shallow water equations system
Su2cor	Quantum physics	2500	Masses of elementary particles, 98.5% vectorizable
Hydro2d	Astrophysics	4500	Calactic jets, 99.5% vectorizable

Table 2: SPEC CFP92 suite (from Table 3 in [8])

Challenges for Validity and Reliability

- Unrepresentative Samples
- Test/Train Splits: Interpolation vs. Extrapolation
- Leakage
- Labeling and Inter-annotator Agreement
- Too many (irrelevant) tasks

Unrepresentative Samples

- Ideally, a benchmark should be
  - a representative sample
  - of a larger population of interest.
- Balanced Corpora:
  - 1960s: Brown Corpus
  - 1990s: British National Corpus
- Current view: catch as catch can
  - *there is no data like more data* – Mercer
- Representative samples → More credible generalizations

	Semantic Scholar		Citation Task	
	%	N	%	N
$\|A\|$	47%	99M	82%	21,885
$\|L\|$	53%	111M	97%	25,850
$\|A \cup L\|$	70%	145M	99%	26,378
$\|A \cap L\|$	31%	65M	80%	21,357
totals	100%	208M	100%	26,657

Table 3: Comparison of Semantic Scholar with a benchmark (SciRepEval Cite task). There are large mismatches in $|A|$ (papers with abstracts) and $|L|$ (papers with links in G).

Test/Train Splits: Interpolation ≠ Extrapolation

It's Difficult to Make Predictions, Especially About the Future

- It is common for graphs to evolve over time.
  - For example, the academic literature is growing very quickly,
  - doubling every nine years
- Benchmarks such as OGB focus on static snapshots from a few years ago,
  - missing opportunities to encourage the community to study growth and timeliness.
- Random splits are common in graph learning benchmarks, e.g., WN18RR,
  - a popular Knowledge Graph Completion (KGC) task based on WordNet

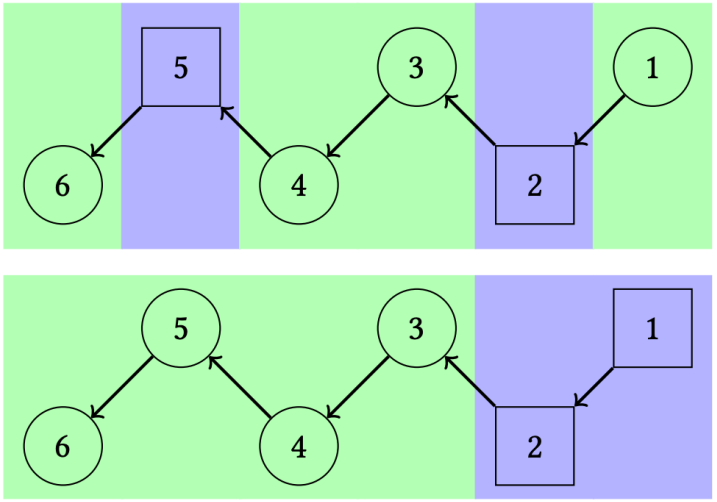


Figure 1: Random Splits (top) vs Causal Split (bottom).

Paper	Year	Title
1	2018	[...] Photograment imaging
2	2016	Convenient probe of S(1D2)[...]
3	2005	Megapixel ion imaging [...]
4	2003	Direct current slide imaging [...]
5	1995	profiles of CI(2Pj) photoframents [...]
6	1988	Adiabatic dissociation of [...]

Table 4: 1 cites 2, 2 cites 3,..., 5 cites 6

Leakage

- Leakage is common in many benchmarks
  - There is considerable discussion of leakage in SciDocs
    - (see 4.2 of https://aclanthology.org/2022.emnlp-main.802/)
  - WN18 → WN18RR (WordNet benchmarks)
    - If x is-a y (a car is a vehicle),
      - then there will be two links between *x* and *y*: hypernym and hyponym
    - Since WN18 randomly splits links into test and train,
      - one of these links is likely to be in test and the other in train
    - Unfortunately, WN18RR corrects some (but not all) of the leakage
      - See table 4 of https://aclanthology.org/2021.emnlp-main.501/
    - Despite this leakage, there are many papers on WN18RR
      - https://paperswithcode.com/dataset/wn18rr

Labeling and Inter-annotator Agreement

- The documentation on SciRepEval makes it clear that some labels are “silver” (less reliable) [underlining added]:
  - *... a new large-scale field of study (FoS) multi-label training set of more than 500K papers with **silver** FoS labels based on publication venue*
- We compared FoS labels in SciRepEval with FoS labels in MAG and found large differences.
  - More agreement in some fields (Computer Science)
    - Less agreement in History, Sociology and Art.
  - It is possible that the annotators
    - are more familiar with Computer Science
    - than History, Sociology and Art.

Too many (irrelevant) tasks

- Example: SciRepEval
  - There are so tasks that
    - some will be more relevant to our use cases,
    - and others will be less relevant
  - The FoS task, for example, classifies documents into 23 fields of study.
    - The FoS task is probably not relevant to recommendation use cases

Principles

- Characterize use case & audience
- Validity:
  - Relevance of task to use case
- Reliability:
  - Inter-rater agreement
- Realistic workloads
- Labeling and annotation
  - Documentation:
    - How was it done?
    - Availability of instructions

- Maintenance
  - Include a feedback loop mechanism to maximize adoption
  - Workload Evolution
  - Lessons learned and addendum(s)
  - Lifecyle and deprecation
- High standards
  - (for high-stakes use cases)
    - Use of established software engineering and data management techniques
      - (e.g., code review, versioning, configurations, dependencies, and testing).
    - How was the data sourced?
      - Provenance?
      - Can we data set be generated again easily?
    - Clean and well-documented data model