A Metrological Framework for Evaluating Crowd-powered Instruments

Chris Welty, Praveen Paritosh, Lora Aroyo
Google

Introduction

One common use of crowdsourced data is to evaluate AI systems that interpret unstructured input (e.g. image, audio, video, text, etc.). In AI, this is by and large the only method of evaluating AI systems – by comparison to humans performing "the same" task. However, most published papers fail to report if the performance differences they find are significant, are within the capability of the evaluation set to measure. How often are published improvements just an accident of the evaluation data set?

In this paper we introduce the idea of using metrology, the science of measurement and its application, as a basis for understanding human (crowd) powered evaluations. We begin with the intuitive observation that evaluating the performance of an AI system is a form of measurement, and in general when we measure something we use an instrument. In the hard sciences, when instruments are used their metrological properties are reported.

Our goal is to provide a valid and rigorous method for characterizing crowd-powered instruments used for evaluation, so that the errors in the instrument can be included in the significance tests, and we can draw a more valid conclusion about the result. Through establishing this method, we expect to learn how to build better instruments and justify allocating resources to improving them.

Benchmarks for ML Performance Assessment

According to (Baker 2016) we experience a "reproducibility crisis" across various scientific fields, e.g. deep learning(Crane 2018), information retrieval (Arguello et al. 2016; Li et al. 2006). In 2017 ACL also acknowledged the need for replicable and reproducible results (Joakim Nivre, 2017). However, none of these efforts focuses on the role of benchmark datasets and their quality as instruments for measuring system performance, and thus also impacting the reliability of results.

A growing number of researchers have recently pointed out numerous problems and inconsistencies in the way the data in such benchmarks is collected (Son 2018; Inel 2017; 2019), and especially in what can be concluded from these evaluations. (Rogers, Drozd, and Li 2017; Gladkova, Drozd, and Matsuoka 2016; Wendlandt, Kummerfeld, and Mihalcea 2018) raise awareness on how properties of the data in benchmark datasets used for evaluation of word embeddings systems play a role in the stability of the results. In this paper we argue that paying attention to the instrument characteristics of a benchmark affects how we make interpretations about the systems being measured, a key question of interest to AI/ML researchers using these benchmarks.

Metrology of Crowd-Powered Instruments

Metrology can help to understand the reliability of measurements well enough for comparison, which is what AI science needs from human computation. We borrow the framework from metrology [BIPM/GUM][1] and modify it to suit crowd-powered instruments. We illustrate this by characterizing the Wordsim-353 (WS-353)[2] benchmark for measuring AI systems' ability to compute lexical similarity of words:

Measurement Procedure

The crowd was tasked to rate 353 word pairs on their similarity, with two special cases to test and calibrate the workers: *a repeated pair (money, cash)*, and a *repeated word (tiger, tiger)*. 13 Workers rated all 353 items, 16 workers rated only 200. WS-353 is typically used as an evaluation instrument by comparing a system's predicted similarity scores to the mean worker scores on each pair using Spearman's rho (rank correlation).

Precision

Precision is the variance of the instrument when measuring the same, unchanging, object. High variance indicates lower precision. The precision of the WS-353 instrument is a vector of per-item standard deviation (σ) scores. The WS-353 word pair σs are normally distributed, with mean $\sigma = 1.7$, and $\sigma(\sigma) = .54$. The highest variance is in the pair *(precedent - example)*, and the lowest aside from *(tiger - tiger)*, is *(king - cabbage)*, which also has the lowest similarity score. In general, the standard deviations have a crescent shaped

[1] https://www.itl.nist.gov/div898/handbook/glossary.htm

[2] https://aclweb.org/aclwiki/WordSimilarity-353_Test_Collection_(State_of_the_art)

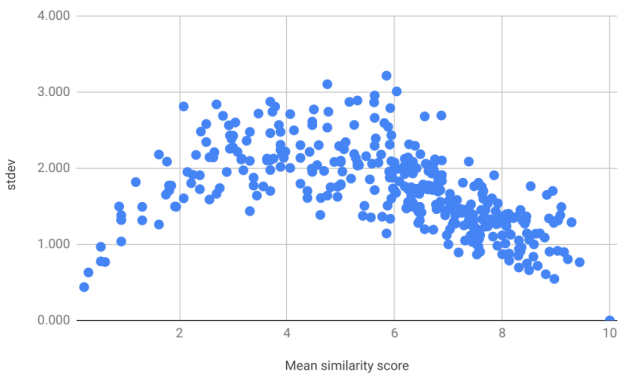

Figure 1: The metrological precision of WordSim-353 can be vizualized by comparing the standard deviation of each pair's votes to the mean.

Sys	WS353	WS353r	WS353r2
ESA	0.739	0.643	0.563
TSA	0.784	0.706	0.635
CLEAR	0.810	0.715	0.642
CLEAR+TSA	0.841	0.748	0.670

Table 1: Spearman's rank correlation of each system's predictions to the mean worker scores of the original and two subsequent reproductions of the WS-353 measurements. While the relative order of the system scores remain the same within each reproduced measurement, the scores change by far more than the differences between systems.

relationship with the mean (Figure 1). This indicates the instrument has low precision.

The analysis shows that 95% of the pairs are statistically equivalent to 10% of their nearest neighbors, making the rank nearly meaningless. *An improvement in rank correlation displayed by one system over another may be statistically accidental.*

Reproducibility and Repeatability

Repeatability is the degree to which measurements under the same conditions agree with each other. Reproducibility is on a different level, for measurements made under different environmental conditions. We utilize Krippendorff's alpha as a summary statistic for repeatability. For the original WS-353 instrument, $\alpha = 0.59$. For reproducability we re-ran the WS-353 set of word pairs with Amazon MTurk using the original guidelines, word pairs, and 13 judgements per pair. Unlike the original, we collected work from 50 different workers to yield the same total number of judgements. These instruments had $\alpha = 0.59$ and $\alpha = 0.52$.

We then compared the results of four word-similarity systems (Halawi et al. 2012) using the original and reproduced instruments. These results are shown in Table 1.

The scores of each system vary dramatically across the three instruments, however the relative ranking of systems remains constant. It is not clear if this is significant, more analysis is required.

	0.0	0.9	1.8	2.7	3.6
CLEAR	0.742	0.782	0.816	0.845	0.867
CLEAR+TSA	0.651	0.675	0.697	0.723	0.744
TSA	0.576	0.599	0.621	0.646	0.668
ESA	0.506	0.530	0.553	0.577	0.598
Count	61425	46375	33359	23280	15693

Table 2: Pearson correlation between (1) the normalized distances between all ws353 pairs and (2) the normalized distances between all the predicted pairs at different values of instrument resolution.

Sensitivity and Resolution

Sensitivity in metrology refers to the rate at which changes in the measured object are reflected by the instrument, it is a ratio of those two changes. Resolution is the smallest change the instrument can detect. It is very difficult to tease these two notions apart when characterizing a crowd-powered instrument. Further study is needed, for the time being we adopt resolution as the characteristic we will specify for the WS-353 instrument.

Table 2 shows the Pearson correlation between: the $l2$-normalized distances between all ws353 pairs and the $l2$-normalized distances between all the experiment pairs. Each column of the table represents the correlations at a particular instrument resolution, such that the correlated distances only include those on pairs for which the ws353 distance is above the resolution. The counts show how many pairs are included at that level of resolution.

We were expecting the results to show that, as the resolution threshold increases, the four systems become decreasingly different. The results do not show this, however they do lend more evidence to support the hypothesis that measuring the differences between these four systems is beyond the *WordSim-353* instrument's capabilities. In particular, the results show that CLEAR, and not CLEAR+TSA, has a higher correlation with the WS-353 pairwise distances.

This experiment has high statistical power because of the number of pairs (bottom row in Table 2), and these differences are significant for that reason.

Discussion

The main contribution of this paper is in introducing elements of metrology to characterize the quality and reliability of crowd-powered instruments. From the results of these experiments we do see promising signals that metrology can help us understand the evaluations, and in particular can help us determine when system performance evaluations show significant improvements. There are still many open questions around how to better characterize crowd-powered instruments, and there is still a large body of future work to do with respect to the characterization of a wide variety of AI benchmarks.

References

Arguello, J.; Crane, M.; Diaz, F.; Lin, J.; and Trotman, A. 2016. Report on the sigir 2015 workshop on reproducibility,

inexplicability, and generalizability of results (rigor). *SIGIR Forum* 49(2):107–116.

Baker, M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533:452–454.

Crane, M. 2018. Questionable answers in question answering research: Reproducibility and variability of published results. *Transactions of the Association for Computational Linguistics* 6:241–252.

Gladkova, A.; Drozd, A.; and Matsuoka, S. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, 8–15. Association for Computational Linguistics.

Halawi, G.; Dror, G.; Gabrilovich, E.; and Koren, Y. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, 1406–1414. ACM.

Inel, O. 2017. Harnessing diversity in crowds and machines for better ner performance. In *The Semantic Web - 14th International Conference, ESWC2017, Part I*, 289–304.

Inel, O. 2019. Validation methodology for expert-annotated datasets: Event annotation case study. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, 12:1–12:15. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Li, Y.; McLean, D.; Bandar, Z. A.; O'Shea, J. D.; and Crockett, K. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering* 18(8):1138–1150.

Rogers, A.; Drozd, A.; and Li, B. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, 135–148. Association for Computational Linguistics.

Son, C. V. 2018. Resource Interoperability for Sustainable Benchmarking: The Case of Events. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Wendlandt, L.; Kummerfeld, J. K.; and Mihalcea, R. 2018. Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2092–2102. Association for Computational Linguistics.