The price of debiasing automatic metrics in natural language evaluation

Arun Tejasvi Chaganty*, Stephen Mussmann*, Percy Liang
Department of Computer Science
Stanford University

Problem: existing automatic metrics are biased.

- Even if automatic metrics correlate well with human judgment at a system-level, they may have poor instance-level correlation.
- We find this is partially explained by the “low recall” of automatic metrics: many examples are systematically scored poorly.
- As a result, it is easy to improve the automatic metric without improving human scores and vice versa [7].

Average human judgment is unbiased

- Let $S(x)$ be the output produced by a system $S$ on input $x \in X$.
- We can measure the quality of $z = (x, S(x)) \in Z$ according to humans: $f(z) \equiv E[Y(z)]$, where $Y(z)$ is any one person’s judgment.
- We are interested in a system’s mean quality: $\mu = E[f(x)]$.
- Any method that matches $\mu$ in expectation is unbiased.
- Given $n$ samples of human judgments, $y^{(i)} = Y(x^{(i)})$, the simple mean estimator is unbiased:
  $$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}.$$

Can we debias automatic metrics with human feedback?

- An equivalent problem is: can we decrease the cost of unbiased human evaluation with an automatic metric?
- The key idea is that the difference between the correlated metric and human judgment will have less variance if they are correlated.
- The control variates estimator exploits this property:
  $$\tilde{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} y^{(i)} - \alpha g(x^{(i)}),$$
  where $\alpha = \text{Cov}(f(z), g(z))$ optimally scales the automatic metric.

The control-variates estimator is the best one can do*: its performance fundamentally limits the cost-savings of using automatic metrics in unbiased evaluation.

* Formally, we prove that among all unbiased estimators using only $y^{(i)}$ and $g(x^{(i)})$, and for all distributions with a given annotator variance, $\gamma = \sigma_f^2/\sigma_y^2$ and metric correlation, $\alpha$, no other estimator has a lower worst-case variance than $\tilde{\mu}_n$.

Cost savings depend only on automatic metric correlation and annotator variance

- The cost of human evaluation can be reduced by decreasing variance and thus decreasing the number of samples required.
- We measure this using data efficiency: the ratio of the variance of $\hat{\mu}_\text{max}$ and $\hat{\mu}_\text{cv}^*$:
  $$\text{DE} = \frac{\text{Var}(\hat{\mu}_\text{max})}{\text{Var}(\hat{\mu}_\text{cv})} = \frac{1 + \gamma}{1 - \rho^2 + \gamma}.$$

Tasks: text summarization and question answering

- We're interested in a system's mean quality: $\mu = E[f(x)]$.
- Any method that matches $\mu$ in expectation is unbiased.
- Given $n$ samples of human judgments, $y^{(i)} = Y(x^{(i)})$, the simple mean estimator is unbiased:
  $$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}.$$