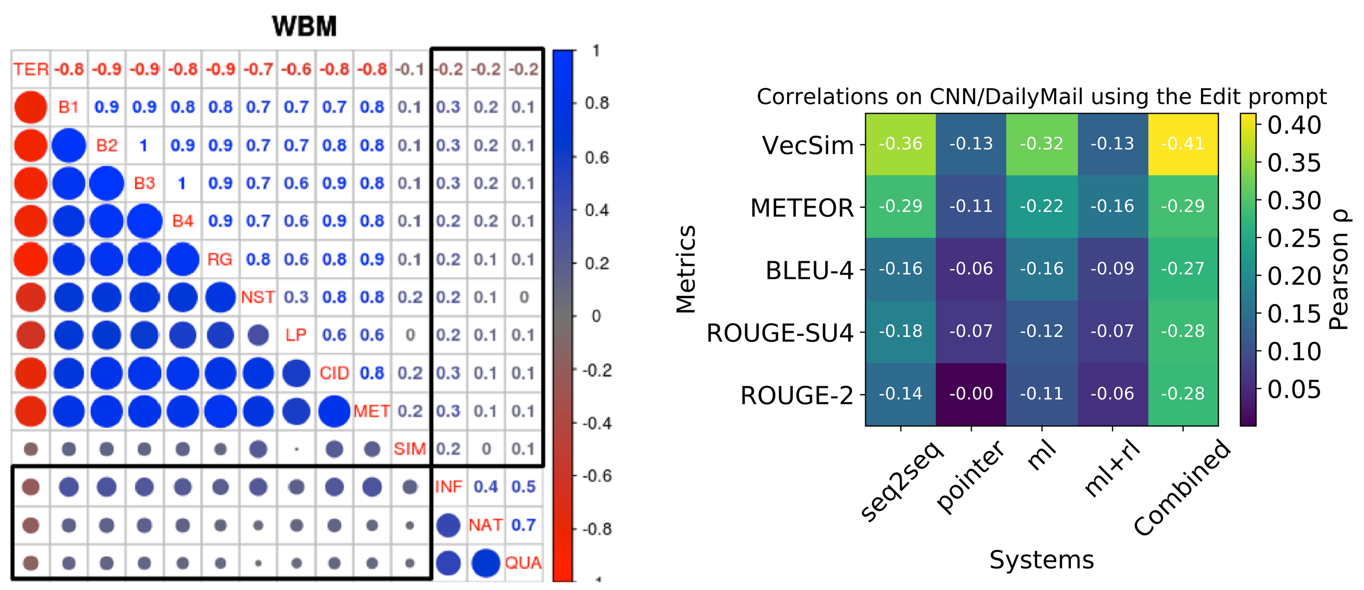
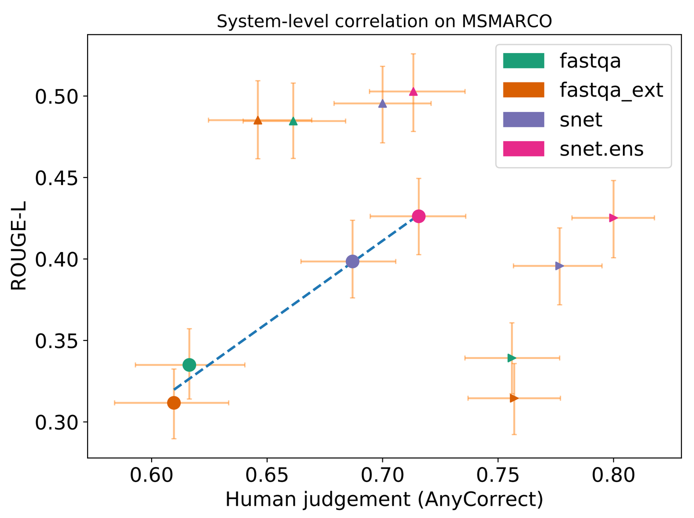
The price of debiasing automatic metrics in natural language evaluation

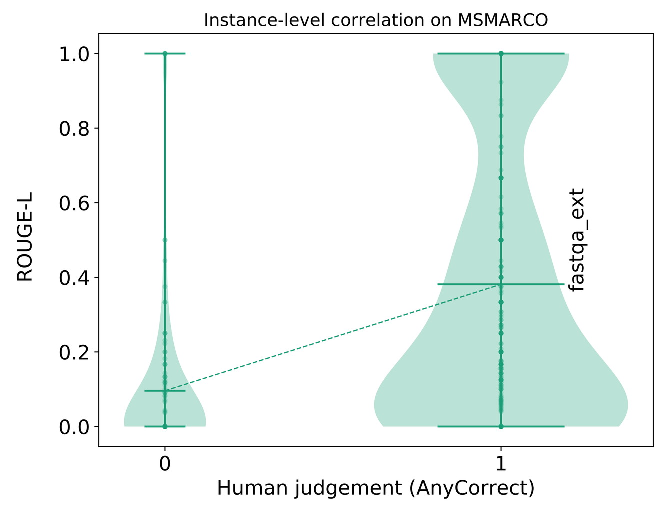
Arun Tejasvi Chaganty* Stephen Mussmann* Percy Liang
Department of Computer Science
Stanford University

Problem: existing automatic metrics are biased.



(a) Automatic metric correlations on a dialog generation task (Novikova et al., 2017). (b) Automatic metric correlations differ significantly across systems.





Correlations between automatic metrics (e.g. ROUGE or BLEU) and human judgments are poor and vary significantly between systems, making automatic evaluation hard to interpret and biased.

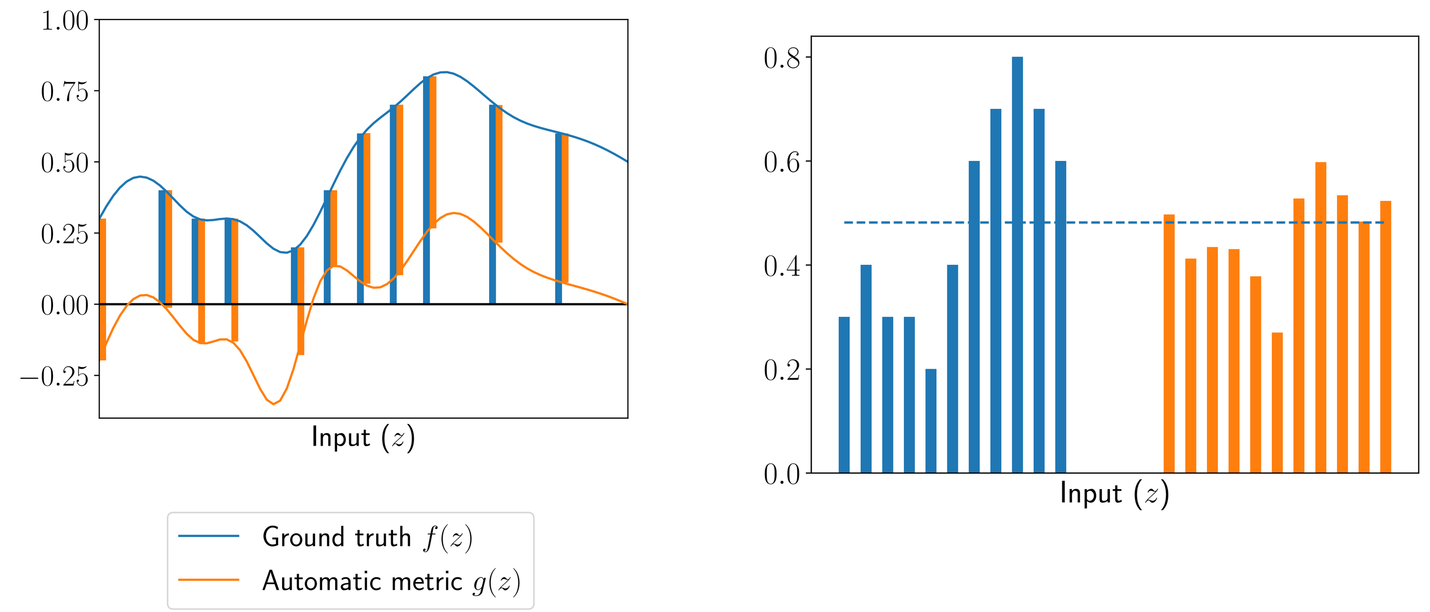
- Even if automatic metrics correlate well with human judgment at a *system-level*, they may have poor *instance-level* correlation.
- We find this is partially explained by the "low recall" of automatic metrics: many examples are systematically scored poorly.
- As a result, it is easy to improve the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

Average human judgment is unbiased

- Let $S(x)$ be the output produced by a system S on input $x \in \mathcal{X}$.
- We can measure the quality of $z = (x, S(x)) \in \mathcal{Z}$ according to humans: $f(z) \stackrel{\text{def}}{=} \mathbb{E}[Y(z)]$, where $Y(z)$ is any one person's judgment.
- We're interested in a system's *mean quality*: $\mu \stackrel{\text{def}}{=} \mathbb{E}_z[f(z)]$.
- Any method that matches μ in expectation is unbiased.
- Given n samples of human judgments, $y^{(i)} = Y(z^{(i)})$, the simple mean estimator is unbiased:

$$\hat{\mu}_{\text{mean}} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} y^{(i)}.$$

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:

$$\hat{\mu}_{\text{cv}} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} y^{(i)} - \alpha g(z^{(i)}),$$

where $\alpha = \mathbf{Cov}(f(z), g(z))$ optimally scales the automatic metric.

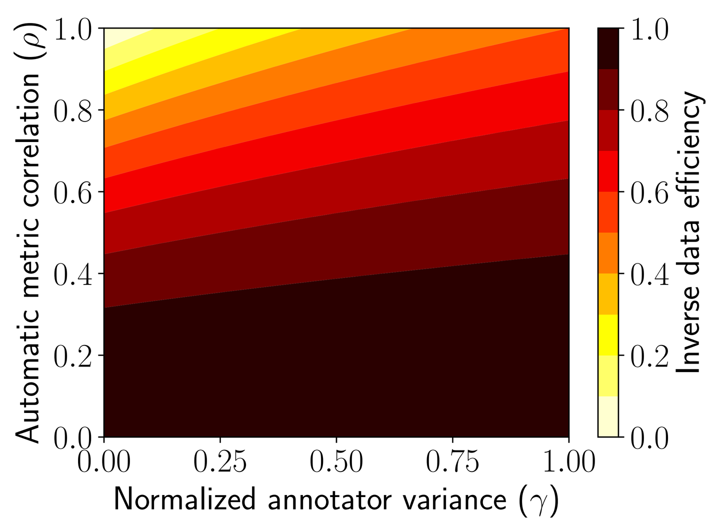
The control-variates estimator is the best one can do*: its performance fundamentally limits the cost-savings of using automatic metrics in unbiased evaluation.

*: formally, we prove that *among all unbiased estimators* using only $y^{(i)}$ and $g(z^{(i)})$, and for all distributions with a given annotator variance, $\gamma \stackrel{\text{def}}{=} \sigma_a^2/\sigma_f^2$, and metric correlation, ρ, *no other estimator has a lower worst-case variance than* $\hat{\mu}_{\text{cv}}$.

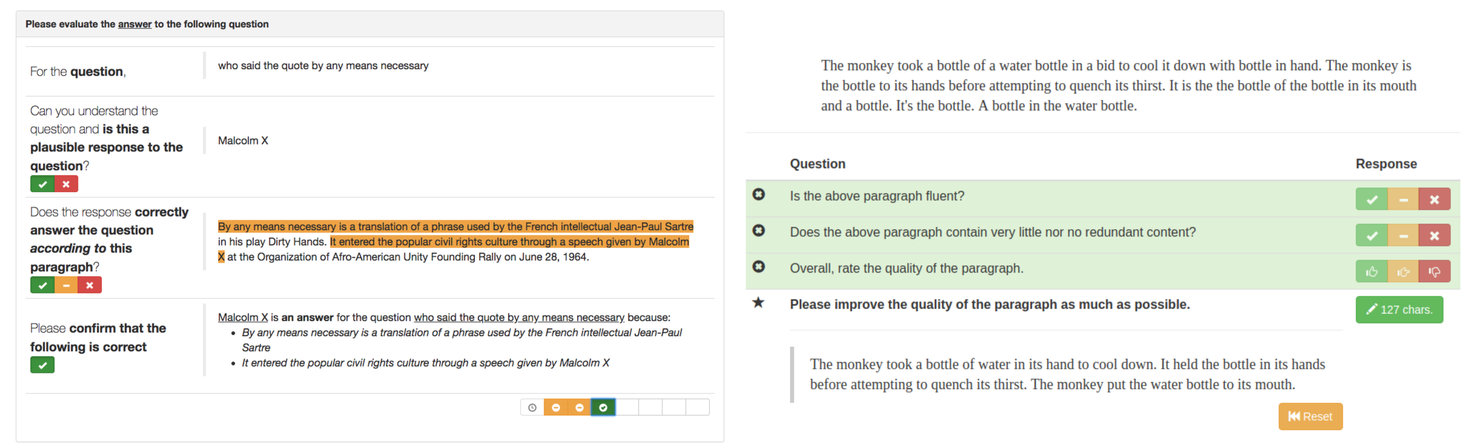
Cost savings depend only on automatic metric correlation and annotator variance

- 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 $\hat{\mu}_{\text{mean}}$ and $\hat{\mu}_{\text{cv}}$:

$$\mathbf{DE} \stackrel{\text{def}}{=} \frac{\mathbf{Var}(\hat{\mu}_{\text{mean}})}{\mathbf{Var}(\hat{\mu}_{\text{cv}})} = \frac{1+\gamma}{1-\rho^2+\gamma}.$$



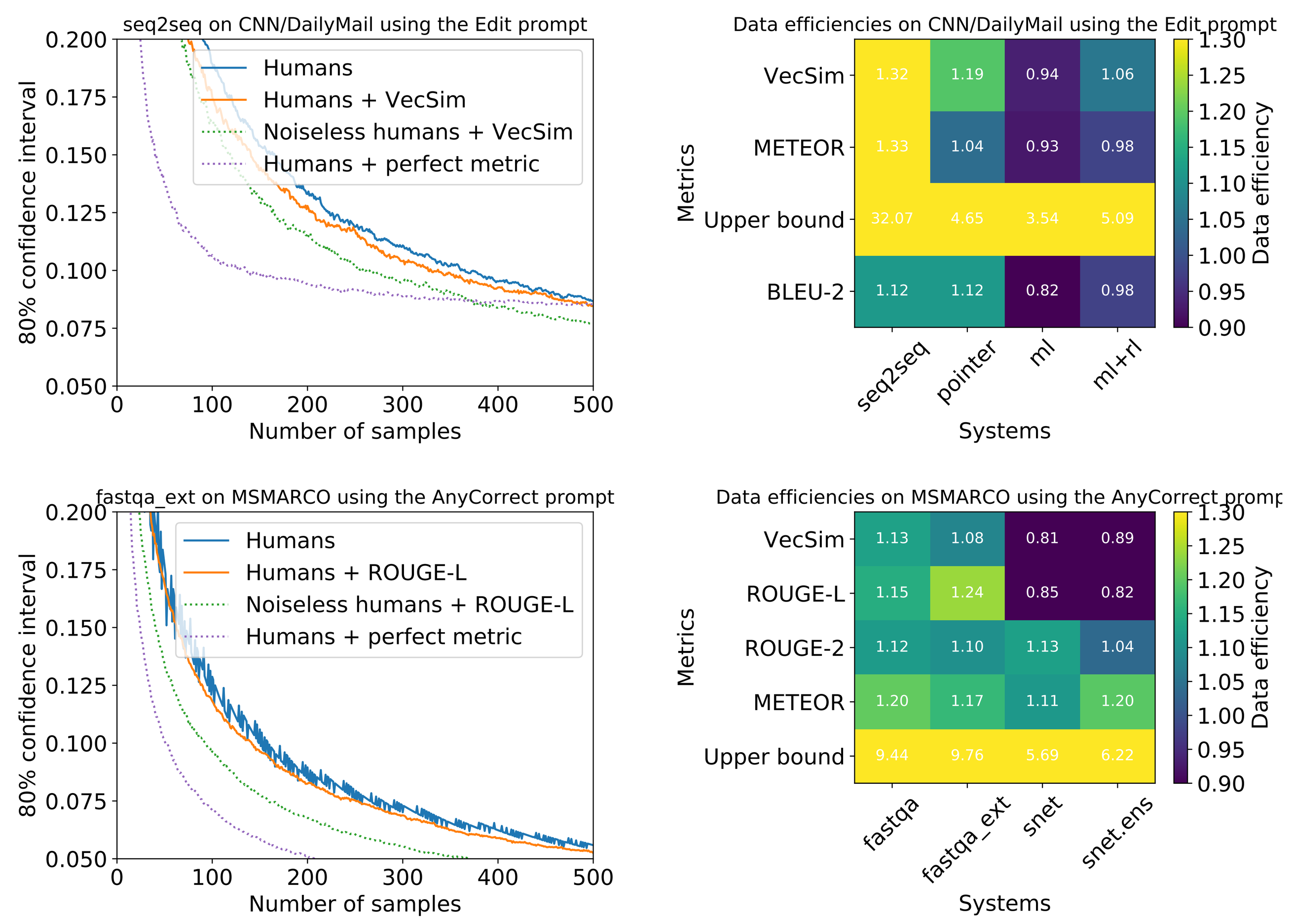
Tasks: text summarization and question answering



(a) MS MARCO ($\gamma = 0.95$) (b) CNN/Daily Mail ($\gamma = 0.36$–1.23)

Post-editing reduces annotator variance by a factor of ∼3 compared to Likert rating.

We are limited to modest data efficiencies



Both automatic metric correlation and annotator variance are important for data efficiency and current metrics and annotation interfaces severely limit possible data efficiency.

The paths forward?

- Theory shows that we can't reduce the costs of *unbiased evaluation* without dramatically improving automatic metrics (probably hard) and annotation prompts (less explored).
- Add inductive bias in how people evalaute output?
- Decompose evaluation so that we can reuse components of human judgment?
- Use stable comparison-based ranking metrics?

{chaganty,mussmann,pliang}@cs.stanford.edu